

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 625 079**

51 Int. Cl.:

C12Q 1/68

(2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **21.11.2012 PCT/US2012/066339**

87 Fecha y número de publicación internacional: **30.01.2014 WO14018080**

96 Fecha de presentación y número de la solicitud europea: **21.11.2012 E 12881774 (9)**

97 Fecha y número de publicación de la concesión europea: **22.03.2017 EP 2847347**

54 Título: **Composiciones y métodos por PCR altamente multiplexada**

30 Prioridad:

24.07.2012 US 201261675020 P

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

18.07.2017

73 Titular/es:

**NATERA, INC. (100.0%)
201 Industrial Road Suite 410
San Carlos, CA 94070, US**

72 Inventor/es:

**ZIMMERMANN, BERNHARD;
HILL, MATTHEW M.;
LACROUTE, PHILIPPE GILBERT y
DODD, MICHAEL**

74 Agente/Representante:

UNGRÍA LÓPEZ, Javier

ES 2 625 079 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

Composiciones y métodos por PCR altamente multiplexada

Referencia cruzada a aplicaciones relacionadas

Esta solicitud reivindica las ventajas y la prioridad de la Solicitud de patente USA con el número de serie 13/683.604 (publicada como US2013/0123120), presentada el 21 de noviembre de 2012, y la Solicitud provisional USA con el número de serie 61/675.020, presentada el 24 de julio de 2012. La Solicitud de patente USA con el número de serie 13/683.604 es una continuación en parte de la Solicitud de patente USA con el número de serie 13/300.235 (publicada como US2012/0270212) presentada el 18 de noviembre de 2011 es una continuación en parte de la Solicitud de patente USA con el número de serie 13/110.685 (publicada como US2011/0288780), presentada el 18 de mayo de 2011, y reivindica las ventajas de la Solicitud provisional USA con el número de serie 61/675.020, presentada el 24 de julio de 2012. La Solicitud de patente USA con el número de serie 13/110.685 reivindica las ventajas de la Solicitud provisional USA con el número de serie 61/395.850, presentada el 18 de mayo de 2010; La Solicitud provisional USA con el número de serie 61/398.159, presentada el 21 de junio de 2010; la Solicitud provisional USA con el número de serie 61/462.972, presentada el 9 de febrero de 2011; la Solicitud provisional USA con el número de serie 61/448.547, presentada el 2 de marzo de 2011; y la Solicitud provisional USA con el número de serie 61/516.996, presentada el 12 de abril de 2011. La Solicitud de patente USA con el número de serie 13/300.235 reivindica las ventajas de la Solicitud provisional USA con el número de serie 61/571.248, presentada el 23 de junio de 2011.

DECLARACIÓN RELATIVA A LA INVESTIGACIÓN O DESARROLLO PATROCINADOS FEDERALMENTE

Este trabajo ha recibido la subvención número 5R44HD60423-3, concedida por el Instituto Nacional de la Salud. El gobierno de los Estados Unidos puede tener derechos sobre cualquier patente que se publique en esta solicitud.

Campo de la invención

Por lo general, la presente invención se refiere a métodos y composiciones para amplificar simultáneamente las regiones de ácido nucleico múltiple de interés en un volumen de reacción.

Antecedentes de la invención

Para aumentar el rendimiento del ensayo y permitir un uso más eficiente las muestras de ácido nucleico, se puede llevar a cabo la amplificación simultánea de muchos ácidos nucleicos de destino en una muestra de interés mediante la combinación de muchos cebadores oligonucleótidos con la muestra y sometiendo posteriormente la muestra a las condiciones de la reacción en cadena de la polimerasa (PCR) en un proceso conocido en la técnica como PCR multiplexada. El uso de PCR multiplexada puede simplificar significativamente los procedimientos experimentales y acortar el tiempo necesario para el análisis y la detección de ácido nucleico. Sin embargo, cuando se agregan varios pares a la misma reacción de PCR, se pueden generar productos de amplificación no diana, como dímeros de cebadores amplificados. El riesgo de generar dichos productos aumenta a medida que el número de cebadores aumenta. Estos amplicones no diana limitan significativamente el uso de los productos amplificados para su ulterior análisis y/o ensayos. Por lo tanto, es necesario mejorar los métodos para reducir la formación de amplicones no diana durante la PCR multiplexada.

Los métodos de PCR multiplexada mejorados serían útiles para una serie de aplicaciones, como el Diagnóstico Genético Prenatal No Invasivo (NPD). En concreto, los métodos actuales de diagnóstico prenatal pueden alertar a los médicos y padres de anomalías en el crecimiento de los fetos. Sin diagnóstico prenatal, uno de cada 50 bebés nace con una grave deficiencia física o mental, y uno de cada 30 tendrá algún tipo de malformación congénita. Desafortunadamente, los métodos estándar tienen una calidad de precisión pobre, o implican un procedimiento invasivo que conlleva un riesgo de aborto. Los métodos basados en niveles de hormonas en la sangre materna o las mediciones por ultrasonidos no son invasivos, sin embargo, también tienen niveles de precisión bajos. Los métodos como la amniocentesis, la biopsia de vellosidades coriónica y las muestras de sangre fetal tiene una alta precisión, pero son invasivos y entrañan riesgos considerables. La amniocentesis se ha realizado en aproximadamente un 3% de todos los embarazos en los EE.UU., aunque su frecuencia de uso ha disminuido durante la última década y media.

Los seres humanos normales tienen dos juegos de 23 cromosomas en cada célula diploide sana, en los que una copia procede de cada progenitor. La aneuploidía, una condición en una célula nuclear donde la célula contiene demasiados y/o muy pocos cromosomas se considera responsable de un gran porcentaje de implantaciones fallidas, abortos, y enfermedades genéticas. La detección de anomalías cromosómicas puede identificar a personas o a embriones con condiciones tales como el síndrome de Down, síndrome de Klinefelter y el síndrome de Turner, entre otros, además de aumentar las probabilidades de tener un embarazo satisfactorio. La realización de pruebas para detectar anomalías cromosómicas es especialmente importante como también lo es la edad de la madre: entre los 35 y los 40 años se estima que al menos el 40% de los embriones son anómalos, y por encima de los 40, más de la mitad de los embriones son anómalos.

Se ha descubierto recientemente que el ADN fetal sin células y las células fetales intactas pueden entrar en la circulación sanguínea materna. Por consiguiente, el análisis de este material genético puede permitir un NPD precoz. Se desea mejorar los métodos para mejorar la sensibilidad y especificidad y reducir el tiempo y los costes necesarios para el NPD.

P. Shen et al., Proceedings of the National Academy of Sciences, vol. 18(16), 5 de abril de 2011, páginas 6549-6554, describe la utilidad de sondas candado largas (LPP) para capturar el 2xón focalizado seguido de una secuenciación basada en matriz.

Fredricksson Simon et al., Nucleic Acids Research, Vol. 35(7), 1 de febrero de 2007, páginas e47.1- e47.6, describe el uso de un método para la amplificación multiplexada de ácidos nucleicos con el fin de amplificar la secuencia codificadora de 10 genes del cáncer humano en un ensayo. Se inició la amplificación mediante PCR multiplexada con 170 pares de cebador. A continuación, cada producto de PCR se circularizó por ligadura. Las moléculas de ADN circulares se enriquecieron todavía más mediante la replicación en círculo rodante cebada aleatoriamente.

Resumen de la invención

La invención se define en las reivindicaciones que se incluyen en el anexo. En un aspecto, la invención presenta métodos de amplificación de loci diana en una muestra de ácido nucleico. En algunas realizaciones, el método consiste en (i) poner en contacto la muestra de ácido nucleico con una biblioteca de cebadores de prueba que, se hibridan simultáneamente a al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes para producir una mezcla de reacción; y (ii) someter la mezcla de reacción a las condiciones de reacción de extensión del cebador para producir productos amplificados que incluyen amplicones diana. En algunas realizaciones, el método también incluye determinar la presencia o ausencia de al menos un amplicón diana (por ejemplo, al menos un 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los amplicones diana). En algunas realizaciones, el método también incluye determinar la secuencia de al menos un amplicón diana (por ejemplo, al menos un 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los amplicones diana).

En varias realizaciones de cualquiera de los aspectos de la invención, al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes se amplifican. En algunas realizaciones, al menos un 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los productos amplificados son amplicones diana. En algunas realizaciones, al menos un 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los loci focalizados son amplificados. En varias realizaciones, menos de un 60, 50, 40, 30, 20, 10, 5, 4, 3, 2, 1, 0,5, 0,25, 0,1, o 0,05% de los productos amplificados son dímeros de cebadores. En algunas realizaciones, la biblioteca de cebadores de prueba incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 pares de cebador de prueba, en donde cada par de cebador de prueba incluye un cebador de prueba directo y un cebador de prueba inverso que se hibridan en el mismo locus diana. En algunas realizaciones,

la biblioteca de cebadores de prueba incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 cebadores de prueba individuales que se hibridan a diferentes loci seleccionados, en donde los cebadores individuales no forman parte de los pares de cebador.

En varias realizaciones de cualquiera de los aspectos de la invención, la concentración de cada cebador de prueba es inferior a 100, 75, 50, 25, 10, 5, 2, o 1 nM. En varias realizaciones, el contenido de GC de los cebadores de prueba está entre 30 y 80%, por ejemplo, entre 40 y 70% o 50 y 60%, incluidos. En algunas realizaciones, el intervalo de las temperaturas de fusión de los cebadores de prueba es inferior a 20, 15, 10, 5, 3, o 1 °C. En algunas realizaciones, la longitud de los cebadores de prueba está entre 15 y 100 nucleótidos, por ejemplo, 15 y 75 nucleótidos, 15 y 40 nucleótidos, 17 y 35 nucleótidos, 18 y 30 nucleótidos, 20 y 65 nucleótidos, incluidos. En algunas realizaciones, los cebadores de prueba incluyen una etiqueta que no es específica de la diana, como una etiqueta que forma una estructura de bucle interno. En algunas realizaciones, la etiqueta está entre dos zonas de unión de ADN. En varias realizaciones, los cebadores de prueba incluyen una región 5' que es específica para un locus diana, una región interna que no es específica para el locus diana y forma una estructura de bucle, y una región 3' que es específica del locus diana. En varias realizaciones, la longitud de la región 3' es al menos de 7 nucleótidos. En algunas realizaciones, la longitud de la región 3' está entre 7 y 20 nucleótidos, por ejemplo, entre 7 y 15 nucleótidos, o 7 y 10 nucleótidos, incluidos. En varias realizaciones, los cebadores de prueba incluyen una región 5' que no es específica para un locus diana (como una etiqueta o un punto de unión de cebador universal) seguido de una región que es específica de un locus diana, una región interna que no es específica para el locus diana y forma una estructura de bucle, y una región 3' que es específica del locus diana. En algunas realizaciones, el intervalo de la longitud de los cebadores de prueba es inferior a 50, 40, 30, 20, 10, o 5 nucleótidos. En algunas realizaciones, la longitud de los amplicones diana está entre 50 y 100 nucleótidos, por ejemplo, entre 60 y 80 nucleótidos, o 60 y 75 nucleótidos, incluidos. En algunas realizaciones, el intervalo de la longitud de los amplicones diana es inferior a 50, 25, 15, 10, o 5 nucleótidos.

En varias realizaciones decualquiera de los aspectos de la invención, las condiciones de reacción de extensión del cebador son condiciones de reacción en cadena de la polimerasa (PCR). En varias realizaciones, la longitud de la fase de reformación térmica es superior a 3, 5, 8, 10, o 15 minutos. En varias realizaciones, la longitud de la fase de extensión es superior a 3, 5, 8, 10, o 15 minutos.

En varias realizaciones de cualquiera de los aspectos de la invención, los cebadores de prueba se utilizan para amplificar simultáneamente al menos 1.000 loci diana diferentes en una muestra que incluye ADN materno de la madre embarazada de un feto y el ADN fetal para determinar la presencia o ausencia de una anomalía cromosómica fetal. En varias realizaciones, el método incluye ligar un punto de unión del cebador universal a las moléculas de ADN en la muestra; amplificando las moléculas de ADN ligadas utilizando al menos 1.000 cebadores específicos y un cebador universal para producir un primer conjunto de productos amplificados; y amplificar el primer conjunto de productos amplificados utilizando al menos 1.000 pares

de cebadores específicos para producir un segundo conjunto de productos amplificados. En varias realizaciones, se utilizan al menos 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 pares de cebador diferentes.

5 En varias realizaciones de cualquiera de los aspectos de la invención, los cebadores de prueba se utilizan para amplificar simultáneamente al menos 1.000 loci diana diferentes en una muestra que incluye ADN de un supuesto padre de un feto y al mismo tiempo amplificar los loci diana seleccionado en una muestra que incluye ADN materno de la madre embarazada del feto y el ADN fetal para determinar si el presunto padre es el padre biológico del feto.

10 En varias realizaciones de cualquiera de los aspectos de la invención, los cebadores de prueba se utilizan para amplificar simultáneamente al menos 1.000 loci diana diferentes en una célula o múltiples células de un embrión para determinar la presencia o ausencia de una anomalía cromosómica. En varias realizaciones, se analizan las células de un conjunto de dos o más embriones, y se selecciona un embrión para la fertilización in vitro.

En varias realizaciones de cualquiera de los aspectos de la invención, los cebadores de prueba se utilizan para amplificar simultáneamente al menos 1.000 loci diana diferentes en una muestra de ácido nucleico forense. En varias realizaciones, la longitud de la fase de reformación térmica es superior a 3, 5, 8, 10, o 15 minutos.

15 En varias realizaciones de cualquiera de los aspectos de la invención, el método implica el uso de los cebadores de prueba para amplificar simultáneamente al menos 1.000 loci diana diferentes en una muestra de ácido nucleico de control para producir un primer conjunto de amplicones diana y para amplificar simultáneamente los loci diana en una muestra de ácido nucleico de prueba para producir un segundo conjunto de amplicones diana. y comparar el primer y el segundo conjuntos de amplicones diana para determinar si un locus diana está presente en una muestra, pero ausente en la otra, o si un locus diana está presente en diferentes niveles en la muestra de control y la muestra de prueba. En varias realizaciones, la muestra de prueba es de un individuo sospechoso de tener una enfermedad o fenotipo de interés (como cáncer), o un mayor riesgo de una enfermedad o fenotipo de interés; y en donde uno o más de los loci diana incluye una secuencia (por ejemplo, un polimorfismo u otra mutación) asociados con un mayor riesgo de desarrollar la enfermedad o fenotipo de interés, o asociados con la enfermedad o fenotipo de interés. En varias realizaciones, el método implica el uso de los cebadores de prueba para amplificar simultáneamente 1.000; loci diana diferentes en una muestra de control que incluye el ARN para producir un primer conjunto de amplicones diana y para amplificar simultáneamente los loci diana en una muestra de prueba que incluye el ARN para producir un segundo conjunto de amplicones diana; y comparar el primer y el segundo conjuntos de amplicones diana para determinar la presencia o ausencia de una diferencia en los niveles de expresión de ARN entre la muestra de control y la muestra de prueba. En varias realizaciones, el ARN es mRNA. En varias realizaciones, la muestra de prueba es de un individuo sospechoso de tener una enfermedad o fenotipo de interés (como cáncer), o un mayor riesgo de una enfermedad o fenotipo de interés (como cáncer); y en el cual uno o más de los loci diana incluye una secuencia (por ejemplo, un polimorfismo u otra mutación) asociados con un mayor riesgo de desarrollar la enfermedad o fenotipo de interés, o asociados con la enfermedad o fenotipo de interés. En algunas realizaciones, la muestra de prueba es de un individuo diagnosticado con una enfermedad o fenotipo de interés (como cáncer); y en donde una diferencia en el nivel de expresión de ARN entre la muestra de control y la muestra de prueba indica que un locus diana incluye una secuencia (por ejemplo, un polimorfismo u otra mutación) asociados con un mayor o menor riesgo de desarrollar la enfermedad o fenotipo de interés.

40 En algunas realizaciones de cualquiera de los aspectos de la invención, los cebadores de prueba se seleccionan de una biblioteca de cebadores candidatos en base a uno o más parámetros, como la selección de cebadores utilizando cualquiera de los métodos de la invención. En algunas realizaciones, los cebadores de prueba se seleccionan de una biblioteca de cebadores candidatos basándose al menos en parte en la capacidad de los cebadores candidatos de formar dímeros de cebadores.

45 En un aspecto, la invención presenta métodos de selección de cebadores de prueba de una biblioteca de cebadores candidatos. En varias realizaciones, la selección implica (i) calcular en un ordenador una puntuación no deseable para la mayoría o todas las combinaciones posibles de dos cebadores candidatos de la biblioteca, donde cada puntuación no deseable se basa, al menos en parte, en la probabilidad de la formación de dímeros entre los dos cebadores candidatos; (ii) extraer el cebador candidato con la puntuación no deseable más alta de la biblioteca de cebadores candidatos; y (iii) si el cebador candidato que se ha extraído en el paso (ii) es un miembro de un par de cebadores, extraer, a continuación, el otro miembro del par de cebadores de la biblioteca de cebadores candidatos; y (iv) opcionalmente repetir los pasos (ii) y (iii), con el fin de seleccionar una biblioteca de cebadores de prueba. En algunas realizaciones, el

método de selección se lleva a cabo hasta que las puntuaciones no deseables para las combinaciones del cebador candidato restantes en la biblioteca son todas iguales o por debajo de un umbral mínimo. En algunas realizaciones, el método de selección se lleva a cabo hasta que el número de cebadores candidatos restantes en la biblioteca se reduce a un número deseado. En varias realizaciones, una puntuación no deseable se calcula para al menos 80, 90, 95, 98, 99, o 99,5% de las combinaciones posibles de cebadores candidatos en la biblioteca. En varias realizaciones, los cebadores candidatos restantes en la biblioteca pueden amplificar simultáneamente al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes. En varias realizaciones, el método también incluye (v) poner en contacto una muestra de ácido nucleico que incluye loci diana con los cebadores candidatos restantes en la biblioteca para producir una mezcla de reacción; y (vi) someter la mezcla de reacción a las condiciones de reacción de extensión del cebador para producir productos amplificados que incluye amplicones diana.

En un aspecto, la invención presenta métodos de selección de cebadores de prueba de una biblioteca de cebadores candidatos. En varias realizaciones, la selección de cebadores de prueba se selecciona de una biblioteca de cebadores candidatos e implica (i) calcular en un ordenador una puntuación no deseable para la mayoría o todas las combinaciones posibles de dos cebadores candidatos de la biblioteca, donde cada puntuación no deseable se basa, al menos en parte, en la probabilidad de la formación de dímeros entre los dos cebadores candidatos; (ii) extraer de la biblioteca de cebadores candidatos el cebador candidato que es parte del mayor número de combinaciones de dos cebadores candidatos con una puntuación no deseable por encima de un primer umbral mínimo; (iii) si el cebador candidato que se ha extraído en el paso (ii) es un miembro de un par de cebadores, extraer, a continuación, el otro miembro del par de cebadores de la biblioteca de cebadores candidatos; y (iv) opcionalmente repetir los pasos (ii) y (iii), con el fin de seleccionar una biblioteca de cebadores de prueba. En algunas realizaciones, el método de selección se lleva a cabo hasta que las puntuaciones no deseables para las combinaciones del cebador candidato restantes en la biblioteca son todas iguales o por debajo del primer umbral mínimo. En algunas realizaciones, el método de selección se lleva a cabo hasta que el número de cebadores candidatos restantes en la biblioteca se reduce a un número deseado. En varias realizaciones, una puntuación no deseable se calcula para al menos 80, 90, 95, 98, 99, o 99,5% de las combinaciones posibles de cebadores candidatos en la biblioteca. En varias realizaciones, los cebadores candidatos restantes en la biblioteca pueden amplificar simultáneamente al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes. En varias realizaciones, el método también incluye (v) poner en contacto una muestra de ácido nucleico que incluye loci diana con los cebadores candidatos restantes en la biblioteca para producir una mezcla de reacción; y (ii) someter la mezcla de reacción a las condiciones de reacción de extensión del cebador para producir productos amplificados que incluye amplicones diana.

En algunas realizaciones de cualquiera de los aspectos de la invención, el método de selección implica reducir todavía más el número de cebadores candidatos restantes en la biblioteca disminuyendo el primer umbral mínimo utilizado en el paso (ii) a un segundo umbral mínimo más bajo y opcionalmente repetir los pasos (ii) y (iii). En algunas realizaciones, el método de selección implica aumentar el primer umbral mínimo utilizado en el paso (ii) a un segundo umbral mínimo más alto y opcionalmente repetir los pasos (ii) y (iii). En algunas realizaciones, el método de selección se lleva a cabo hasta que las puntuaciones no deseables para las combinaciones del cebador candidato restantes en la biblioteca son todas iguales o por debajo del segundo umbral mínimo, o hasta que el número de cebadores candidato restantes en la biblioteca se reduce a un número deseado.

En varias realizaciones de cualquiera de los aspectos de la invención, el método implica, antes del paso (i), identificar o seleccionar los cebadores que se hibridan al loci diana. En algunas realizaciones, múltiples cebadores (o pares de cebador) se hibridan al mismo locus diana, y el método de selección se utiliza para seleccionar un cebador (o un par de cebadores) para este locus diana en base a uno o más parámetros. En varias realizaciones, el método implica, antes del paso (ii), quitar un par de cebadores de la biblioteca que produce un amplicón diana que se solapa con un amplicón diana producido por otro par de cebadores. En varias realizaciones, se selecciona un cebador candidato de un grupo de dos o más cebadores candidatos con puntuaciones no deseables iguales para quitarlos de la biblioteca de cebadores candidatos en base a uno o más parámetros distintos. En algunas realizaciones, los cebadores candidatos restantes en la biblioteca se utilizan como una biblioteca de cebadores de prueba en cualquiera de los métodos de la invención. En algunas realizaciones, la biblioteca de cebadores de prueba resultante incluye cualquiera de los cebadores de prueba de la invención.

En varias realizaciones de cualquiera de los aspectos de la invención, las puntuaciones no deseables se basan, al menos en parte, en uno o más parámetros seleccionados del grupo compuesto por la tasa de heterocigosidad del locus diana, el predominio de la enfermedad asociada con una secuencia (por ejemplo, un polimorfismo) en el locus diana, la penetrancia de la enfermedad asociada con una secuencia (por ejemplo, un polimorfismo) en el locus diana, la especificidad del cebador candidato para el locus diana, el tamaño del cebador candidato, la temperatura de fusión del amplicón diana, el contenido de GC del amplicón diana, la eficiencia de la amplificación del amplicón diana, y el tamaño del amplicón diana.

En varias realizaciones de cualquiera de los aspectos de la invención, las puntuaciones no deseables se basan, al menos en parte, en uno o más parámetros seleccionados del grupo compuesto por la tasa de heterocigosidad del locus diana, la especificidad del cebador candidato para el locus diana; el tamaño del cebador candidato, la temperatura de fusión del amplicón diana, el contenido de GC del amplicón diana, la eficiencia de la amplificación del amplicón diana, y el tamaño del amplicón diana; y los cebadores de prueba se utilizan para amplificar simultáneamente al menos 1.000 loci diana diferentes en una muestra que incluye ADN materno de la madre embarazada de un feto y el ADN fetal para determinar la presencia o ausencia de una anomalía cromosómica fetal. En varias realizaciones, el método incluye ligar un punto de unión del cebador universal a las moléculas de ADN en la muestra; amplificar las moléculas de ADN ligadas utilizando al menos 1.000 cebadores específicos y un cebador universal para producir un primer conjunto de productos amplificados; y amplificar el primer conjunto de productos amplificados utilizando al menos 1.000 pares de cebadores específicos para producir un segundo conjunto de productos amplificados. En varias realizaciones, se utilizan al menos 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 pares de cebador diferentes. En varias realizaciones, al menos 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes se amplifican.

En varias realizaciones de cualquiera de los aspectos de la invención, las puntuaciones no deseables se basan, al menos en parte, en uno o más parámetros seleccionados del grupo compuesto por la tasa de heterocigosidad del locus diana, la especificidad del cebador candidato para el locus diana; el tamaño del cebador candidato, la temperatura de

fusión del amplicón diana, el contenido de GC del amplicón diana, la eficiencia de la amplificación del amplicón diana, y el tamaño del amplicón diana; y los cebadores de prueba se utilizan para amplificar simultáneamente al menos 1.000 loci diana diferentes en una muestra que incluye ADN de un supuesto padre de un feto y al mismo tiempo amplificar los loci diana en una muestra que incluye ADN materno de la madre embarazada del feto y el ADN fetal para determinar si el presunto padre es el padre biológico del feto. En varias realizaciones, al menos 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes se amplifican.

En varias realizaciones de cualquiera de los aspectos de la invención, las puntuaciones no deseables se basan, al menos en parte, en uno o más parámetros seleccionados del grupo compuesto por la tasa de heterocigosidad del locus diana, la especificidad del cebador candidato para el locus diana; el tamaño del cebador candidato, la temperatura de fusión del amplicón diana, el contenido de GC del amplicón diana, la eficiencia de la amplificación del amplicón diana, y el tamaño del amplicón diana; y los cebadores de prueba se utilizan para amplificar simultáneamente al menos 1.000 loci diana diferentes en una célula o múltiples células de un embrión para determinar la presencia o ausencia de una anomalía cromosómica. En varias realizaciones, se analizan las células de un conjunto de dos o más embriones, y se selecciona un embrión para la fertilización.

En varias realizaciones, al menos 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes se amplifican.

En varias realizaciones de cualquiera de los aspectos de la invención, las puntuaciones no deseables se basan, al menos en parte, en uno o más parámetros seleccionados del grupo compuesto por la tasa de heterocigosidad del locus diana, la especificidad del cebador candidato para el locus diana; el tamaño del cebador candidato, la temperatura de fusión del amplicón diana, el contenido de GC del amplicón diana, la eficiencia de la amplificación del amplicón diana, y el tamaño del amplicón diana; y los cebadores de prueba se utilizan para amplificar simultáneamente al menos 1.000 loci diana diferentes en una muestra de ácido nucleico forense. En varias realizaciones, la longitud de la fase de reformación térmica es superior a 3, 5, 8, 10, o 15 minutos. En varias realizaciones, al menos 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes se amplifican.

En varias realizaciones de cualquiera de los aspectos de la invención, las puntuaciones no deseables se basan, al menos en parte, en uno o más parámetros seleccionados del grupo compuesto por la tasa de heterocigosidad del locus diana, el predominio de la enfermedad asociada con una secuencia (por ejemplo, un polimorfismo) en el locus diana, la penetrancia de la enfermedad asociada con una secuencia (por ejemplo, un polimorfismo) en el locus diana, la especificidad del cebador candidato para el locus diana, el tamaño del cebador candidato, la temperatura de fusión del amplicón diana, el contenido de GC del amplicón diana, la eficiencia de la amplificación del amplicón diana, y el tamaño del amplicón diana; y el método implica el uso de los cebadores de prueba para amplificar simultáneamente al menos 1.000 loci diana diferentes en una muestra de ácido nucleico de control para producir un primer conjunto de amplicones diana y para amplificar simultáneamente los loci diana en una muestra de ácido nucleico de prueba para producir un segundo conjunto de amplicones diana; y comparar el primer y el segundo conjuntos de amplicones diana para determinar si un locus diana está presente en una muestra, pero ausente en la otra, o si un locus diana está presente en diferentes niveles en la muestra de control y la muestra de prueba. En varias realizaciones, la muestra de prueba es de un individuo sospechoso de tener una enfermedad o fenotipo de interés, o un mayor riesgo de una enfermedad o fenotipo de interés; y en donde uno o más de los loci diana incluye una secuencia (por ejemplo, un polimorfismo) en el locus diana asociado con un mayor riesgo de desarrollar la enfermedad o fenotipo de interés, o asociado con la enfermedad o fenotipo de interés. En varias realizaciones, al menos 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes se amplifican.

En varias realizaciones de cualquiera de los aspectos de la invención, las puntuaciones no deseables se basan, al menos en parte, en uno o más parámetros seleccionados del grupo compuesto por la tasa de heterocigosidad del locus diana, el predominio de la enfermedad asociada con una secuencia (por ejemplo, un polimorfismo) en el locus diana, la penetrancia de la enfermedad asociada con una secuencia (por ejemplo, un polimorfismo) en el locus diana, la especificidad del cebador candidato para el locus diana, el tamaño del cebador candidato, la temperatura de fusión del amplicón diana, el contenido de GC del amplicón diana, la eficiencia de la amplificación del amplicón diana, y el tamaño del amplicón diana; y, el método implica el uso de los cebadores de prueba para amplificar simultáneamente 1.000; diferentes loci diana en una muestra de control que incluye el ARN para producir un primer conjunto de amplicones diana y para amplificar simultáneamente los loci diana en una muestra de prueba que incluye el ARN para producir un segundo conjunto de amplicones diana; y comparar el primer y el segundo conjuntos de amplicones diana para determinar la presencia o ausencia de una diferencia en los niveles de expresión de ARN entre la muestra de control y la muestra de prueba. En varias realizaciones, el ARN es mRNA. En varias realizaciones, la muestra de prueba es de un individuo sospechoso de tener una enfermedad o fenotipo de interés (como cáncer), o un mayor riesgo de una enfermedad o fenotipo de interés (como cáncer); y en donde uno o más de los loci diana incluye una secuencia (por ejemplo, un polimorfismo u otra mutación) asociados con un mayor riesgo de desarrollar la enfermedad o fenotipo de interés, o asociados con la enfermedad o fenotipo de interés. En algunas realizaciones, la muestra de prueba es de un individuo diagnosticado con una enfermedad o fenotipo de interés (como cáncer); y en donde una diferencia en el nivel de expresión de ARN entre la muestra de control y la muestra de prueba indica que un locus diana incluye una secuencia (por ejemplo, un polimorfismo u otra mutación) asociados con un mayor o menor riesgo de desarrollar la enfermedad o fenotipo de interés. En varias realizaciones, al menos 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes se amplifican.

En un aspecto, la invención presenta bibliotecas de cebadores. En algunas realizaciones, los cebadores se seleccionan de una biblioteca de cebadores candidatos utilizando cualquiera de los métodos de la invención. En algunas realizaciones, la biblioteca incluye cebadores que se hibridan simultáneamente hasta al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes. En algunas realizaciones, la biblioteca incluye cebadores que amplifican simultáneamente al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes. En algunas realizaciones, la biblioteca incluye cebadores que amplifican simultáneamente al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes, de tal modo que menos de 60, 40, 30, 20, 10, 5, 4, 3, 2, 1, 0,5, 0,25, 0,1, o 0,05% de los productos amplificados son dímeros de cebadores. En algunas realizaciones, la biblioteca incluye cebadores que amplifican simultáneamente 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes, de tal modo que al menos 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los productos amplificados son amplicones diana. En algunas realizaciones, la biblioteca incluye cebadores que amplifican simultáneamente loci focalizados de tal modo que al menos 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los loci focalizados de 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 75.000; o 100.000 loci diana diferentes se amplifican. En algunas realizaciones, la biblioteca de cebadores incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 pares de cebador de prueba, en donde cada par de cebadores incluye un cebador de prueba directo y un cebador de prueba inverso en el que cada par de cebadores de prueba se hibrida en un locus diana. En algunas realizaciones, la biblioteca de cebadores incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 cebadores de prueba individuales que se hibridan cada uno de ellos a diferentes locus diana, en donde los cebadores individuales no forman parte de los pares de cebador.

En varias realizaciones de cualquiera de los aspectos de la invención, la concentración de cada cebador es inferior a 100, 75, 50, 25, 10, 5, 2, o 1 nM. En varias realizaciones, el contenido de GC de los cebadores está entre 30 y 80%, por ejemplo, entre 40 y 70% o 50 y 60%, incluidos. En algunas realizaciones, el intervalo de contenido de GC de los cebadores es inferior a 30, 20, 10, o 5%. En varias realizaciones, la temperatura de fusión de los cebadores está entre 40 y 80 °C, por ejemplo, entre 50 y 70 °C, 55 y 65 °C, o 57 y 60,5 °C, incluidos. En algunas realizaciones, el intervalo de la temperatura de fusión de los cebadores es inferior a 15, 10, 5, 3, o 1 °C. En algunas realizaciones, la longitud de los cebadores está entre 15 y 100 nucleótidos, por ejemplo, 15 y 75 nucleótidos, 15 y 40 nucleótidos, 17 y 35 nucleótidos, 18 y 30 nucleótidos, o 20 y 65 nucleótidos, incluidos. En algunas realizaciones, los cebadores incluyen una etiqueta que no es específica de la diana, como una etiqueta que forma una estructura de bucle interno. En algunas realizaciones, la etiqueta está entre dos zonas de unión de ADN. En varias realizaciones, los cebadores incluyen una región 5' que es específica para un locus diana, una región interna que no es específica para el locus diana y forma una estructura de bucle, y una región 3' que es específica del locus diana. En varias realizaciones, la longitud de la región 3' es al menos de 7 nucleótidos. En algunas realizaciones, la longitud de la región 3' está entre 7 y 20 nucleótidos, por ejemplo, entre 7 y 15 nucleótidos, o 7 y 10 nucleótidos, incluidos. En varias realizaciones, los cebadores incluyen una región 5' que no es específica para un locus diana (como otra etiqueta o un punto de unión del cebador universal) seguido de una región que es específica de un locus diana, una región interna que no es específica para el locus diana y forma una estructura de bucle, y una región 3' que es específica del locus diana. En algunas realizaciones, el intervalo de la longitud de los cebadores es inferior a 50, 40, 30, 20, 10, o 5 nucleótidos. En algunas realizaciones, la longitud de los amplicones diana está entre 50 y 100 nucleótidos, por ejemplo, entre 60 y 80 nucleótidos, o 60 y 75 nucleótidos, incluidos. En algunas realizaciones, el intervalo de la longitud de los amplicones diana es inferior a 50, 25, 15, 10, o 5 nucleótidos.

En un aspecto, la presente divulgación proporciona kits que incluyen cualquier de las bibliotecas de cebadores de la invención para amplificar loci diana en una muestra de ácido nucleico. En algunas realizaciones, el kit incluye instrucciones para utilizar la biblioteca para amplificar los loci diana.

En un aspecto, la presente divulgación presenta métodos para determinar un estado de ploidía de cromosoma en un feto en gestación. En algunas realizaciones, el método consiste en poner en contacto una muestra de ácido nucleico con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes para producir una mezcla de reacción; en donde la muestra de ácido nucleico incluye ADN materno de la madre del feto y ADN fetal del feto. En algunas realizaciones, la mezcla de reacción se somete a condiciones de reacción de extensión del cebador para producir productos amplificados; los productos amplificados se miden con un secuenciador de alto rendimiento

para producir datos de secuenciación; los recuentos de alelos en los loci polimórficos se calculan en un ordenador en base a los datos de secuenciación; en un ordenador se crean una pluralidad de hipótesis de ploidía cada una de ellas perteneciente a un estado de ploidía posible diferente del cromosoma; un modelo de distribución conjunto para los recuentos de alelos esperados en los loci polimórficos en el cromosoma se genera en un ordenador para cada hipótesis de ploidía; una probabilidad relativa de cada hipótesis de ploidía se determina en un ordenador utilizando el modelo de distribución conjunto y los recuentos de alelos; y el estado de ploidía del feto se determina seleccionando el estado de ploidía correspondiente a la hipótesis con la mayor probabilidad.

En un aspecto, la divulgación presenta métodos para determinar un estado de ploidía de un cromosoma en un feto en gestación. En una realización un método para determinar un estado de ploidía de un cromosoma en un feto en gestación incluye la obtención de una primera muestra de ADN que comprende ADN materno de la madre del feto y ADN fetal del feto, la preparación de la primera muestra aislando el ADN para obtener una muestra preparada, la medición del ADN en la muestra preparada a una pluralidad de loci polimórficos en el cromosoma, el cálculo, en un

ordenador, de los recuentos de alelos en la pluralidad de loci polimórficos de las mediciones de ADN realizadas en la muestra preparada, la creación, en un ordenador, de una pluralidad de hipótesis de ploidía cada una de ellas perteneciente a un estado de ploidía posible diferente del cromosoma, la creación, en un ordenador, de un

modelo de distribución conjunto para los recuentos de alelos esperados en la pluralidad de loci polimórficos en el cromosoma para cada hipótesis de ploidía, la determinación, en un ordenador, de una probabilidad relativa de cada una de las hipótesis de ploidía utilizando el modelo de distribución conjunto y los recuentos de alelos medidos en la muestra preparada, y la determinación del estado de ploidía del feto seleccionando el estado de ploidía correspondiente a la hipótesis con la mayor probabilidad.

En un aspecto, la divulgación presenta métodos de realizar pruebas para determinar la distribución anómala de un cromosoma en una muestra que incluye una mezcla de ADN materno y fetal. En algunas realizaciones, el método consiste en (i) poner en contacto la muestra con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000. 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes para producir una mezcla de reacción; en la que los loci diana son de una pluralidad de cromosomas diferentes; y en la que la pluralidad de cromosomas diferentes incluye al menos un primer cromosoma sospechoso de tener una distribución anómala en la muestra y al menos un segundo cromosoma que se supone que está distribuido normalmente en la muestra; (ii) someter la mezcla de reacción a las condiciones de reacción de extensión del cebador para producir productos amplificados; (iii) secuenciar los productos amplificados para obtener una pluralidad de etiquetas de secuencia que se alineen con los loci diana; en donde las etiquetas de secuencia tienen la longitud suficiente para ser asignadas a un locus diana específico; (iv) asignar en un ordenador la pluralidad de etiquetas de secuencia a sus loci diana correspondientes; (v) determinar en un ordenador un número de etiquetas de secuencia que se alinean con los loci diana del primer cromosoma y un número de etiquetas de secuencia que se alinean con los loci diana del segundo cromosoma; y (vi) comparar en un ordenador los números del paso (v) para determinar la presencia o ausencia de una distribución anómala del primer cromosoma.

En un aspecto, la divulgación proporciona métodos para detectar la presencia o ausencia de una aneuploidía fetal. En algunas realizaciones, el método consiste en (i) poner en contacto una muestra que incluye una mezcla de ADN materno y fetal con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000. 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana no polimórficos diferentes para producir una mezcla de reacción; en la que los loci diana son de una pluralidad de cromosomas diferentes; (ii) someter la mezcla de reacción a las condiciones de reacción de extensión del cebador para producir productos amplificados que incluyen amplicones diana; (iii) cuantificar en un ordenador una frecuencia relativa de los

amplicones diana del primer y segundo cromosomas de interés; (iv) comparar en un ordenador la frecuencia relativa de los amplicones diana del primer y segundo cromosoma de interés; e (v) identificar la presencia o ausencia de una aneuploidía en base a las frecuencias relativas comparadas del primer y segundo cromosoma de interés. En algunas realizaciones, el primer cromosoma es un cromosoma sospechoso de ser euploide. En algunas realizaciones, el segundo cromosoma es un cromosoma sospechoso de ser aneuploide.

En un aspecto, se describe un método para determinar la presencia o ausencia de aneuploidía fetal en una muestra de tejido materno que comprende ADN genómico materno y fetal, el método incluye (a) la obtención de una mezcla de ADN genómico materno y fetal de dicha muestra de tejido materno, (b) la realización de una secuenciación de ADN paralela masivamente de fragmentos de ADN seleccionados aleatoriamente de la mezcla de ADN genómico materno y fetal del paso (a) para determinar la secuencia de dichos fragmentos de ADN, (c) la identificación de los cromosomas a los que las secuencias obtenidas en el paso (b) pertenecen, (d) la utilización de los datos del paso (c) para determinar una cantidad de al menos un primer cromosoma en dicha mezcla de ADN genómico materno y fetal, en donde al menos dicho primer cromosoma se supone que es euploide en el feto, (e) la utilización de los datos del paso (c) para determinar una cantidad de un segundo cromosoma en dicha mezcla de ADN genómico materno y fetal, en donde dicho segundo cromosoma se sospecha que es aneuploide en el feto, (f) el cálculo de la fracción de ADN fetal en la mezcla de ADN materno y fetal, (g) el cálculo de una distribución esperada de la cantidad del segundo cromosoma diana si el segundo cromosoma diana es euploide, la utilización del número en el paso (d), (h) el cálculo de una distribución esperada de la cantidad del segundo cromosoma diana si el segundo cromosoma diana es aneuploide, la utilización del primer número del paso (d) y la fracción calculada de ADN fetal en la mezcla de ADN materno y fetal en el paso (f), y (i) utilizando una probabilidad máxima o el enfoque a posteriori máximo para determinar si la cantidad del segundo cromosoma según se ha determinado en el paso (e) es más probable que sea parte de la distribución calculada en el paso (g) o la distribución calculada en el paso (h); indicando, por tanto, la presencia o ausencia de una aneuploidía fetal.

En varias realizaciones de cualquiera de los aspectos de la divulgación, el método también incluye la obtención de datos genotípicos de uno o ambos padres del feto. En algunas realizaciones, obtener datos genotípicos de uno o ambos padres del feto incluye preparar el ADN de los padres en donde la preparación comprende preferentemente enriquecer el ADN en la pluralidad de loci polimórficos para dar ADN de los padres preparado, opcionalmente amplificar el ADN de los padres preparado, y medir el ADN de los padres en la muestra preparada en la pluralidad de loci polimórficos.

En varias realizaciones de cualquiera de los aspectos de la divulgación, la creación de un modelo de distribución conjunto para las probabilidades de recuento de alelos esperado de la pluralidad de loci polimórficos en el cromosoma se lleva a cabo utilizando los datos genéticos obtenidos de uno o de ambos padres. En algunas realizaciones, la muestra (por ejemplo, la primera muestra) se ha aislado del plasma materno y, cuando se lleva a cabo la obtención de

datos genotípicos de la madre, estimando los datos genotípicos maternos de las mediciones de ADN realizadas en la muestra preparada.

En un aspecto, se describe un cuadro de diagnóstico para ayudar a determinar un estado de ploidía de un cromosoma en un feto en gestación donde el cuadro de diagnóstico es capaz de ejecutar los pasos de preparación y medición de cualquiera de los métodos de la divulgación.

En varias realizaciones de cualquiera de los aspectos de la divulgación, los recuentos de alelos son probabilísticos en lugar de binarios. En algunas realizaciones, las mediciones del ADN en la muestra preparada en la pluralidad de loci polimórficos también se utilizan para determinar si el feto ha heredado o no uno o una pluralidad de haplotipos vinculados a la enfermedad.

En varias realizaciones de cualquiera de los aspectos de la divulgación, la creación de un modelo de distribución conjunto para las probabilidades del recuento de alelos se realiza utilizando datos sobre la probabilidad de que los cromosomas se crucen en diferentes ubicaciones en un cromosoma para modelar la dependencia entre alelos polimórficos en el cromosoma. En algunas realizaciones, la creación de un modelo de distribución conjunto para los recuentos de alelos y el paso de determinar la probabilidad relativa de cada hipótesis se realiza utilizando un método que no requiere el uso de un cromosoma de referencia.

En varias realizaciones de cualquiera de los aspectos de la divulgación, la determinación de la probabilidad relativa de cada hipótesis hace uso de una fracción estimada de ADN fetal en la muestra preparada. En algunas realizaciones, las mediciones de ADN de la muestra preparada utilizada en el cálculo de las probabilidades del recuento de alelos y la determinación de la probabilidad relativa de cada hipótesis comprende datos genéticos primarios. En algunas realizaciones, la selección del estado de ploidía correspondiente a la hipótesis con la mayor probabilidad se lleva a cabo utilizando las estimaciones de probabilidad máximas o las estimaciones a posteriori máximas.

En varias realizaciones de cualquiera de los aspectos de la divulgación, la determinación del estado de ploidía del feto también incluye la combinación de las probabilidades relativas de cada una de las hipótesis de ploidía determinadas utilizando el modelo de distribución conjunto y las probabilidades del recuento de alelos con las probabilidades relativas de cada una de las hipótesis de ploidía que se calculan utilizando técnicas estadísticas tomadas de un grupo que se compone de un análisis de recuento de lectura, la comparación de tasas de heterocigosidad, una estadística que sólo está disponible cuando se utiliza información genética de los padres, la probabilidad de señales genotípicas normalizadas para determinados contextos de los progenitores, una estadística que se calcula utilizando una fracción fetal estimada de la muestra (por ejemplo, la primera muestra) o la muestra preparada, y combinaciones de estas.

En varias realizaciones de cualquiera de los aspectos de la divulgación, se calcula una estimación de certeza para el estado de ploidía determinado. En algunas realizaciones, el método también incluye tomar una acción clínica en base al estado de ploidía determinado del feto, en donde la acción clínica se selecciona de o bien la finalización del embarazo o mantenerlo.

En varias realizaciones de cualquiera de los aspectos de la divulgación, el método se puede realizar para fetos entre 4 y 5 semanas de gestación; entre 5 y 6 semanas de gestación; entre 6 y 7 semanas de gestación; entre 7 y 8 semanas de gestación; entre 8 y 9 semanas de gestación; entre 9 y 10 semanas de gestación; entre 10 y 12 semanas de gestación; entre 12 y 14 semanas de gestación; entre 14 y 20 semanas de gestación; entre 20 y 40 semanas de gestación; en el primer trimestre; en el segundo trimestre; en el tercer trimestre; o combinaciones de estos.

En varias realizaciones de cualquiera de los aspectos de la divulgación, un informe que muestra un estado de ploidía determinado de un cromosoma en un feto en gestación generado utilizando el método. En algunas realizaciones, se describe un kit para determinar un estado de ploidía de un cromosoma diana en un feto en gestación diseñado para ser utilizado con cualquiera de los métodos de la invención, el kit incluye una pluralidad de cebadores directos interiores y opcionalmente la pluralidad de cebadores inversos interiores, donde cada uno de los cebadores está diseñado para que se hibride en la región de ADN inmediatamente corriente arriba y/o corriente abajo de uno de los lugares polimórficos en el cromosoma diana, y opcionalmente en cromosomas adicionales, donde la región de hibridación está separada del lugar polimórfico por un número de bases pequeño, en el que el número pequeño se selecciona del grupo compuesto de 1, 2, 3, 4, 5, 6 a 10, 11 a 15, 16 a 20, 21 a 25, 26 a 30, 31 a 60, y combinaciones de estos.

En un aspecto, la divulgación presenta métodos para establecer si un presunto padre es el padre biológico de un feto que se está gestando en una madre embarazada. En algunas realizaciones, el método consiste en (i) amplificar simultáneamente una pluralidad de loci polimórficos que incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes en material genético del presunto padre para producir un primer conjunto de productos amplificados; (ii) amplificar simultáneamente la pluralidad de loci polimórficos correspondiente en una muestra mezclada de ADN que se origina a partir de una muestra de sangre de la madre embarazada para producir un segundo conjunto de productos amplificados; en donde la muestra mezclada de ADN incluye ADN fetal y ADN materno; (iii) determinar en un ordenador la probabilidad de que el presunto padre sea el padre biológico del feto utilizando mediciones genotípicas en base al primer y segundo conjuntos de productos amplificados; y (iv) establecer si el presunto padre es el padre biológico del feto utilizando la probabilidad determinada de que el presunto padre es el padre biológico del feto. En varias realizaciones, el método incluye además amplificar simultáneamente la pluralidad de loci polimórficos correspondiente en el material genético de la madre para producir un tercer conjunto de productos amplificados; en donde la probabilidad de que el presunto padre sea el padre biológico del

feto se determina utilizando mediciones genotípicas en base al primer, segundo y tercer conjuntos de productos amplificados.

En un aspecto, la divulgación proporciona métodos de estimar las probabilidades relativas de que cada embrión de un conjunto de embriones se desarrollará como se desea. En algunas realizaciones, el método consiste en poner en contacto una muestra de cada embrión con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes para producir una mezcla de reacción para cada embrión, en donde cada una de las muestras se deriva de una o más células de un embrión. En algunas realizaciones, cada mezcla de reacción se somete a condiciones de reacción de extensión del cebador para producir productos amplificados. En algunas realizaciones, el método incluye determinar en un ordenador una o más características de al menos una célula de cada embrión en base a los productos amplificados; y estimar en un ordenador las probabilidades relativas de que cada embrión se desarrollará según se desee, en base a una o más características de al menos una célula para cada embrión.

En un aspecto, la divulgación presenta métodos de medir la cantidad de dos o más loci diana en una muestra de ácido nucleico. En algunas realizaciones, el método implica (i) utilizar una PCR para amplificar una muestra de ácido nucleico que incluye un primer locus estándar, un segundo locus estándar, un primer locus diana, y un segundo locus diana para formar productos amplificados; donde el primer locus estándar y el primer locus diana tienen el mismo número de nucleótidos pero tienen una secuencia que difiere en uno o más nucleótidos; y donde el segundo locus estándar y el segundo locus diana tienen el mismo número de nucleótidos pero tienen una secuencia que difiere en uno o más nucleótidos; (ii) secuenciar los productos amplificados para determinar una relación estándar que compare la cantidad relativa del primer locus estándar amplificado en comparación con el segundo locus estándar amplificado; en donde una relación estándar indica la diferencia en la eficiencia de la PCR para la amplificación del primer locus estándar y el segundo locus estándar; (iii) determinar una relación diana que compare la cantidad relativa del primer locus diana amplificado en comparación con el segundo locus diana amplificado; y (iv) ajustar la relación diana del paso (iii) en base a la relación estándar del paso (ii) para determinar la cantidad relativa del primer locus diana y el segundo locus diana en la muestra. En varias realizaciones, el método implica determinar la cantidad absoluta del primer locus diana y el segundo locus diana en la muestra. En varias realizaciones, el método incluye además determinar la presencia o ausencia de un locus diana (por ejemplo, al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes en la muestra. En varias realizaciones, el método implica utilizar cualquiera de las bibliotecas de cebadores de la invención. En varias realizaciones, el método implica amplificar simultáneamente 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes.

En un aspecto, la divulgación presenta métodos de medir cuantitativamente una pluralidad de dianas genéticas en una muestra para análisis. En algunas realizaciones, el método incluye (i) mezclar material genético derivado de la muestra para análisis con una pluralidad de reactivos de amplificación específicos diana, y una pluralidad de secuencias estándar correspondientes a las dianas de los reactivos de amplificación específicos diana; (ii) amplificar zonas diana del material genético y las secuencias estándar para producir amplicones diana y amplicones de secuencia estándar; y (iii) medir la cantidad de amplicones diana y amplicones de secuencia estándar producidos. En algunas realizaciones, el material genético está presente en una biblioteca genética. En algunas realizaciones, las dianas genéticas son loci polimórficos (como, por ejemplo, SNP). En algunas realizaciones, la medición de la cantidad se consigue contando las secuencias. En algunas realizaciones, el método incluye además la determinación del número de copias estimada de al menos un cromosoma en una muestra del que la biblioteca genética se ha derivado, en donde la determinación implica comparar el número de lecturas de secuencia de un amplicón diana con el número de lecturas de secuencia de un amplicón estándar. En algunas realizaciones, las secuencias estándar y la biblioteca genética

incluye puntos de cebado universal capaces de ser cebados por el mismo cebador. En algunas realizaciones, el paso de mezcla incluye al menos 10; 100, 500; 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 reactivos de amplificación específicos diana diferentes y al menos 10; 100, 500; 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 secuencias estándar. En varias realizaciones, el método implica utilizar cualquiera de las bibliotecas de cebadores de la invención. En varias realizaciones, el método implica amplificar simultáneamente 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 zonas diana diferentes. En algunas realizaciones, se conocen las cantidades relativas de cada una de las secuencias estándar. En algunas realizaciones, las cantidades relativas de cada una de las secuencias se ha calibrado con respecto a un genoma de referencia. En algunas realizaciones, la muestra para análisis incluye una mezcla de genomas fetales y maternos. En algunas realizaciones, la muestra para análisis se deriva de la sangre de una mujer embarazada o se deriva del plasma sanguíneo. En algunas realizaciones, el genoma de referencia tiene al menos una aneuploidía, como una aneuploidía en el cromosoma 13, 18, 21, X, o Y. En algunas realizaciones, el genoma de referencia es diploide.

En un aspecto, la divulgación presenta una mezcla que incluye una pluralidad de secuencias estándar genéticas, en donde la cantidad relativa de cada secuencia estándar genética en la mezcla se ha determinado mediante la calibración a un genoma de referencia. En varias realizaciones, la mezcla incluye al menos 10; 100, 500; 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 secuencias estándar genéticas. En varias realizaciones, las secuencias estándar genéticas incluyen un primer punto de cebado universal, un segundo punto de cebado universal, un primer punto de cebado específico diana, un segundo punto de cebado específico diana, y una secuencia de marcador situada entre el primer y el segundo puntos de cebado específicos diana, en donde el primer

punto de cebado específico diana y el segundo punto de cebado específico diana están situados entre el primer y el segundo puntos de cebado universal. En varias realizaciones, la calibración implica utilizar cualquiera de las bibliotecas de cebadores de la invención. En varias realizaciones, la calibración implica amplificar simultáneamente 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 zonas diana diferentes. En algunas realizaciones, el genoma de referencia tiene al menos una aneuploidía, como una aneuploidía en el cromosoma 13, 18, 21, X, o Y. En algunas realizaciones, el genoma de referencia es diploide.

En un aspecto, la divulgación presenta métodos de producir un conjunto de secuencias estándar genéticas calibradas. En algunas realizaciones, el método incluye (i) formar una mezcla de reacción de amplificación que incluye una biblioteca genética preparada a partir de una genoma de referencia, una pluralidad de conjuntos de reactivos del cebador específicos diana, y una pluralidad de secuencias estándar genéticas correspondientes a los conjuntos de reactivos de amplificación específicos diana, (ii) amplificar la biblioteca genética y las secuencias estándar genéticas para producir amplicones de las secuencias diana y amplicones de las secuencias estándar genéticas, (iii) medir la cantidad de los amplicones de las secuencias diana y los amplicones de las secuencias estándar genéticas, y (iv) determinar la cantidad relativa de cada una de las secuencias estándar genéticas con respecto a la otra, por la cual se calibra la pluralidad de secuencias estándar genéticas. En varias realizaciones, se utilizan al menos 10; 100; 500; 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 secuencias estándar genéticas. En varias realizaciones, el método implica utilizar cualquiera de las bibliotecas de cebadores de la invención. En varias realizaciones, el método implica amplificar simultáneamente 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o secuencias diferentes. En algunas realizaciones, el genoma de referencia tiene al menos una aneuploidía, como una aneuploidía en el cromosoma 13, 18, 21, X, o Y. En algunas realizaciones, el genoma de referencia es diploide.

En un aspecto, la divulgación proporciona un conjunto de secuencias estándar genéticas que han sido calibradas según cualquiera de los métodos de la invención. En un aspecto, la invención proporciona un conjunto de secuencias estándar genéticas que se han podido calibrar antes, durante o después de que se haya llevado a cabo el método.

En un aspecto, la divulgación presenta métodos de medir el número de copias de un gen de interés que tiene al menos un alelo que tiene una eliminación. En algunas realizaciones, el método incluye (i) mezclar material genético derivado a partir de una muestra para análisis con un reactivo de amplificación específico del gen de interés y que no puede amplificar significativamente la eliminación que comprende el alelo del gen de interés, una secuencia estándar correspondiente al gen de interés, un reactivo de amplificación específico para una secuencia de referencia, y una secuencia estándar correspondiente a la secuencia de referencia; (ii) amplificar la secuencia de genes de interés, la secuencia estándar correspondiente al gen de interés, la secuencia de referencia, y la secuencia estándar correspondiente a la secuencia de referencia para producir genes de los amplicones de interés, los amplicones de secuencia de referencia, y los amplicones de secuencia estándar; y (iii) medir la cantidad de amplicones diana y amplicones de secuencia estándar producidos. En algunas realizaciones, la medición de la cantidad se consigue contando las lecturas de secuencia. En algunas realizaciones, el método incluye además la determinación del número de copias estimada de al menos un cromosoma en una muestra del que la biblioteca genética se ha derivado, en donde la determinación implica comparar el número de secuencias de amplicones diana con el número de secuencias de amplicones estándar. En algunas realizaciones, las secuencias estándar y la biblioteca genética incluyen puntos de cebado universal capaces de ser cebados por el mismo cebador. En algunas realizaciones, las cantidades relativas de cada una de las secuencias se han calibrado con respecto a un genoma de referencia. En varias realizaciones, se utilizan al menos 10; 100; 500; 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 secuencias estándar genéticas. En varias realizaciones, el método implica utilizar cualquiera de las bibliotecas de cebadores de la invención. En varias realizaciones, el método implica amplificar simultáneamente 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 zonas diana diferentes. En algunas realizaciones, el genoma de referencia es diploide. En algunas realizaciones, la muestra para análisis está derivada de la sangre.

En algunas realizaciones de cualquiera de los aspectos de la divulgación, que enriquecen preferentemente el ADN en la muestra (por ejemplo, la primera muestra) en los loci diana (por ejemplo, la pluralidad de loci polimórficos) incluye la obtención de una pluralidad de sondas precircularizadas donde cada sonda está dirigida a uno de los loci (por ejemplo, loci polimórficos), donde el extremo 3' y 5' de las sondas está diseñado preferiblemente para que se hibride en una región de ADN que está separada del punto polimórfico del locus por un pequeño número de bases, donde el pequeño número es 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 a 25, 26 a 30, 31 a 60, o una combinación de estos, hibridando las sondas precircularizadas al ADN de la muestra (por ejemplo, la primera muestra), que llena el hueco entre los extremos de la sonda hibridada utilizando ADN polimerasa, circularizando la sonda precircularizada, y amplificando la sonda circularizada.

En algunas realizaciones de cualquiera de los aspectos de la divulgación, que enriquecen preferentemente el ADN en los loci diana (por ejemplo., la pluralidad de loci polimórficos) incluye la obtención de una pluralidad de sondas para PCR mediadas por unión donde cada sonda para PCR focaliza uno de los loci diana (por ejemplo, los loci polimórficos), y donde las sondas para PCR corriente arriba y corriente abajo están diseñadas para que se hibriden en una región de ADN en una hebra de ADN que está separada preferiblemente del punto polimórfico del locus por un pequeño número de bases, donde el pequeño número es 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 a 25, 26 a 30, 31 a 60, o una combinación de estos, hibridando las sondas para PCR mediadas por unión al ADN de la muestra (por ejemplo, la primera muestra), que llena el hueco entre los extremos de la sonda para PCR mediada por unión

utilizando ADN polimerasa, ligando las sondas para PCR mediadas por unión, y amplificando las sondas para PCR mediadas por unión.

En algunas realizaciones de varios aspectos de la divulgación, que enriquece preferentemente el ADN en los loci diana (por ejemplo, la pluralidad de loci polimórficos) incluye la obtención de una pluralidad de sondas de captura híbridas que focalizan los loci (por ejemplo, los loci polimórficos), hibridando las sondas de captura híbridas en el ADN en la muestra (por ejemplo, la primera muestra) y eliminan físicamente algunos o todo el ADN no hibridado de la muestra (por ejemplo, la primera muestra) de ADN.

En algunas realizaciones de cualquiera de los aspectos de la divulgación, las sondas de captura híbridas están diseñadas para hibridarse en una región que flanquea pero no se solapa con el punto polimórfico. En algunas realizaciones, las sondas de captura híbridas están diseñadas para hibridarse con una región que flanquea pero no se solapa con el punto polimórfico, y donde la longitud de la sonda de captura de flanco se puede seleccionar del grupo compuesto por menos de unas 120 bases, menos de unas 110 bases, menos de unas 100 bases, menos de unas 90 bases, menos de unas 80 bases, menos de unas 70 bases, menos de unas 60 bases, menos de unas 50 bases, menos de unas 40 bases, menos de unas 30 bases, y menos de unas 25 bases. En algunas realizaciones, las sondas de captura híbrida están diseñadas para que se hibriden en una región que se solapa con el punto polimórfico, y donde la pluralidad de sondas de captura híbrida comprende al menos dos sondas de captura híbrida para cada loci polimórfico, y donde cada sonda de captura híbrida está diseñada para ser complementaria de un alelo diferente en el locus polimórfico.

En algunas realizaciones de cualquiera de los aspectos de la divulgación, el enriquecimiento preferente del ADN en una pluralidad de loci polimórficos incluye la obtención de una pluralidad de cebadores directos internos donde cada cebador se dirige a uno de los loci polimórficos, y donde el extremo 3' de los cebadores directos internos está diseñado para hibridarse con una región de ADN corriente arriba del punto polimórfico, y separada del punto polimórfico por un pequeño número de bases, donde el pequeño número de bases se selecciona del grupo compuesto por 1, 2, 3, 4, 5, 6 a 10, 11 a 15, 16 a 20, 21 a 25, 26 a 30, o 31 a 60 pares de bases, opcionalmente obteniendo una pluralidad de cebadores inversos internos donde cada cebador se dirige a uno de los loci polimórficos, y donde el extremo 3' de los cebadores inversos internos está diseñado parahibridarse con una región de ADN corriente arriba del sitio polimórfico, y separada del sitio polimórfico por un pequeño número de bases, donde el pequeño número de bases se selecciona del grupo compuesto por 1, 2, 3, 4, 5, 6 a 10, 11 a 15, 16 a 20, 21 a 25, 26 a 30, o 31 a 60 pares de bases, hibridando los cebadores internos con el ADN, y amplificando el ADN utilizando la reacción en cadena de la polimerasa para formar amplicones.

En algunas realizaciones de cualquiera de los aspectos de la divulgación, el método también incluye la obtención de una pluralidad de cebadores directos exteriores en donde cada cebador focaliza uno de los loci diana (por ejemplo, loci polimórficos), y donde los cebadores directos exteriores están diseñados para que se hibriden en la región de ADN corriente arriba del cebador directo interior, obteniendo opcionalmente una pluralidad de cebadores inversos exteriores donde cada cebador focaliza uno de los loci diana (por ejemplo, loci polimórficos), y donde los cebadores inversos exteriores están diseñados para hibridarse en la región de ADN inmediatamente corriente abajo del cebador inverso interior, hibridando los primeros cebadores en el ADN, y amplificando el ADN utilizando la reacción en cadena de polimerasa.

En algunas realizaciones de cualquiera de los aspectos de la divulgación, el método también incluye la obtención de una pluralidad de cebadores inversos exteriores donde cada cebador focaliza uno de los loci polimórficos, y donde los cebadores inversos exteriores están diseñados para que se hibriden en la región de ADN inmediatamente corriente abajo del cebador inverso interior, la obtención opcionalmente de una pluralidad de cebadores directos exteriores donde cada cebador focaliza uno de los loci diana (por ejemplo, los loci polimórficos), y donde los cebadores directos exteriores están diseñados para hibridarse en la región de ADN corriente arriba del cebador directo interior, hibridando los primeros cebadores en el ADN, y amplificando el ADN utilizando la reacción en cadena de polimerasa.

En algunas realizaciones de cualquiera de los aspectos de la divulgación, la preparación de la muestra (por ejemplo, la primera muestra) incluye además anexar adaptadores universales al ADN en la muestra (por ejemplo, la primera muestra) y amplificar el ADN en la muestra (por ejemplo, la primera muestra) utilizando la reacción en cadena de polimerasa. En algunas realizaciones, al menos una fracción de los amplicones que se amplifican es inferior a 100 bp, inferior a 90 bp, inferior a 80 bp, inferior a 70 bp, inferior a 65 bp, inferior a 60 bp, inferior a 55 bp, inferior a 50 bp, o inferior a 45 bp, y donde la fracción es un 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, o 99%.

En algunas realizaciones de cualquiera de los aspectos de la invención, la amplificación del ADN se realiza en uno o en varios volúmenes de reacción individual, y en donde cada volumen de reacción individual contiene más de 100 pares de cebador directos e inversos diferentes, más de 200 pares de cebador directos e inversos diferentes, más de 500 pares de cebador directos e inversos diferentes, más de 1.000 pares de cebador directos e inversos diferentes, más de 2.000 pares de cebador directos e inversos diferentes, más de 5.000 pares de cebador directos e inversos diferentes, más de 10.000 pares de cebador directos e inversos diferentes, más de 20.000 pares de cebador directos e inversos diferentes, más de 50.000 pares de cebador directos e inversos diferentes, o más de 100.000 pares de cebador directos e inversos diferentes.

En algunas realizaciones de cualquiera de los aspectos de la invención, la preparación de la muestra (por ejemplo, la primera muestra) comprende además dividir la muestra (por ejemplo, la primera muestra) en una pluralidad de porciones, y donde el ADN en cada porción está enriquecido preferentemente en un subconjunto de los loci diana (por ejemplo, una pluralidad de loci polimórficos). En algunas realizaciones, los cebadores internos se seleccionan identificando pares de cebadores que es probable que formen dúplex de cebadores no deseados y eliminando de la pluralidad de cebadores al menos uno de los pares de cebadores identificados por su probabilidad de formar dúplex de cebadores no deseados. En algunas realizaciones, los cebadores interiores contienen una región que está diseñada para que se hibride o bien corriente arriba o corriente abajo del locus focalizado (por ejemplo, el locus polimórfico), y opcionalmente contienen una secuencia de cebado universal diseñada para permitir la amplificación por PCR. En algunas realizaciones, al menos algunos de los cebadores contienen además una región aleatoria que difiere para cada molécula del cebador. En algunas realizaciones, al menos algunos de los cebadores contienen además un código de barras molecular.

En algunas realizaciones de cualquiera de los aspectos de la divulgación, el enriquecimiento preferente resulta en un grado medio de sesgo alélico entre la muestra preparada y la muestra (por ejemplo, la primera muestra) de un factor seleccionado del grupo que se compone de no más de un factor de 2, no más de un factor de 1,5, no más de un factor de 1,2, no más de un factor de 1,1, no más de un factor de 1,05, no más de un factor de 1,02, no más de un factor de 1,01, no más de un factor de 1,005, no más de un factor de 1,002, no más de un factor de 1,001 y no más de un factor de 1,0001. En algunas realizaciones, la pluralidad de loci polimórficos son SNP. En algunas realizaciones, la medición del ADN en la muestra preparada se realiza mediante secuenciación.

En algunas realizaciones de cualquiera de los aspectos de la invención, los loci diana están presentes en el mismo ácido nucleico de interés (por ejemplo, el mismo cromosoma o la misma región de un cromosoma). En algunas realizaciones, al menos algunos de los loci diana están presentes en diferentes ácidos nucleicos de interés (por ejemplo, diferentes cromosomas). En algunas realizaciones, la muestra de ácido nucleico incluye ácidos nucleicos fragmentados o digeridos. En algunas realizaciones, la muestra de ácido nucleico incluye ADN, cDNA, o mRNA genómico. En algunas realizaciones, la muestra de ácido nucleico incluye ADN de una sola célula. En algunas realizaciones, la muestra de ácido nucleico es una muestra de sangre o plasma que está prácticamente libre de células. En algunas realizaciones, la muestra de ácido nucleico incluye o está derivada de sangre, plasma, saliva, semen, esperma, cultivo celular flotante, secreción mucosa, placa dental, tejido del tracto gastrointestinal, heces, orina, pelo, hueso, fluidos corporales, lágrimas, tejido, piel, uñas, blastómeros, embriones, líquido amniótico, muestras de vellosidades coriónicas, bilis, linfa, mucosidad cervical, o una muestra forense. En algunas realizaciones, los loci diana son segmentos de ácidos nucleicos humanos. En algunas realizaciones, los loci diana comprenden o se componen de polimorfismos de un solo nucleótido (SNP). En algunas realizaciones, los cebadores son moléculas de ADN.

En algunas realizaciones de cualquiera de los aspectos de la invención, el ADN en la muestra (por ejemplo, la primera muestra) se origina a partir de plasma materno. En algunas realizaciones, la preparación de la muestra (por ejemplo, la primera muestra) comprende además la amplificación del ADN. En algunas realizaciones, la preparación de la muestra (por ejemplo, la primera muestra) comprende además enriquecer preferentemente el ADN en la muestra (por ejemplo, la primera muestra) en los loci diana (por ejemplo, una pluralidad de loci polimórficos).

En varias realizaciones, la reacción de extensión del cebador o la reacción en cadena de la polimerasa incluye la adición de uno o más nucleótidos mediante una polimerasa. En varias realizaciones, la reacción de extensión del cebador o la reacción en cadena de la polimerasa no incluye PCR mediado por unión. En varias realizaciones, la reacción de extensión del cebador o la reacción en cadena de la polimerasa no incluye la unión de dos cebadores por una ligasa. En varias realizaciones, los cebadores no incluyen sondas invertidas vinculadas (LIP), que también se pueden denominar sondas precircularizadas, sondas precircularizantes, sondas circularizantes, sondas candado, o sondas de inversión molecular (MIP).

Se entiende que los aspectos y realizaciones de la invención que se describen en el presente documento incluyen aspectos y realizaciones "que comprenden," "que se componen," y "que se componen básicamente de" (fórmula).

Definiciones

Polimorfismo de un solo nucleótido (SNP) hace referencia a un solo nucleótido que puede diferir entre los genomas de dos miembros de la misma especie. El empleo del término no debe implicar ninguna limitación en la frecuencia con que se produce cada variante.

Secuencia puede referirse a una secuencia de ADN o una secuencia genética. Puede referirse a la estructura física, primaria de la molécula o cadena de ADN en un individuo. Puede referirse a la secuencia de nucleótidos que se encuentra en esa molécula de ADN o a la cadena complementaria de la molécula de ADN. Puede referirse a la información contenida en la molécula de ADN como su representación in silico.

Locus se refiere a una región determinada de interés en el ADN de un individuo, que puede referirse a un SNP, el punto de una posible inserción o delección, o el punto de otra variación genética relevante. Los SNP vinculados a enfermedad pueden referirse también a loci vinculados a enfermedad. Alelo polimórfico, también "locus polimórfico," se refiere a un alelo o locus en el que el genotipo varía entre individuos de una determinada especie. Algunos ejemplos de alelos polimórficos incluyen polimorfismos de un solo nucleótido, repeticiones en tándem cortas, eliminaciones, duplicaciones e inversiones.

Punto polimórfico hace referencia a los nucleótidos específicos encontrados en una región polimórfica que varía entre individuos.

Alelo

Datos genéticos, también "datos genotípicos", se refiere a los datos que describen aspectos del genoma de uno o más individuos. Pueden referirse a uno o un conjunto de loci, secuencias parciales o completas, cromosomas parciales o completos o al genoma completo. Se puede referir a la identidad de uno o una pluralidad de nucleótidos; puede referirse a un conjunto de nucleótidos secuenciales o nucleótidos de diferentes ubicaciones en el genoma, o una combinación de estos.

Los datos genotípicos se utilizan típicamente in silico; sin embargo, también es posible considerar nucleótidos físicos en una secuencia como datos genéticos químicamente codificados. Se puede decir que los datos genotípicos son "sobre", "de" o "procedentes del" individuo o los individuos. Los datos genotípicos se pueden referir a medidas resultantes de una plataforma de determinación genotípica cuyas mediciones se realizan con material genético.

Material genético también "Muestra genética," se refiere a materia física, como tejido o sangre, de uno o más individuos que comprenden datos genéticos ruidosos de ADN o ARN, se refiere a datos genéticos con cualquiera de los siguientes elementos: pérdidas de alelos, mediciones inciertas de pares de bases, mediciones incorrectas de pares de bases, mediciones ausentes de pares de bases,

mediciones inciertas de inserciones o deleciones,

mediciones inciertas de número de copias de segmentos de cromosomas, señales falsas, mediciones ausentes, otros errores, o combinaciones de estos.

Certeza se refiere a la probabilidad estadística que el SNP, el alelo, el conjunto de alelos determinados, o el número determinado de copias de segmento de cromosoma represente correctamente el estado genético real del individuo.

Determinación de ploidía, también "Determinación del Número de Copia de Cromosoma" o "Determinación del Número de Copia" (CNC), puede hacer referencia al acto de determinar la cantidad y/o identidad cromosómica de uno o más cromosomas presentes en una célula.

Aneuploidía se refiere al estado en que están presentes en una célula un número erróneo de cromosomas (por ejemplo, el número erróneo de cromosomas completos o el número erróneo de segmentos de cromosomas como, por ejemplo, la presencia de deleciones o duplicaciones de un segmento de cromosoma). En el caso de una célula somática humana, puede referirse al caso de que una célula no contenga 22 pares de cromosomas autosómicos y un par de cromosomas sexuales. En un gameto humano, puede referirse al caso de que una célula no contenga uno de cada uno de los 23 cromosomas. En el caso de un tipo de cromosoma individual, se puede referir al caso donde más o menos de dos copias de cromosomas homólogos pero no idénticos están presentes, o donde hay dos copias de cromosomas presentes que se originan del mismo progenitor. En algunas realizaciones, la eliminación de un segmento de cromosoma es una microeliminación.

Estado de ploidía se refiere a la cantidad e identidad cromosómica de uno o más tipos de cromosomas en una célula.

Cromosoma se puede referir a una única copia del cromosoma, lo que significa una única molécula de ADN de las que hay 46 en una célula somática normal; un ejemplo es el "cromosoma de origen materno 18". Cromosoma también se puede referir al tipo de cromosoma, del que hay 23 en una célula somática humana normal; un ejemplo es "cromosoma 18".

Identidad cromosómica puede referirse al número de cromosomas referente, es decir, el tipo de cromosoma. Los humanos normales tienen 22 tipos de cromosomas autosómicos numerados, y dos tipos de cromosomas sexuales. Puede referirse también al origen parental del cromosoma. Puede también referirse a un cromosoma específico heredado del progenitor. También puede referirse a otras características identificadoras de un cromosoma.

El estado del material genético o simplemente "estado genético" puede referirse a la identidad de un conjunto de SNP en el ADN, puede referirse a los haplotipos por fases del material genético, y puede referirse a la secuencia del ADN, incluyendo inserciones, deleciones, repeticiones y mutaciones. También puede referirse al estado de ploidía de uno o más cromosomas, segmentos cromosómicos o conjunto de segmentos cromosómicos.

Datos alélicos se refiere a un conjunto de datos genotípicos respecto a un conjunto de uno o más alelos. Puede referirse a los datos haplotípicos por fases. Puede referirse a identidades de SNP, y puede referirse a los datos de secuencia del ADN, incluyendo inserciones, deleciones, repeticiones y mutaciones. Puede incluir el origen parental de cada alelo.

Estado alélico se refiere al estado real de los genes en un conjunto de uno o más alelos. Puede referirse al estado real de los genes descrito por los datos alélicos.

Ratio alélico o ratio de alelos, se refiere al ratio entre la cantidad de cada alelo en un locus que se encuentra presente en una muestra o en un individuo. Cuando la muestra se ha medido mediante secuenciación, el ratio alélico se puede referir al ratio de lecturas de secuencias que corresponden a cada alelo del locus. Cuando la muestra se ha medido por un método de medición basado en la intensidad, el ratio del alelo se puede referir al ratio de las cantidades de cada alelo presente en ese locus estimado por el método de medición.

Recuento de alelos se refiere al número de secuencias que corresponden a un determinado locus y, si ese locus es polimórfico, se refiere al número de secuencias que corresponden a cada uno de los alelos. Si cada alelo se somete a conteo de forma binaria, entonces el recuento de alelos será el número entero. Si los alelos se someten a conteo probabilístico, entonces el recuento de alelos podrá ser una fracción.

Probabilidad del recuento de alelos se refiere al número de secuencias que es probable que correspondan a un locus determinado o a un conjunto de alelos en un locus polimórfico, combinado con la probabilidad de la correspondencia. Cabe señalar que los recuentos de alelos son equivalentes a las probabilidades del recuento de alelos, donde la probabilidad de correspondencia para cada secuencia recontada es binaria (cero o uno). En algunas realizaciones, las probabilidades del recuento de alelos pueden ser binarias. En algunas realizaciones, las probabilidades del recuento de alelos puede establecerse que sean iguales a las mediciones de ADN.

Distribución alélica, o 'distribución de recuento de alelos' hace referencia a la cantidad relativa de cada alelo que está presente para cada locus en un conjunto de loci. Una distribución alélica puede hacer referencia a un individuo, a una muestra, o a un conjunto de mediciones realizada en una muestra. En el contexto de la secuenciación, la distribución alélica hace referencia al número o número probable de lecturas que corresponden a un alelo determinado para cada alelo en un conjunto de loci polimórficos. Las mediciones de alelos se pueden tratar probabilísticamente, es decir, la probabilidad de que un alelo determinado esté presente para una lectura de secuencia determinada es una fracción entre 0 y 1, o se pueden tratar de un modo binario, es decir, cualquier lectura determinada se considera que es exactamente cero o una copia de un alelo determinado.

Patrón de distribución alélica se refiere a un conjunto de distribuciones alélicas diferentes para distintos contextos parentales. Determinados patrones de distribución alélica pueden ser indicativos de determinados estados de ploidía.

Sesgo alélico se refiere al grado en el que el ratio medido de alelos de un locus heterocigoto difiere del ratio que estaba presente en la muestra original de ADN. El grado de sesgo alélico en un determinado locus es igual al ratio alélico observado en ese locus, medido, dividido por el ratio de alelos de la muestra de ADN original en ese locus. El sesgo alélico se puede definir para que sea superior a uno, de forma que si el cálculo del grado de sesgo alélico da un valor, x , que es inferior a 1, entonces el grado de sesgo alélico se puede reajustar como $1/x$. El sesgo alélico se puede deber a un sesgo de amplificación, un sesgo de purificación o algún otro fenómeno que afecta a distintos alelos de forma diferente.

Cebador, también "sonda para PCR" hace referencia a una sola molécula de ADN (un oligómero de ADN) o una colección de moléculas de ADN (oligómeros de ADN) donde las moléculas de ADN son idénticas, o casi idénticas, y donde el cebador contiene una región que está diseñada para que se hibride a un locus focalizado (por ejemplo, un locus polimórfico focalizado o un locus no polimórfico), y puede contener una secuencia de cebado diseñada para permitir la amplificación por PCR. Un cebador también puede contener un código de barras molecular. Un cebador puede contener una región aleatoria que difiere para cada molécula individual. Los términos "cebador de prueba" y "cebador candidato" no son limitativos y pueden referirse a cualquiera de los cebadores que se describen en el presente documento.

Biblioteca de cebadores se refiere a una población de dos o más cebadores. En varias realizaciones, la biblioteca incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 cebadores diferentes. En varias realizaciones, la biblioteca incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 pares de cebadores diferentes, en donde cada par de cebadores incluye un cebador de prueba directo y un cebador de prueba inverso en el que cada par de cebadores de prueba se hibrida en un locus diana. En algunas realizaciones, la biblioteca de cebadores incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 cebadores individuales diferentes que se hibridan cada uno de ellos a diferentes locus diana, en donde los cebadores individuales no forman parte de los pares de cebador. En algunas realizaciones, la biblioteca tiene tanto (i) pares de cebadores como (ii) cebadores individuales (como cebadores universales) que no forman parte de los pares de cebadores.

Sonda de captura híbrida hace referencia a cualquier secuencia de ácido nucleico, posiblemente modificada, que se genera mediante diversos métodos como PCR o síntesis directa y destinada a ser complementaria de una hebra de una secuencia de ADN diana específica en una muestra. Las sondas de captura híbridas exógenas se pueden añadir a una muestra preparada e hibridada a través de un proceso de desnaturalización-rehibridación para formar dúplex de fragmentos exógenos-endógenos. Posteriormente, estos dúplex se pueden separar físicamente de la muestra a través de diversos medios.

Lectura de secuencia se refiere a los datos que representan una secuencia de bases de nucleótidos que han sido medidas utilizando un método de secuenciación clónico. La secuenciación clónica puede producir datos de la secuencia que representan una única molécula o clones o agrupaciones de una molécula de ADN original. Una lectura de secuencia también puede tener una puntuación de calidad asociada en cada posición de base de la secuencia indicando la probabilidad de que el nucleótido ha sido determinado correctamente.

Establecer la correspondencia de una lectura de secuencia es el proceso de determinar la ubicación de origen de la lectura de secuencia en la secuencia del genoma de un organismo concreto. La ubicación de origen de las lecturas de secuencia se basa en la similaridad de la secuencia nucleotídica de la lectura y la secuencia del genoma.

Error de copia emparejada, también “aneuploidía de cromosoma emparejado”, o “MCA”, se refiere a un estado de aneuploidía donde una célula contiene dos cromosomas idénticos o casi idénticos. Este tipo de aneuploidía puede surgir durante la formación de los gametos en la meiosis, y puede ser denominada error de no disyunción meiótica. Este tipo de error se puede producir en la mitosis. La trisomía emparejada se puede referir al caso en el que tres copias de un determinado cromosoma se encuentran presentes en un individuo y dos de las copias son idénticas.

Error de copia no emparejada, también “Aneuploidía de Cromosoma Único” o “UCA”, se refiere a un estado de aneuploidía en el que una célula contiene dos cromosomas que proceden del mismo progenitor, y que pueden ser homólogos pero no idénticos. Este tipo de aneuploidía puede surgir durante la meiosis, y puede ser denominado error meiótico. La trisomía no emparejada se puede referir al caso en el que tres copias de un determinado cromosoma se encuentran presentes en un individuo y dos de las copias son del mismo progenitor, y son homólogas, pero no idénticas. Cabe señalar que la trisomía no emparejada se puede referir al caso en el que dos cromosomas homólogos de un progenitor están presentes, y donde algunos segmentos de los cromosomas son idénticos mientras que otros segmentos son simplemente homólogos.

Cromosomas homólogos se refiere a copias de cromosomas que contienen el mismo conjunto de genes que normalmente se emparejan durante la meiosis.

Cromosomas idénticos se refiere a copias de cromosomas que contienen el mismo conjunto de genes, y para cada gen tienen el mismo conjunto de alelos que son idénticos, o casi idénticos.

Pérdida de alelos (Allele Drop Out, ADO) se refiere a la situación en la que al menos no se detecta uno de los pares de bases en un conjunto de pares de bases de cromosomas homólogos en un alelo determinado.

Pérdida de locus (Locus Drop Out, LDO) se refiere a la situación en la que no se detectan ambos pares de bases en un conjunto de pares de bases de cromosomas homólogos en un alelo determinado.

Homocigoto se refiere a tener alelos similares en loci cromosómicos correspondientes.

Heterocigoto puede referirse a tener alelos distintos en loci cromosómicos correspondientes.

Tasa de heterocigosidad hace referencia a la tasa de individuos en la población que tienen alelos heterocigóticos en un locus determinado. La tasa de heterocigosidad también se puede referir al ratio previsto o medido de alelos, en un determinado locus de un individuo o una muestra de ADN.

Polimorfismo de un solo nucleótido altamente informativo (HISNP) se refiere a un SNP donde el feto tiene un alelo que no se encuentra presente en el genotipo de la madre.

Región cromosómica se refiere a un segmento de un cromosoma o un cromosoma completo.

Segmento de un cromosoma se refiere a una sección de un cromosoma que en cuanto a tamaño puede ir de un par de bases al cromosoma entero.

Cromosoma se refiere a un cromosoma completo, o también a un segmento o sección de un cromosoma.

Copias se refiere al número de copias de un segmento de cromosoma. Puede referirse a copias idénticas, o puede referirse a copias no idénticas homólogas de un segmento de cromosoma, donde las distintas copias del segmento de cromosoma contienen un conjunto de loci sustancialmente similar, y donde uno o más de los alelos son distintos. Hay que advertir que, en algunos casos de aneuploidía, tales como el error de copia M2, es posible tener algunas copias del segmento de cromosoma determinado que sean idénticas, así como algunas copias del mismo segmento de cromosoma que no sean idénticas.

Haplotipo se refiere a una combinación de alelos en múltiples loci que son típicamente heredados juntos en el mismo cromosoma. El haplotipo puede referirse a solamente dos loci o a un cromosoma completo, dependiendo del número de eventos de recombinación que se han producido entre un conjunto de loci determinado. El haplotipo puede referirse también a un conjunto de polimorfismos de un solo nucleótido (SNP) en una sola cromátida que están asociados estadísticamente.

Datos haplotípicos llamados también “datos por fases” o “datos genéticos ordenados”; puede referirse a datos de un solo cromosoma en un genoma diploide o poliploide; es decir, la copia materna o paterna aislada de un cromosoma en un genoma diploide.

Ajuste por fases se refiere a la acción de determinar los datos genéticos haplotípicos de un individuo concreto no ordenados, datos genéticos diploides (o poliploides). Puede referirse a la acción de determinar cuál de dos genes en un alelo, para un conjunto de alelos hallado en un cromosoma, está asociado con cada uno de los dos cromosomas homólogos en un individuo.

Datos por fases se refiere a los datos genéticos donde se han determinado uno o más haplotipos.

Hipótesis se refiere a un posible estado de ploidía en un conjunto de cromosomas determinado, o un conjunto de estados alélicos posibles en un conjunto de loci determinado. El conjunto de posibilidades puede contener uno o más elementos.

Hipótesis de número de copias, también “hipótesis de estado de ploidía”, se refiere a una hipótesis sobre el número de copias de un cromosoma determinado en un individuo. Puede referirse también a una hipótesis sobre la identidad de cada uno de los cromosomas, incluyendo el progenitor de origen de cada cromosoma, y cuál de los dos cromosomas del progenitor está presente en el individuo. También puede referirse a una hipótesis sobre qué cromosomas, o segmentos de cromosomas, de haberlos, de un individuo relacionado, se corresponden genéticamente con un cromosoma determinado de un individuo

Individuo diana se refiere al individuo cuyo estado genético está siendo determinado. En algunas realizaciones, sólo una cantidad limitada de ADN está disponible para el individuo diana. En algunas realizaciones, el individuo diana es un feto. En algunas realizaciones, puede haber más de un individuo diana. En algunas realizaciones, cada feto que se originó a partir de un par de padres puede considerarse individuos diana. En algunas realizaciones, los datos genéticos que se están determinando son unos de un conjunto de determinaciones de alelos. En algunas realizaciones, los datos genéticos que se están determinando son una determinación de ploidía.

Individuo relacionado hace referencia a cualquier individuo que está relacionado genéticamente y, por este motivo, comparte bloques de haplotipos con el individuo diana. En un contexto, el individuo diana puede ser un progenitor genético del individuo diana, o cualquier material genético derivado de un progenitor, como esperma, un cuerpo polar, un embrión, un feto o un niño. Puede referirse también a un hermano, progenitor o abuelo.

Hermano se refiere a cualquier individuo cuyos padres genéticos sean los mismos que los del individuo en cuestión. En algunas realizaciones, se puede referir a un recién nacido, un embrión, o un feto, o una o más células que se originan a partir de un recién nacido, un embrión, o un feto. Un hermano puede referirse también a un individuo haploide procedente de uno de los progenitores, como esperma, un cuerpo polar, o cualquier otro conjunto de materia genética haplotípica. Un individuo puede ser considerado hermano de sí mismo.

Fetal se refiere “al feto” o “a la región de la placenta que es genéticamente similar al feto”. En una mujer embarazada, alguna porción de la placenta es genéticamente similar al feto, y el ADN fetal flotante libre que se encuentra en la sangre materna puede haberse originado en la porción de la placenta con un genotipo que coincide con el del feto. Cabe señalar que la información genética de la mitad de los cromosomas de un feto se hereda de la madre del feto. En algunas realizaciones, el ADN de estos cromosomas heredados de la madre que procedía de una célula fetal se considera que es “de origen fetal,” no “de origen materno.”

ADN de origen fetal se refiere al ADN que era originalmente parte de la célula cuyo genotipo era esencialmente equivalente al del feto.

ADN de origen materno se refiere al ADN que era originalmente parte de la célula cuyo genotipo era esencialmente equivalente al de la madre.

Niño se refiere a un embrión, un blastómero o un feto. Cabe señalar que en las realizaciones divulgadas en el presente documento, los conceptos descritos se aplican igualmente bien a los individuos que son un niño ya nacido, un feto, un embrión o un conjunto de células de estos. El uso del término niño puede simplemente indicar que el individuo designado como el niño es la descendencia genética de los padres.

Progenitor se refiere a la madre o el padre genéticos de un individuo. Un individuo tiene típicamente dos progenitores, una madre y un padre, aunque esto no tiene que ser necesariamente así, por ejemplo, en el quimerismo cromosómico o genético. Un padre puede ser considerado un individuo.

Contexto parental se refiere al estado genético de un SNP determinado, en cada uno de los dos cromosomas relevantes para cada uno de los dos progenitores de la diana.

Desarrollo según deseado, también “desarrollo normal”, se refiere a un embrión viable implantado en un útero y que resulte en un embarazo y/o a un embarazo que sigue y que resulta en un nacimiento vivo y/o a que el niño nacido carece de anomalías cromosómicas y/o a que el niño nacido carece de otros estados genéticos no deseados, tales como genes vinculados a enfermedad. El término “desarrollo según deseado” comprende todo aquello que puedan desear los padres o el personal sanitario. En algunos casos “desarrollo según deseado” puede referirse a un embrión viable o no viable que resulte útil para la investigación médica u otros fines.

Inserción en un útero se refiere al proceso de transferencia de un embrión a la cavidad uterina, en el contexto de la fertilización in vitro.

Plasma materno se refiere a la porción de plasma de la sangre de una mujer que está embarazada.

Decisión clínica se refiere a cualquier decisión de tomar o no tomar una medida que tiene un resultado que afecta a la salud o a la supervivencia de un individuo. En el contexto del diagnóstico prenatal, una decisión clínica puede referirse a la decisión de abortar o no un feto. Una decisión clínica puede referirse también a la decisión de realizar más pruebas, de emprender acciones para mitigar un fenotipo no deseado o de emprender acciones para prepararse para el nacimiento de un niño con anomalías.

Cuadro de diagnóstico se refiere a una o una combinación de máquinas diseñadas para realizar uno o una pluralidad de aspectos de los métodos divulgados en el presente documento. En una realización, el cuadro de diagnóstico puede estar colocado en un punto de atención al paciente. En una realización, el cuadro de diagnóstico puede realizar una

amplificación focalizada seguida de una secuenciación. En una realización, el cuadro de diagnóstico puede funcionar solo o con la ayuda de un técnico.

Método basado en informática se refiere a un método que confía sustancialmente en la estadística para dar sentido a una gran cantidad de datos. En el contexto del diagnóstico prenatal, se refiere a un método diseñado para determinar el estado de ploidía de uno o más cromosomas, o el estado alélico de uno o más alelos, deduciendo estadísticamente el estado más probable, en lugar de medir directamente de forma física el estado, partiendo de una gran cantidad de datos genéticos, por ejemplo, una secuenciación o array molecular. En una realización de la presente divulgación, la técnica basada en informática puede ser una divulgada en esta patente. En una realización de la presente divulgación puede ser PARENTAL SUPPORT™.

Datos genéticos primarios se refiere a las señales de intensidad análogas que son producidas por una plataforma de determinación del genotipo. En el contexto de matrices de SNP, los datos genéticos primarios hacen referencia a las señales de intensidad antes de que se haya realizado cualquier determinación de genotipo. En el contexto de la secuenciación, los datos genéticos primarios se refieren a las mediciones análogas, análogas al cromatograma, que proceden del secuenciador antes de que se haya determinado la identidad de cualesquiera pares de bases y antes de que la secuencia se haya correlacionado con el genoma.

Datos genéticos secundarios se refiere a los datos genéticos procesados que son producidos por una plataforma de determinación del genotipo. En el contexto de un array del SNP, los datos genéticos secundarios se refieren a las determinaciones del alelo realizadas por el software asociado con el lector de arrays del SNP, donde el software ha realizado una determinación con independencia de que un determinado alelo se encuentre presente o no en la muestra. En el contexto de la secuenciación, los datos genéticos secundarios se refieren a las identidades del par de bases de las secuencias que se han determinado y posiblemente también donde las secuencias se han correlacionado con el genoma.

Diagnóstico prenatal no invasivo (NPD), o también "análisis prenatal no invasivo" (NPS), se refiere a un método para determinar el estado genético de un feto que se está gestando en una madre, utilizando material genético que se encuentra en la sangre de la madre, donde el material genético se obtiene extrayendo sangre intravenosa de la madre.

Enriquecimiento preferente de ADN que corresponde a un locus, o enriquecimiento preferente de ADN en un locus, se refiere a cualquier método que resulta en un aumento del porcentaje de moléculas de ADN en la mezcla de ADN que corresponde al locus tras el enriquecimiento respecto del porcentaje de moléculas de ADN existente en la mezcla de ADN que corresponde al locus antes del enriquecimiento. El método puede implicar la amplificación selectiva de moléculas de ADN que corresponden a un locus. El método puede implicar la eliminación de moléculas de ADN que no corresponden al locus. El método puede implicar una combinación de métodos. El grado de enriquecimiento se define como el porcentaje de moléculas de ADN tras el enriquecimiento de la mezcla que corresponde al locus dividido por el porcentaje de moléculas de ADN antes del enriquecimiento de la mezcla que corresponde al locus. El enriquecimiento preferente se puede realizar en una pluralidad de loci. En algunas realizaciones de la presente divulgación, el grado de enriquecimiento es mayor que 20. En algunas realizaciones de la presente divulgación, el grado de enriquecimiento es mayor que 200. En algunas realizaciones de la presente divulgación, el grado de enriquecimiento es mayor que 2000. Cuando se realiza un enriquecimiento preferente en una pluralidad de loci, el grado de enriquecimiento se puede referir al grado medio de enriquecimiento de todos los loci que componen el conjunto de loci.

Amplificación se refiere a un método que aumenta el número de copias de una molécula de ADN.

Amplificación selectiva se puede referir a un método que aumenta el número de copias de una molécula concreta de ADN o moléculas de ADN que corresponden a una

región concreta del ADN. También se puede referir a un método que aumenta el número de copias de una molécula de ADN focalizada, o una región focalizada de ADN determinada más de lo que aumenta en las moléculas o regiones de ADN no focalizadas. La amplificación selectiva puede ser un método de enriquecimiento preferente.

Secuencia de cebado universal se refiere a una secuencia de ADN que se puede unir a una población de moléculas de ADN diana, por ejemplo, mediante enlace, PCR o PCR mediada por enlace. Una vez añadida a la población de moléculas diana, se pueden utilizar cebadores específicos para las secuencias de cebado universal con el fin de amplificar la población diana utilizando un único par de cebadores de amplificación. Las secuencias de cebado universal típicamente no están relacionadas con las secuencias diana.

Adaptadores universales, o "adaptadores de unión" o "etiquetas de biblioteca" son moléculas de ADN que contienen una secuencia de cebado universal que se puede unir mediante enlace covalente al extremo 5' y 3' de una población de moléculas de ADN de doble cadena diana. La adición de los adaptadores proporciona secuencias de cebado universal a los extremos 5' y 3' de la población diana con la que se puede producir la amplificación por PCR, amplificando todas las moléculas de la población diana con un único par de cebadores de amplificación.

Focalización se refiere a un método utilizado para amplificar de forma selectiva o enriquecer de forma preferente las moléculas de ADN que corresponden a un conjunto de loci en una mezcla de ADN.

Modelo de distribución conjunto se refiere a un modelo que define la probabilidad de eventos definidos en términos de múltiples variables aleatorias, dada una pluralidad de variables aleatorias definidas en el mismo espacio de probabilidad,

donde las probabilidades de la variable están vinculadas. En algunas realizaciones, se puede utilizar el caso degenerado donde las probabilidades de las variables no están vinculadas.

Breve descripción de las ilustraciones

- Las realizaciones descritas en el presente documento se explicarán más detalladamente con referencia a los dibujos adjuntos, en donde se hace referencia a dichas estructuras mediante números a través de diferentes vistas. Los dibujos que se muestran no son necesariamente a escala, sino que se hace hincapié generalmente en ilustrar los principios de las realizaciones que se describen en el presente documento.
- Figura 1: Representación gráfica del método de mini-PCR multiplexada directa.
- Figura 2: Representación gráfica del método de mini-PCR semi-anidada.
- Figura 3: Representación gráfica del método de mini-PCR anidada.
- Figura 4: Representación gráfica del método de mini-PCR hemi-anidada.
- Figura 5: Representación gráfica del método de mini-PCR triplemente hemi-anidada.
- Figura 6: Representación gráfica del método de mini-PCR anidado unilateral.
- Figura 7: Representación gráfica del método de mini-PCR unilateral.
- Figura 8: Representación gráfica del método de mini-PCR semi-anidada inversa.
- Figura 9: Algunos flujos de trabajo posibles de los métodos semi-anidados.
- Figura 10: Representación gráfica de adaptadores de unión de bucle.
- Figura 11: Representación gráfica de cebadores etiquetados internamente.
- Figura 12: Un ejemplo de algunos cebadores con etiquetas internas.
- Figura 13: Representación gráfica de un método utilizando cebadores con una región de unión al adaptador de unión.
- Figura 14: Precisiones simuladas de la determinación del estado de ploidía para el método del recuento con dos técnicas de análisis diferentes.
- Figura 15: Ratio de dos alelos para una pluralidad de SNP de una línea celular del Experimento 4.
- Figura 16: Ratio de dos alelos para una pluralidad de SNP de una línea celular del Experimento 4 clasificadas por cromosoma.
- Figuras 17A-D: Ratio de dos alelos para una pluralidad de SNP en cinco muestras de plasma de cuatro mujeres embarazadas, clasificadas por cromosoma.
- Figura 18: Fracción de datos que se puede explicar por la varianza binomial antes y después de la corrección de datos.
- Figura 19: Gráfico que muestra el enriquecimiento relativo del ADN fetal de las muestras siguiendo un breve protocolo de preparación de bibliotecas.
- Figura 20: Profundidad del gráfico de lectura que compara los métodos de PCR directa y semi-anidada.
- Figura 21: Comparación de la profundidad de lectura para la PCR directa de tres muestras genómicas.
- Figura 22: Comparación de la profundidad de lectura para la mini-PCR semi-anidada de tres muestras.
- Figura 23: Comparación de la profundidad de lectura para reacciones 1200-plex y 9600-plex.
- Figura 24: Ratios del recuento de lectura para seis células de tres cromosomas.
- Figura 25: Ratios de alelos para dos reacciones de tres células y una tercera reacción realizada con 1 ng de ADN genómico en tres cromosomas.
- Figura 26: Ratios de alelos para dos reacciones de una única célula en tres cromosomas.
- Figura 27: Comparación de dos bibliotecas de cebadores que muestran el número de loci con una determinada frecuencia de alelos menor que son focalizados por cada biblioteca de cebadores.
- Figura 28A: Gráfica de la electroforesis de los productos de PCR. Las Figuras 28B-28M son electroferogramas de las líneas 1-12, respectivamente, en la Figura 28A.
- Figuras 29A-29E: La representación en dibujos de un método de la divulgación para la determinación de una aneuploidía fetal (Figura 29A). Los datos del genotipo maternos y paternos (de sangre o hisopos bucales) y los datos de frecuencia de cruce de la base de datos HapMap se utilizan para generar (Figura 29B) múltiples hipótesis independientes para cada estado de ploidía fetal potencial in silico (Figura 29C). Cada una de estas hipótesis se amplía para incluir sub-hipótesis que tienen en cuenta los diferentes puntos de cruce. El modelo de datos predice el aspecto

que tendrán los datos de secuenciación (las distribuciones alélicas esperadas) dado cada genotipo fetal hipotético y a diferentes fracciones ADN libre de células fetales, y se compara con los datos de secuenciación reales; la probabilidad para cada hipótesis se determina utilizando la estadística Bayesiana. En este ejemplo hipotético, se determina la hipótesis con la probabilidad más alta (euploidía) (Figura 29D). Las probabilidades individuales de la Figura 29C se suman para cada familia de hipótesis de número de copia (monosomía, disomía o triploidía). La hipótesis con la máxima probabilidad se denomina estado de ploidía, revela la fracción fetal y representa la precisión calculada específica de la muestra (Figura 29E).

Figuras 30A-30H: Representaciones gráficas típicas de euploidía (Figuras 30A-30C),

monosomía (Figura 30D:), y trisomía (Figuras 30E-30H). Para todos los gráficos, el eje x representa la posición lineal de los loci polimórficos individuales a lo largo de cada cromosoma (como se indica debajo de los gráficos), y el eje y representa el número de lecturas de un alelo A como una fracción del total de lecturas de alelos (A+B). Los genotipos maternos y fetales, además de la posición del eje y alrededor del cual están centradas las bandas, se indican a la derecha de los gráficos. Si se desea para facilitar la visualización, los gráficos pueden estar codificados por colores según el genotipo materno, de modo que el rojo indique un genotipo materno de AA, el azul indique un genotipo materno de BB, y el verde indique un genotipo materno de AB. Si se desea, las contribuciones de alelo materno se pueden indicar en color en la columna "Genotipo fetal". Las contribuciones de alelo se indican como maternas/fetales, de modo que los alelos para los que la madre es AA y el feto es AB se indican como AA|AB. Figura 30A: Gráficos generados cuando dos cromosomas están presentes y la fracción ADN libre de células fetal es el 0%. Este gráfico es de una mujer que no está embarazada y, por tanto, representa el patrón cuando el genotipo es completamente materno. Las agrupaciones de alelos, por tanto, están centradas alrededor de 1 (AA alelos), 0,5 (AB alelos), y 0 (BB alelos). Figura 30B: Gráfico generado cuando dos cromosomas están presentes y la fracción fetal es el 12%. La contribución de alelos fetales a la fracción de alelo A lee los cambios de posición de algunos puntos de alelo arriba y abajo a lo largo del eje y, de modo que las bandas están centradas alrededor de 1 (alelos AA|AA), 0,94 (alelos AA|AB), 0,56 (alelos AB|AA), 0,50 (alelos AB|AB), 0,44 (alelos AB|BB), 0,06 (alelos BB|AB), y 0 (alelos BB|BB). Figura 30C: Gráfico generado cuando dos cromosomas están presentes y la fracción fetal es el 26%. Este patrón, que incluye dos bandas rojas y dos bandas azules periféricas y un trío de bandas verdes en el centro, se puede ver rápidamente (no se muestra el color). Las bandas están centradas alrededor de 1 (alelos AA|AA), 0,87 (alelos AA|AB), 0,63 (alelos AB|AA alelos), 0,50 (alelos AB|AB), 0,37 (alelos AB|BB), 0,13 (alelos BB|AB), y 0 (alelos BB|BB). Figura 30D: Gráfico generado cuando un cromosoma está presente y la fracción fetal es el 26%. El patrón distintivo de una banda periférica externa roja y una externa azul además de dos bandas verdes en el centro indica monosomía heredada de la madre (no se muestra el color). Puesto que el feto sólo contribuye un único alelo (A o B) a las lecturas de alelos, las bandas periféricas roja y azul internas no están presentes, y el trío de bandas del centro se condensa en dos bandas (no se muestra el color). Las bandas están centradas alrededor de 1 (alelos AA|A), 0,57 (alelos AB|A), 0,43 (alelos AB|B), y 0 (alelos BB|B). Figura 30E: Gráfico generado cuando tres cromosomas están presentes y la fracción fetal es el 27%. Este patrón de dos bandas rojas y dos azules periféricas además de dos bandas verdes en el centro indica trisomía mitótica heredada de la madre (no se muestra el color). Las bandas están centradas alrededor de 1 (alelos AA|AAA), 0,88 (alelos AA|AAB), 0,56 (alelos AB|AAB), 0,44 (alelos AB|ABB), 0,12 (alelos BB|ABB)

y 0 (alelos BB|BBB). Figura 30F: Gráfico generado cuando tres cromosomas están presentes y la fracción fetal es el 14%. Este patrón de tres bandas rojas y tres azules periféricas además de dos bandas verdes en el centro indica trisomía mitótica heredada del padre (no se muestra el color). Las bandas están centradas alrededor de 1 (alelos AA|AAA), 0,93 (alelos AA|AAB), 0,87 (alelos AA|ABB), 0,60 (alelos AB|AAA), 0,53 (alelos AB|AAB), 0,47 (alelos AB|ABB), 0,40 (alelos AB|BBB), 0,13 (alelos BB|AAB), 0,07 (alelos BB|ABB), y 0 (alelos BB|BBB). Figura 30G: Gráfico generado cuando tres cromosomas están presentes y la fracción fetal es el 35%). Este patrón de dos bandas rojas y dos bandas azules periféricas y cuatro bandas verdes en el centro indica trisomía mitótica heredada de la madre (no se muestra el color). Las bandas están centradas alrededor de 1 (alelos AA|AAA), 0,85 (alelos AA|AAB), 0,72 (alelos AB|AAA), 0,57 (alelos AB|AAB), 0,43 (alelos AB|ABB), 0,28 (alelos AB|BBB), 0,15 (alelos BB|ABB), y 0 (alelos BB|BBB). Figura 30H: Gráfico generado cuando tres cromosomas están presentes y la fracción fetal es el 25%. Este patrón de dos bandas rojas y dos azules periféricas además de cuatro bandas verdes en el centro indica trisomía mitótica heredada del padre (no se muestra el color). Este patrón se puede distinguir del de la trisomía mitótica heredada de la madre (como en la Figura 30G:) por la posición de las bandas periféricas internas. Específicamente, las bandas están centradas alrededor de 1 (alelos AA|AAA), 0,78 (alelos AA|ABB), 0,67 (alelos AB|AAA), 0,56 (alelos AB|AAB), 0,44 (alelos AB|ABB), 0,33 (alelos AB|BBB), 0,22 (alelos BB|AAB), y 0 (alelos BB|BBB).

Figura 31: Representaciones gráficas de (Figura 31 A) muestras de prueba euploide, (Figura 31B) T13, (Figura 31C) T18, (Figura 31D) T21, (Figura 31E) 45,X, y (Figura 31F) 47,XXY según se indica. Cada cromosoma se indica en la parte superior del gráfico, los genotipos fetales y maternos se indican a la derecha de los gráficos, el eje x representa la posición lineal de los SNP a lo largo de cada cromosoma, y el eje y indica el número de lecturas de alelos A como una fracción de las lecturas totales. Cabe señalar la posición de la agrupación alterada basada en la fracción fetal que se describe en el presente documento. Cada punto representa un solo locus SNP. Los genotipos fetales y maternos se indican a la derecha del gráfico, y las identidades de los cromosomas se indican en la parte superior de los gráficos.

Figura 32: El predominio de aneuploidías cromosómicas sexuales combinada en el nacimiento es mayor que la de aneuploidías autosómicas.

A pesar de que los dibujos anteriormente identificados exponen las realizaciones divulgadas en el presente documento, también se contemplan otras realizaciones, tal y como se hace constar en la exposición. Esta realización presenta realizaciones ilustrativas mediante una representación.

Descripción detallada

5 La invención se define en las reivindicaciones que se incluyen en el anexo. La presente invención se basa en parte en el sorprendente descubrimiento de que a menudo sólo un número relativamente pequeño de cebadores en una biblioteca de cebadores son responsables de una cantidad sustancial de los dímeros de cebadores amplificados que se forman durante reacciones PCR multiplexadas. Se han desarrollado métodos para seleccionar los cebadores menos deseados para eliminarlos de una

10 biblioteca de cebadores candidatos. Al reducir la cantidad de dímeros de cebadores a una cantidad negligible (-0,1% de los productos de PCR, estos métodos permiten a las bibliotecas de cebadores resultantes amplificar simultáneamente un gran número de loci diana en una sola reacción PCR multiplexada. Puesto que los cebadores se hibridan con los loci diana y los amplifican en lugar de hibridarse con otros cebadores y formar dímeros de cebadores amplificados, el número de loci diana diferentes que se pueden amplificar aumenta. También se ha descubierto que utilizando

15 concentraciones de cebador más bajas y tiempos de reformación térmica mucho más largos de lo normal aumenta la probabilidad de que los cebadores se hibriden en los loci diana en lugar de hibridarse entre sí y formen dímeros de cebadores.

Durante la amplificación por PCR y la secuenciación de 19.488 loci diana en una muestra genómica, 99,4- 99,7% de las lecturas de secuenciación se correlacionaron con el genoma, de esas, 99,99% de las correlacionadas con los loci focalizados. Para las muestras de plasma con 10 millones de lecturas de secuenciación, de forma típica al menos 19.350 de los 19.488 loci focalizados (99,3 %) se amplificaron y se secuenciaron. Ser capaz de amplificar simultáneamente un gran número de loci diana a la vez reduce enormemente la cantidad de tiempo y la cantidad de ADN necesarios para analizar miles de loci diana. Por ejemplo, el ADN de una sola célula es suficiente para analizar simultáneamente miles de loci diana, lo cual es importante para aplicaciones en las que la cantidad de ADN es baja

20 como, por ejemplo, las pruebas genéticas de una sola célula de un embrión antes de la fertilización in vitro o la realización de pruebas genéticas de una muestra forense con poco ADN. Además, poder analizar los loci diana en un volumen de reacción (como una cámara o un pocillo) en lugar de dividir la muestra en múltiples reacciones diferentes reduce la variabilidad que puede ocurrir entre reacciones. Además, se han desarrollado métodos para utilizar estándares de referencia para corregir el sesgo de amplificación que puede ocurrir entre loci diana diferentes. Por ejemplo, diferencias en la eficiencia de la amplificación entre loci diana debidas a factores como el contenido de GC pueden producir diferentes cantidades de productos de PCR para ser producidos para loci diana que están realmente presentes en la misma cantidad. El uso de estándares de referencia similares a los loci diana permite la detección de dicho sesgo de amplificación de modo que se pueda corregir durante la cuantificación de los loci diana.

Durante la secuenciación de productos de PCR, se detectan artefactos como dímeros de cebadores y, por tanto, inhiben la detección de amplicones diana. Debido a esta limitación, se suelen utilizar microarrays con sondas de hibridación para la detección ya que las microarrays son menos sensibles a las interferencias de los dímeros de cebadores. El alto nivel de multiplexado con amplicones diana mínimos que ahora se ha conseguido permite la PCR seguida de una secuenciación para ser utilizado como una alternativa a los microarrays.

Los métodos de PCR multiplexada de la invención pueden estar en una variedad de aplicaciones como, por ejemplo, determinación del genotipo, detección de anomalías cromosómicas (como aneuploidía de cromosoma fetal), mutación de genes y análisis de polimorfismos (como, por ejemplo, polimorfismos de un solo nucleótido, SNP), análisis de delección de genes, determinación de la paternidad, análisis de diferencias genéticas entre las poblaciones, análisis forense, medición de la predisposición a una enfermedad, análisis cuantitativo de mRNA, y detección e identificación de agentes infecciosos (como bacterias, parásitos, y virus). Los métodos de PCR multiplexada también se pueden utilizar para pruebas prenatales no invasivas, como pruebas de paternidad o la detección de anomalías cromosómicas fetales.

Ejemplos de métodos de diseño de cebadores

La PCR altamente multiplexada a menudo puede dar como resultado la producción de una proporción muy alta de ADN de producto que genera reacciones secundarias improductivas como la formación de dímeros de cebadores. En una realización, los cebadores particulares que son más propensos a causar reacciones secundarias improductivas se pueden quitar de la biblioteca de cebadores para dar una biblioteca de cebadores que dará como resultado una mayor proporción de ADN amplificado que se corresponde con el genoma. El paso de quitar cebadores problemáticos, es decir, aquellos cebadores que son especialmente propensos a formar dímeros ha permitido inesperadamente niveles de multiplexado para PCR extremadamente altos para un posterior análisis por secuenciación. En sistemas como el de secuenciación, donde el rendimiento se degrada significativamente por los dímeros de cebadores y/o otros productos problemáticos, se ha conseguido un multiplexado mayor que 10, mayor que 50, y mayor que 100 veces más alto que otros multiplexados descritos. Cabe señalar que esto se opone a los métodos de detección basados en una sonda, por ejemplo, microarrays, TAQMAN, PCR etc. donde un exceso de dímeros de cebadores no afectará al resultado de forma apreciable. También cabe señalar que la creencia general en la técnica es que multiplexar la PCR para la secuenciación está limitado a unos 100 ensayos en el mismo pocillo. Fluidigm y Rain Dance ofrecen plataformas para realizar 48 o

50 1000s ensayos por PCR en reacciones paralelas para una muestra.

Hay varias formas de elegir cebadores para una biblioteca donde la cantidad de dímeros de cebadores no correspondiente u otros productos problemáticos de cebador se minimizan. Los datos empíricos indican que un pequeño número de cebadores ‘malos’ son responsables de una gran cantidad de reacciones secundarias de dímeros de cebadores no correspondientes. Quitar estos cebadores ‘malos’ puede aumentar el porcentaje de lecturas de secuencia que corresponden a los loci focalizados. Una manera de identificar los cebadores ‘malos’ es mirar a los datos de secuenciación de ADN que se amplificó mediante la amplificación focalizada; estos dímeros de cebadores que se ven con mayor frecuencia se pueden quitar para dar una biblioteca de cebadores que es significativamente menos probable de producir ADN de producto secundario que no corresponde con el genoma. También hay programas disponibles públicamente que pueden calcular la energía de unión de varias combinaciones de cebadores, y quitar aquellos con la energía de unión más alta también darán una biblioteca de cebadores que es significativamente menos probable de producir ADN de producto secundario que no corresponde con el genoma.

En algunas realizaciones para seleccionar cebadores, se crea una biblioteca inicial de cebadores candidatos designando uno o más cebadores o pares de cebadores a loci diana candidatos. Un conjunto de loci diana candidatos (como los SNP) se puede seleccionar en base a información disponible públicamente sobre parámetros deseados para los loci diana, como frecuencia de SNP dentro de una población diana o la tasa de heterocigosidad de los SNP. En una realización, los cebadores para PCR pueden estar diseñados utilizando el programa Primer3 (página web primer3.sourceforge.net; libprimer3 versión 2.2.3). Si lo desea, los cebadores se pueden diseñar para reformar térmicamente en un intervalo de temperatura de reformación térmica determinado, tener un intervalo determinado de contenido de GC, tener un intervalo de tamaño determinado, producir amplicones diana en un intervalo de tamaño determinado, y/o tener otras características de parámetros. Empezar con múltiples cebadores o pares de cebadores por locus diana candidato aumenta la probabilidad de que un cebador o par de cebadores se mantenga en la biblioteca para la mayoría de loci diana o para todos ellos. En una realización, los criterios de selección pueden requerir que al menos un par de cebadores por locus diana permanezca en la biblioteca. De este modo, la mayoría o todos los loci diana se amplificarán cuando se utilice la biblioteca de cebadores final. Esto es deseable para aplicaciones como análisis para eliminaciones o duplicaciones de un gran número de ubicaciones en el genoma o análisis para un gran número de secuencias (como polimorfismos u otras mutaciones) asociadas con una enfermedad o un mayor riesgo para una enfermedad. Si un par de cebadores de la biblioteca produjese un amplicón diana que se solapa con un amplicón diana producido por otro par de cebadores, uno de los pares de cebadores se puede quitar de la biblioteca para evitar interferencias.

En algunas realizaciones, las “puntuaciones no deseables” (la puntuación más alta representa la menos deseable) se calcula (como el cálculo en un ordenador) para la mayoría o todas las combinaciones posibles de dos cebadores de una biblioteca de cebadores candidatos. En varias realizaciones, una puntuación no deseable se calcula para al menos 80, 90, 95, 98, 99, o 99,5% de las combinaciones posibles de cebadores candidatos en la biblioteca. Cada puntuación no deseable se basa al menos en parte en la probabilidad de formación de dímeros entre los dos cebadores candidatos. Es deseable, que la puntuación no deseable también se puede basar en uno o más parámetros seleccionados del grupo compuesto por la tasa de heterocigosidad del locus diana, el predominio de la enfermedad asociada con una secuencia (por ejemplo, un polimorfismo) en el locus diana, la penetrancia de la enfermedad asociada con una secuencia (por ejemplo, un polimorfismo) en el locus diana, especificidad del cebador candidato para el locus diana, tamaño del cebador candidato, la temperatura de fusión del amplicón diana, el contenido de GC del amplicón diana, la eficiencia de la amplificación del amplicón diana, y el tamaño del amplicón diana. Si se consideran múltiples factores, la puntuación no deseable se puede calcular en base a una media ponderada de los diversos parámetros. A los parámetros se les puede asignar diferentes pesos en base a su importancia para la aplicación particular para la que se utilizarán los cebadores. En algunas realizaciones, el cebador con la puntuación no deseable más alta se quita de la biblioteca. Si el cebador que se ha quitado es un miembro de un par de cebadores que se hibrida con un locus diana, entonces el otro miembro del par de cebadores se puede quitar de la biblioteca. El proceso de quitar cebadores se puede repetir según se desee. En algunas realizaciones, el método de selección se lleva a cabo hasta que las puntuaciones no deseables para las combinaciones del cebador candidato restantes en la biblioteca son todas iguales o por debajo de un umbral mínimo. En algunas realizaciones, el método de selección se lleva a cabo hasta que el número de cebadores candidatos restantes en la biblioteca se reduce a un número deseado.

En varias realizaciones, después de que se hayan calculado las puntuaciones no deseables, el cebador candidato que es parte del mayor número de combinaciones de dos cebadores candidatos con una puntuación no deseable por encima de un primer umbral mínimo se quita de la biblioteca. Este paso omite las interacciones iguales o por debajo del primer umbral mínimo ya que estas interacciones son menos significativas. Si el cebador que se ha quitado es un miembro de un par de cebadores que se hibrida con un locus diana, entonces el otro miembro del par de cebadores se puede quitar de la biblioteca. El proceso de quitar cebadores se puede repetir según se desee. En algunas realizaciones, el método de selección se lleva a cabo hasta que las puntuaciones no deseables para las combinaciones del cebador candidato restantes en la biblioteca son todas iguales o por debajo del primer umbral mínimo. Si el número de cebadores candidatos restantes en la biblioteca es superior al deseado, el número de cebadores se puede reducir disminuyendo el primer umbral mínimo a un segundo umbral mínimo más bajo y repetir el proceso de quitar cebadores. Si el número de cebadores candidatos restantes en la biblioteca es inferior al deseado, el método se puede continuar aumentando el primer umbral mínimo a un segundo umbral mínimo más alto y repetir el proceso de quitar cebadores utilizando la biblioteca de cebadores candidatos original, permitiendo, de este modo, que más cebadores candidatos permanezcan en la biblioteca. En algunas realizaciones, el método de selección se lleva a cabo hasta que las puntuaciones no deseables para las combinaciones del cebador candidato restantes en la biblioteca son todas iguales o

por debajo del segundo umbral mínimo, o hasta que el número de cebadores candidato restantes en la biblioteca se reduce a un número deseado.

Si lo desea, los pares de cebador que producen un amplicón diana que se solapa con un amplicón diana producido por otro par de cebadores se pueden dividir en reacciones de amplificación separadas. Las reacciones de amplificación por PCR múltiples pueden ser deseables para aplicaciones en las que es deseable analizar todos los loci diana candidatos (en lugar de omitir loci diana candidatos del análisis debido a la superposición de amplicones diana).

Estos métodos de selección minimizan el número de cebadores candidatos que se tienen que quitar de la biblioteca para conseguir la reducción deseada en dímeros de cebadores. Al quitar un número más pequeño de cebadores candidatos de la biblioteca, se pueden amplificar más loci diana (o todos) utilizando la biblioteca de cebadores resultante. El multiplexado de grandes números de cebadores impone restricciones considerables en los ensayos que se pueden incluir. Los ensayos que interactúan involuntariamente dan como resultado productos de amplificación

falsos. Las restricciones de tamaño de miniPCR pueden dar como resultado más restricciones. En una realización, es posible empezar con un gran número de SNP diana potenciales (entre aproximadamente 500 a más de 1 millón) e intentar designar cebadores para amplificar cada SNP. Cuando se pueden diseñar cebadores es posible intentar identificar pares de cebador que es muy probable que puedan formar productos falsos evaluando la probabilidad de formación doble de cebadores falsos entre todos los pares de cebadores posibles utilizando parámetros termodinámicos publicados para la formación doble de ADN. Las interacciones del cebador se pueden clasificar mediante una función de puntuación relacionada con la interacción y los cebadores con las peores puntuaciones de interacción se eliminan hasta que se cumpla el número de cebadores deseados. En los casos en los que las SNP que tienen probabilidades de ser heterocigóticas son muy útiles, es posible también clasificar la lista de ensayos y seleccionar los ensayos más compatibles con heterocigotos. Los experimentos han validado que los cebadores con puntuaciones de interacción altas son más propensos a formar dímeros de cebadores. Con una multiplexación alta no es posible eliminar todas las interacciones falsas, pero es esencial quitar los cebadores o pares de cebadores con las puntuaciones de interacción más altas in silico ya que pueden dominar toda una reacción, limitando en gran medida la amplificación para las dianas previstas. Hemos realizado este procedimiento para crear conjuntos de cebadores multiplexados de hasta y en algunos casos más de 100.0 cebadores. La mejora debida a este procedimiento es sustancial, ya que permite la amplificación de más del 80%, más del 90%, más del 95%, más del 98%, e incluso más del 99% en los productos diana determinada por la secuenciación de todos los productos de PCR, en comparación con un 10% a partir de una reacción en la que los peores cebadores no se quitaron. Cuando se combina con un enfoque semianidado parcial como se ha descrito previamente, más de un 90%, e incluso más de un 95% de los amplicones pueden corresponder a las secuencias focalizadas.

Tenga en cuenta que hay otros métodos para determinar que sondas para PCR es probable que formen dímeros. En una realización, el análisis de un conjunto de ADN que se ha amplificado utilizando un conjunto de cebadores no optimizados puede ser suficiente para determinar cebadores problemáticos. Por ejemplo, el análisis se puede realizar utilizando secuenciación, y aquellos dímeros que están presentes en mayor número están determinados a ser aquellos que tienen más probabilidades de formar dímeros, y se pueden quitar.

Este método tiene un número de aplicaciones potenciales, por ejemplo, para la determinación del genotipo de SNP, la determinación de la tasa de heterocigosidad, la medición del número de copias, y otras aplicaciones de secuenciación focalizadas. En una realización, el método de diseño de cebador se puede utilizar en combinación con el método de mini-PCR que se describe en cualquier otra parte en este documento. En algunas realizaciones, el método de diseño del cebador se puede utilizar como parte de un método de PCR multiplexada masivo.

El uso de etiquetas en los cebadores puede reducir la amplificación y secuenciación de productos de dímeros de cebadores. En algunas realizaciones, el cebador contiene una región interna que forma una estructura en bucle con una etiqueta. En realizaciones específicas, los cebadores incluyen una región 5' que es específica para un locus diana, una región interna que no es específica para el locus diana y forma una estructura de bucle, y una región 3' que es específica para el locus diana. En algunas realizaciones, la región de bucle puede encontrarse entre dos regiones de unión donde las dos regiones de unión están diseñadas para unirse a regiones continuas o colindantes de ADN de la plantilla. En varias realizaciones, la longitud de la región 3' es al menos de 7 nucleótidos. En algunas realizaciones, la longitud de la región 3' está entre 7 y 20 nucleótidos, por ejemplo, entre 7 y 15 nucleótidos, o 7 y 10 nucleótidos, incluidos. En varias realizaciones, los cebadores incluyen una región 5' que no es específica para un locus diana (como una etiqueta o un punto de unión del cebador universal) seguido de una región que es específica de un locus diana, una región interna que no es específica para el locus diana y forma una estructura de bucle, y una región 3' que es específica del locus diana. Los cebadores de etiquetas se pueden utilizar para acortar secuencias específicas diana necesarias por debajo de 20, por debajo de 15, por debajo de 12, e incluso por debajo de 10 pares base. Esto puede ser fortuito con un diseño de cebador estándar cuando la secuencia diana se fragmenta dentro del punto de unión del cebador, o se puede diseñar en el diseño del cebador. Las ventajas de este método incluyen: aumenta el número de ensayos que se pueden diseñar para una determinada longitud de amplicón máxima, y acorta la secuenciación "no

informativa" de la secuencia del cebador. También se puede utilizar en combinación con el etiquetado interno (véase en otra parte de este documento).

En una realización, la cantidad relativa de productos no productivos en la amplificación por PCR focalizada multiplexada se puede reducir subiendo la temperatura de reformación térmica. En los casos en que se amplifican bibliotecas con la

misma etiqueta que los cebadores específicos diana, la temperatura de reformación térmica se puede aumentar en comparación con el ADN genómico ya que las etiquetas contribuirán a la unión del cebador. En algunas realizaciones utilizamos concentraciones de cebador considerablemente más bajas que las que se han informado anteriormente junto con la utilización de tiempos de reformación térmica más largos que los que se han informado en cualquier otra parte.

5 En algunas realizaciones los tiempos de reformación térmica pueden durar más de 3 minutos, más de 5 minutos, más de 8 minutos, más de 10 minutos, más de 15 minutos, más de 20 minutos, más de 30 minutos, más de 60 minutos, más de 120 minutos, más de 240 minutos, más de 480 minutos, e incluso más de 960 minutos. En una realización, se utilizan tiempos de reformación térmica más largos que en informes anteriores, lo que permite concentraciones de cebador más bajas. En varias realizaciones, se utilizan tiempos de reformación térmica más largos de lo normal, por ejemplo, superiores a 3, 5, 8, 10, o 15 minutos. En algunas realizaciones, las concentraciones de cebador son tan bajas como 50 nM, 20 nM, 10 nM, 5 nM, 1 nM, e inferiores a 1 uM. Esto sorprendentemente da como resultado un rendimiento robusto para las reacciones altamente multiplexadas, por ejemplo, reacciones de 1.000-plex, reacciones de 2.000-plex, reacciones de 5.000-plex, reacciones de 10.000-plex, reacciones de 20.000-plex, reacciones de 50.000-plex, e incluso reacciones de 100.000-plex. En una realización, la amplificación utiliza uno, dos, tres, cuatro o cinco ciclos ejecutados con tiempos de reformación térmica prolongados, seguida de ciclos de PCR con tiempos de reformación térmica más habituales con cebadores con etiquetas.

Para seleccionar ubicaciones diana, se puede empezar con un conjunto de diseños de par de cebadores candidatos y crear un modelo termodinámico de interacciones potencialmente adversas entre pares de cebadores y, a continuación, utilizar el modelo para eliminar diseños que son incompatibles con otros diseños del conjunto.

20 Después del proceso de selección, los cebadores restantes en la biblioteca se pueden utilizar en cualquiera de los métodos de la invención.

Ejemplos de bibliotecas de cebadores

En un aspecto, la invención presenta bibliotecas de cebadores, dichos cebadores se seleccionan de una biblioteca de cebadores candidatos utilizando cualquiera de los métodos de la invención. En algunas realizaciones, la biblioteca incluye cebadores que se hibridan simultáneamente (o son capaces de hibridarse simultáneamente) o que se amplifican simultáneamente (o son capaces de amplificarse simultáneamente) al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes en un volumen de reacción. En varias realizaciones, la biblioteca incluye cebadores que amplifican simultáneamente (o son capaces de amplificar simultáneamente) entre 1.000 y 2.000; 2.000 y 5.000; 5.000 y 7.500; 7.500 y 10.000; 10.000 y 20.000; 20.000 y 25.000; 25.000 y 30.000; 30.000 y 40.000; 40.000 y 50.000; 50.000 y 75.000; o 75.000 y 100.000 loci diana diferentes en un volumen de reacción. En varias realizaciones, la biblioteca incluye cebadores que amplifican simultáneamente (o son capaces de amplificar simultáneamente) entre 1.000 y 100.000; loci diana diferentes en un volumen de reacción, como, por ejemplo, entre 1.000 y 50.000; 1.000 y 30.000; 1.000 y 20.000; 1.000 y 10.000; 2.000 y 30.000; 2.000 y 20.000; 2.000 y 10.000; 5.000 y 30.000; 5.000 y 20.000; o 5.000 y 10.000 loci diana diferentes, incluidos. En algunas realizaciones, la biblioteca incluye cebadores que amplifican simultáneamente (o pueden amplificar simultáneamente) los loci diana en un volumen de reacción de tal modo que menos de un 60, 40, 30, 20, 10, 5, 4, 3, 2, 1, 0,5, 0,25, 0,1, o 250,5% de los productos amplificados son dímeros de cebadores. Las diversas realizaciones, la cantidad de productos amplificados que son dímeros de cebadores está entre 0,5 y 60%, como, por ejemplo, entre 0,1 y 40%, 0,1 y 20%, 0,25 y 20%, 0,25 y 10%, 0,5 y 20%, 0,5 y 10%, 1 y 20%, o 1 y 10%, incluido. En algunas realizaciones, los cebadores amplifican simultáneamente (o pueden amplificar simultáneamente) los loci diana en un volumen de reacción de tal modo que al menos un 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los productos amplificados son amplicones diana. En varias realizaciones, la cantidad de productos amplificados que son amplicones diana está entre 50 y 99,5%, como, por ejemplo, entre 60 y 99%, 70 y 98%, 80 y 98%, 90 y 99,5%, o 95 y 99,5%, incluido. En algunas realizaciones, los cebadores amplifican simultáneamente (o pueden amplificar simultáneamente) los loci diana en un volumen de reacción de tal modo que al menos un 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los loci focalizados son amplificados. En varias realizaciones, la cantidad de loci diana que se amplifican está entre 50 y 99,5%, como, por ejemplo, entre 60 y 99%, 70 y 98%, 80 y 99%, 90 y 99,5%, 95 y 99,9%, o 98 y 99,99% incluido. En algunas

realizaciones, la biblioteca de cebadores incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 pares de cebador de prueba, en donde cada par de cebadores incluye un cebador de prueba directo y un cebador de prueba inverso en el que cada par de cebadores de prueba se hibrida en un locus diana. En algunas realizaciones, la biblioteca de cebadores incluye al menos 101.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 cebadores de prueba individuales que se hibridan cada uno de ellos a diferentes locus diana, en donde los cebadores individuales no forman parte de los pares de cebador.

En varias realizaciones, la concentración de cada cebador es inferior a 100, 75, 50, 25, 20, 10, 5, 2, o 1 nM, o inferior a 500, 100, 10, o 1 uM. En varias realizaciones, la concentración de cada cebador está entre 1 uM y 100 nM, como, por ejemplo, entre 1 uM y 1 nM, 1 y 75 nM, 2 y 50 nM o 5 y 50 nM, incluido. En varias realizaciones, el contenido de GC de los cebadores está entre 30 y 80%, como, por ejemplo, entre 40 y 70% o 50 y 60%, incluidos. En algunas realizaciones, el intervalo de contenido de GC de los cebadores es inferior a 30, 20, 10, o 5%. En algunas realizaciones, el intervalo de contenido de GC de los cebadores está entre 5 y 30%, como, por ejemplo, 5 y 20% o 5 y 10%), incluido. En algunas realizaciones, la temperatura de fusión (T_m) de los cebadores de prueba es entre 40 y 80 °C, como, por ejemplo, 50 y 70 °C, 55 y 65 °C, o 57 y 60,5 °C, incluido. En algunas realizaciones, la T_m se calcula utilizando el programa Primer3 (libprimer3 versión 2.2.3) utilizando los parámetros SantaLucia integrados (página web primer3.sourceforge.net). En

algunas realizaciones, el intervalo de la temperatura de fusión de los cebadores es inferior a 15, 10, 5, 3, o 1 °C. 25En algunas realizaciones, el intervalo de la temperatura de fusión de los cebadores está entre 1 y 15 °C, como, por ejemplo, entre 1 y 10 °C, 1 y 5 °C, o 1 y 3 °C, incluido. En algunas realizaciones, la longitud de los cebadores está entre 15 y 100 nucleótidos, como, por ejemplo, entre 15 y 75 nucleótidos,

- 5 15 y 40 nucleótidos, 17 y 35 nucleótidos, 18 y 30 nucleótidos, 20 y 65 nucleótidos, incluido. En algunas realizaciones, el intervalo de la longitud de los cebadores es inferior a 50, 40, 30, 20, 10, o 5 nucleótidos. En algunas realizaciones, el intervalo de la longitud de los cebadores está entre 5 y 50 nucleótidos, como, por ejemplo, 5 y 40 nucleótidos, 5 y 20 nucleótidos, o 5 y 10 nucleótidos, incluidos. En algunas realizaciones, la longitud de los amplicones diana está entre 50 y 100 nucleótidos, por ejemplo, entre 60 y 80 nucleótidos, o 60 y 75 nucleótidos, incluidos. En algunas realizaciones, el
10 intervalo de la longitud de los amplicones diana es inferior a 50, 25, 15, 10, o 5 nucleótidos. En algunas realizaciones, el intervalo de la longitud de los amplicones diana está entre 5 y 50 nucleótidos, como, por ejemplo, 5 y 25 nucleótidos, 5 y 15 nucleótidos, o 5 y 10 nucleótidos, incluidos.

Estas bibliotecas de cebadores se pueden utilizar en cualquiera de los métodos de la invención.

Ejemplos de kits de cebadores

- 15 En un aspecto, la invención presenta un kit (como, por ejemplo, kits para amplificar loci diana en una muestra de ácido nucleico) que incluye cualquiera de las bibliotecas de cebadores de la invención. En algunas realizaciones, puede formularse un kit que comprende una pluralidad de cebadores diseñado para alcanzar los métodos descritos en esta divulgación. Los cebadores pueden ser cebadores directos e inversos exteriores, cebadores directos e inversos
20 interiores como se divulga en el presente documento, pueden ser cebadores que han sido diseñados para tener una afinidad de unión baja con otros cebadores en el kit como se describe en la sección sobre diseño de cebadores, pueden ser sondas de captura híbridas o sondas precircularizadas como se describe en las secciones pertinentes, o alguna combinación de estas. En una realización, se puede formular un kit para determinar un estado de ploidía de un cromosoma diana en un feto en gestación diseñado para ser utilizado con los métodos divulgados en el presente documento, el kit comprende una pluralidad de cebadores directos interiores y opcionalmente una pluralidad de
25 cebadores inversos interiores, y opcionalmente cebadores directos exteriores y cebadores inversos exteriores, donde cada uno de los cebadores está diseñado para que se hibride en la región del ADN inmediatamente corriente arriba y/o corriente abajo de uno de los puntos diana (por ejemplo, puntos polimórficos) en el cromosoma diana, y opcionalmente cromosomas adicionales. En una realización, el kit de cebador se puede utilizar en combinación con el cuadro de diagnóstico que se describe en otra parte de este documento. En algunas realizaciones, el kit incluye instrucciones para
30 utilizar la biblioteca para amplificar los loci diana.

Ejemplos de métodos de PCR multiplexada

- En un aspecto, la invención presenta métodos de amplificar loci diana en una muestra de ácido nucleico que implica (i) poner en contacto la muestra de ácido nucleico con una biblioteca de cebadores que se hibridan simultáneamente a
35 menos de 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes para producir una mezcla de reacción; y (ii) someter la mezcla de reacción a las condiciones de reacción de extensión del cebador (como, por ejemplo, las condiciones de PCR) para producir productos amplificados que incluyen amplicones diana. En algunas realizaciones, el método también incluye determinar la presencia o ausencia de al menos un amplicón diana (por ejemplo, al menos un 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los amplicones diana). En algunas realizaciones, el método también incluye determinar la secuencia de al menos un amplicón diana (por
40 ejemplo, al menos un 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los amplicones diana). En algunas realizaciones, al menos un 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, o 99,5% de los loci diana son amplificados. En varias realizaciones, menos de un 60, 50, 40, 30, 20, 10, 5, 4, 3,2, 1, 0,5, 0,25, 0,1, o 0,05% de los productos amplificados son dímeros de cebadores.

- En una realización, un método divulgado en el presente documento utiliza la PCR focalizada altamente multiplexada y altamente eficiente para amplificar ADN, seguida de una secuenciación de alto rendimiento para determinar las
45 frecuencias alélicas en cada locus diana. La capacidad para multiplexar más de unos 50 o 100 cebadores para PCR en un volumen de reacción de forma que la mayoría de las lecturas de secuencia resultantes que se correspondan con los loci diana sea algo novedoso y no obvio. Una técnica que permite realizar una PCR focalizada altamente multiplexada de forma altamente eficiente implica el diseño de cebadores que es poco probable que se hibriden entre sí. Las sondas para PCR, a las que de forma típica se hace referencia como cebadores, se seleccionan creando un modelo termodinámico de interacciones potencialmente adversas entre al menos 500; al menos 1.000; al menos 2.000; al
50 menos 5.000; al menos 7.500; al menos 10.000; al menos 20.000; al menos 25.000; al menos 30.000; al menos 40.000; al menos 50.000; al menos 75.000; o al menos 100.000 pares de cebadores potenciales, o interacciones no intencionadas entre cebadores y el ADN de la muestra y, a continuación, utilizando el modelo para eliminar diseños que son incompatibles con otros diseños del conjunto. Otra técnica que permite realizar una PCR focalizada altamente multiplexada de forma altamente eficiente consiste en utilizar un planteamiento de anidado parcial o total en la PCR focalizada. El uso de uno o una combinación de estos planteamientos permiten el multiplexado de al menos 300, al
55 menos 800, al menos 1200, al menos 4000 o al menos 10.000 cebadores en un único conjunto con el ADN amplificado resultante que comprende una mayoría de moléculas de ADN que, cuando se secuencia, se corresponderá con los loci diana. El uso de uno o una combinación de estos planteamientos permite el multiplexado de un gran número de cebadores en un único conjunto con el ADN amplificado resultante que comprende más de un 50%, más de un 60%,
60

más de un 67%, más de un 80%, más de un 90%, más de un 95%, más de un 96%, más de un 97%, más de un 98%, más de un 99%, o más de un 99,5% de moléculas de ADN que representan un mapa de los loci diana.

En algunas realizaciones la detección del material genético diana se puede realizar de forma multiplexada. El número de secuencias diana genéticas que se pueden ejecutar en paralelo puede ir de uno a diez, de diez a cien, de cien a mil, de mil a diez mil, de diez mil a cien mil, de cien mil a un millón, o de un millón a diez millones. Los intentos anteriores de multiplexar más de 100 cebadores por conjunto han dado como resultado problemas significativos con reacciones secundarias no deseadas tales como la formación de dímeros de cebador.

PCR focalizada

En algunas realizaciones, la PCR se puede utilizar para focalizar ubicaciones específicas del genoma. En muestras de plasma, el ADN original está altamente fragmentado (de forma típica es inferior a 500 bp, con una longitud media inferior a 200 bp). En la PCR, tanto los cebadores directos como los inversos se recuecen en el mismo fragmento para permitir la amplificación. Por tanto, si los fragmentos son cortos, los ensayos por PCR deben amplificar también las regiones relativamente cortas. Al igual que las MIPS, si las posiciones polimórficas están demasiado cerca del punto de unión de la polimerasa, podría dar como resultado ciertos sesgos en la amplificación de alelos diferentes. Actualmente, los cebadores para PCR que focalizan regiones polimórficas, como aquellas que contienen SNP, están diseñadas de forma típica de tal modo que el extremo 3' del cebador se hibridará en la base inmediatamente adyacente a la base o bases polimórficas. En una realización de la presente divulgación, los extremos 3' tanto del cebador para PCR directo como del cebador para PCR inverso están diseñados para hibridarse a bases que están a una o varias posiciones de distancia de las posiciones de la variante (puntos polimórficos) del alelo focalizado. El número de bases entre el punto polimórfico (SNP u otro) y la base a la que el extremo 3' del cebador está diseñado para que se hibride puede ser una base, pueden ser dos bases, pueden ser tres bases, pueden ser cuatro bases, pueden ser cinco bases, pueden ser seis bases, pueden ser de siete a diez bases, pueden ser de once a quince bases, o pueden ser de dieciséis a veinte bases. Los cebadores directo e inverso pueden estar diseñados para hibridar un número diferente de bases lejos del punto polimórfico.

El ensayo de PCR se puede generar en grandes cantidades, sin embargo, las interacciones entre diferentes ensayos de PCR dificulta multiplexarlos más allá de unos cien ensayos. Se pueden utilizar varios planteamientos moleculares complejos para aumentar el nivel de multiplexado, aunque puede seguir estando limitado a poco más de 100, quizá 200, o posiblemente 500 ensayos por reacción. Se pueden dividir muestras con grandes cantidades de ADN entre múltiples sub-reacciones y después recombinarse antes de la secuenciación. Para muestras donde o bien la muestra en conjunto o una subpoblación de moléculas de ADN es limitada, la división de la muestra introduciría ruido estático. En una realización, una cantidad de ADN pequeña o limitada puede referirse a una cantidad por debajo de 10 pg, entre 10 y 100 pg, entre 100 pg y 1 ng, entre 1 y 10 ng, o entre 10 y 100 ng. Cabe señalar que mientras que este método es especialmente útil en pequeñas cantidades de ADN donde

otros métodos que implican la división en múltiples conjuntos pueden producir problemas significativos relacionados con el ruido estocástico introducido, este método sigue proporcionando la ventaja de minimizar el sesgo cuando se ejecuta en muestras de cualquier cantidad de ADN. En estas situaciones se puede utilizar un paso de pre-amplificación universal para aumentar la cantidad de la muestra en conjunto. De forma ideal, este paso de pre-amplificación no debería alterar apreciablemente las distribuciones alélicas.

En una realización, un método de la presente divulgación puede generar productos de PCR que son específicos para un gran número de loci focalizados, específicamente de 1.000 a 5.000 loci, de 5.000 a 10.000 loci o más de 10.000 loci, para la determinación del genotipo por secuenciación o cualquier otro método de determinación del genotipo, de muestras limitadas como células individuales o ADN de fluidos corporales. Actualmente, la realización de reacciones de PCR multiplexada de más de 5 a 10 dianas presenta un importante reto y a menudo se ve obstaculizada por productos secundarios del cebador, como dímeros de cebadores, y otros artefactos. Cuando se detectan secuencias diana utilizando microarrays con sondas de hibridación, se pueden omitir los dímeros de cebadores y otros artefactos, ya que no son detectados. Sin embargo, cuando se utiliza la secuenciación como un método de detección, la amplia mayoría de las lecturas de secuenciación secuenciaría estos artefactos y no las secuencias diana deseadas en una muestra. Los métodos descritos en la técnica utilizados para multiplexar más de 50 o 100 reacciones en un volumen de reacción seguidos de secuenciación darán como resultado de forma típica más de un 20%, y a menudo más de un 50%, en muchos casos más de un 80% y en algunos casos más de un 90% de lecturas de secuencia no diana.

En general, para realizar una secuenciación focalizada de múltiples dianas (n) de una muestra (superior a 50, superior a 100, superior a 500, o superior a 1.000), se puede dividir la muestra en un número de reacciones paralelas que amplifican una diana individual. Esto se ha realizado en placas multipocillo de PCR o se puede realizar en plataformas comerciales como FLUIDIGM ACCESS ARRAY (48 reacciones por muestra en chips microfluídicos) o DROPLET PCR de RAIN DANCE TECHNOLOGY (de cientos a miles de dianas). Lamentablemente, estos métodos de división y agrupación son problemáticos para las muestras con una cantidad limitada de ADN ya que, a menudo, no hay suficientes copias del genoma para garantizar que hay una copia de cada región del genoma en cada pocillo. Este es un problema especialmente grave cuando los loci polimórficos están focalizados, y las proporciones relativas de los alelos en los loci polimórficos son necesarias, ya que el ruido estocástico introducido por la división y el agrupamiento producirá unas mediciones muy poco precisas de las proporciones de los alelos que estaban presentes en la muestra original de ADN. En el presente documento se describe un método de amplificar de un modo eficaz y efectivo muchas

reacciones de PCR que es aplicable a casos donde sólo está disponible una cantidad limitada de ADN. En una realización, el método se puede

aplicar para el análisis de células individuales, fluidos corporales, mezclas de ADN como el ADN flotante libre que se encuentra en el plasma materno, biopsias, muestras del medio ambiente y/o forenses.

- 5 En una realización, la secuenciación focalizada puede implicar uno, una pluralidad, o todos los pasos siguientes, a) Generar y amplificar una biblioteca con secuencias de adaptador en ambos extremos de fragmentos de ADN, b) Dividir en múltiples reacciones después de la amplificación de la biblioteca, c) Generar y opcionalmente amplificar una biblioteca con secuencias de adaptador en ambos extremos de los fragmentos de ADN, d) Realizar la amplificación de 1000- a 10.000-plex de dianas seleccionadas utilizando un cebador "Directo" específico diana por diana y un cebador específico de etiqueta, e) Realizar una segunda amplificación de este producto utilizando cebadores "Inversos" específicos diana y uno (o más) cebadores específicos para una etiqueta universal que se introdujo como parte de los cebadores directos específicos diana en la primera ronda, f) Realizar una preamplificación de 1000-plex de dianas seleccionadas para un número limitado de ciclos, g) Dividir el producto en múltiples alícuotas y amplificar subconjuntos de dianas en reacciones individuales (por ejemplo, de 50 a 500-plex, aunque esto se puede utilizar hasta un plexado individual. h) Agrupar productos de reacciones de subconjuntos paralelos, i) Durante estas amplificaciones los cebadores pueden llevar etiquetas compatibles de secuenciación (parciales o completas) de tal modo que los productos se puedan secuenciar.

PCR altamente multiplexada

- 20 En el presente documento se divulgan métodos que permiten la amplificación focalizada de más de un centenar a decenas de miles de secuencias diana (por ejemplo, loci SNP) de una muestra de ácido nucleico como ADN genómico obtenido de plasma. La muestra amplificada puede estar relativamente libre de productos de dímeros de cebadores y tener un sesgo alélico bajo en los loci diana. Si durante o después de la amplificación los productos se anexan con adaptadores compatibles de secuenciación, el análisis de estos productos se puede realizar por secuenciación.

- 25 La realización de una amplificación por PCR altamente multiplexada utilizando métodos conocidos en la técnica da como resultado la generación de productos de dímeros de cebadores que superan a los productos de amplificación deseados y no son adecuados para la secuenciación. Esto se puede reducir empíricamente eliminando cebadores que forman estos productos, o realizando la selección in silico de cebadores. Sin embargo, cuanto mayor sea el número de ensayos, más difícil se vuelve el problema.

- 30 Una solución es dividir la reacción 5000-plex en varias amplificaciones con un menor plexado, por ejemplo, reacciones de cien 50-plex o cincuenta 100-plex, o utilizar microfluídicos o incluso dividir la muestra en reacciones PCR individuales. Sin embargo, si el ADN de la muestra es limitado, como en el diagnóstico prenatal no invasivo del plasma de embarazo, se debe evitar dividir la muestra entre múltiples reacciones ya que esto resultará en un cuello de botella.

- 35 En el presente documento se describen métodos para amplificar primero globalmente el ADN de plasma de una muestra y, a continuación, dividir la muestra en múltiples reacciones de enriquecimiento diana multiplexadas con números más moderados de secuencias diana por reacción. En una realización, un método de la presente divulgación se puede utilizar para enriquecer preferentemente una mezcla de ADN en una pluralidad de loci, el método comprende uno o más de los siguientes pasos: la generación y amplificación de una biblioteca a partir de una mezcla de ADN donde las moléculas en la biblioteca tienen secuencias de adaptador ligadas en ambos extremos de los fragmentos de ADN, la división de la biblioteca amplificada en múltiples reacciones, la realización de una primera ronda de amplificación multiplexada de dianas seleccionadas utilizando un cebador "directo" específico diana por diana y uno o una pluralidad de cebadores "inversos" universales específicos de adaptador. En una realización, un método de la presente divulgación incluye además la realización de una segunda amplificación utilizando cebadores específicos diana "inversos" y uno o una pluralidad de cebadores específicos a una etiqueta universal que se introdujo como parte de los cebadores directos específicos diana en la primera ronda. En una realización, el método puede implicar un planteamiento de PCR totalmente anidado, hemi-anidado, semi-anidado, completamente anidado unilateral, hemi-anidado unilateral, o semi-anidado unilateral. En una realización, un método de la presente divulgación se utiliza para enriquecer preferentemente una mezcla de ADN en una pluralidad de loci, el método comprende la realización de una preamplificación de dianas seleccionadas para un número limitado de ciclos, dividiendo el producto en múltiples alícuotas y amplificando subconjuntos de dianas en reacciones individuales, y la agrupación de productos de reacciones de subconjuntos paralelos. Cabe señalar que este planteamiento podría utilizarse para realizar la amplificación focalizada de tal manera que daría como resultado niveles bajos de sesgo alélico para 50-500 loci, para 500 a 5.000 loci, para 5.000 a 50.000 loci, o incluso para 50.000 a 500.000 loci. En una realización, los cebadores llevan etiquetas compatibles de secuenciación parciales o completas.

- 55 El flujo de trabajo puede implicar (1) la extracción de ADN como ADN de plasma, (2) la preparación de la biblioteca de fragmentos con adaptadores universales en ambos extremos de los fragmentos, (3) la amplificación de la biblioteca utilizando cebadores universales específicos para los adaptadores, (4) la división de la "biblioteca" de muestras amplificada en múltiples alícuotas, (5) la realización de amplificaciones multiplexadas (por ejemplo, unos 100-plex, 1.000, o 10.000-plex con un cebador específico diana por diana y un cebador específico de etiqueta) en alícuotas, (6) la agrupación de alícuotas de una muestra, (7) la codificación por medio de código de barras de la muestra, (8) la mezcla de las muestras y el ajuste de la concentración, (9) la secuenciación de la muestra. El flujo de trabajo puede comprender múltiples sub-pasos que contienen uno de los pasos listados (Por ejemplo, el paso (2) de la preparación del paso de la

biblioteca podría implicar tres pasos enzimáticos (final repentino, factor de cola dA y unión de adaptador) y pasos de purificación). Los pasos del flujo de trabajo se pueden combinar, dividir o realizar en un orden diferente (por ejemplo, codificación por medio de barras y agrupación de muestras).

Es importante señalar que la amplificación de una biblioteca se puede realizar de tal manera que esté sesgada para amplificar fragmentos cortos más eficientemente. De esta manera es posible amplificar preferentemente secuencias más cortas, por ejemplo, fragmentos de ADN mono-nucleosomal como el ADN fetal libre de células (de origen placentario) que se encuentra en la circulación de la mujer embarazada. Cabe señalar que los ensayos por PCR pueden tener las etiquetas, por ejemplo, etiquetas de secuenciación, (normalmente una forma truncada de 15-25 bases). Después del multiplexado, los multiplexados de la PCR de una muestra se agrupan y, a continuación, se completan las etiquetas (incluida la codificación por medio de barras) por una PCR específica de etiqueta (también se podría hacer mediante unión). Además, las etiquetas de secuenciación completa se pueden añadir en la misma reacción que el multiplexado. En los primeros ciclos las dianas se pueden amplificar con los cebadores específicos diana, posteriormente los cebadores específicos etiquetados se encargan de completar la secuencia del adaptador SQ. Los cebadores para PCR pueden no llevar etiquetas. Las etiquetas de secuenciación se pueden anexas a los productos de amplificación por unión.

En una realización, la PCR altamente multiplexada seguida de una evaluación de material amplificado por secuenciación clónica se puede utilizar para varias aplicaciones como la detección de aneuploidía fetal. Mientras que las PCR multiplexadas tradicionales evalúan hasta cincuenta loci simultáneamente, el enfoque que se describe en el presente documento se puede utilizar para permitir la evaluación simultánea de más de 50 loci simultáneamente, más de 100 loci simultáneamente, más de 500 loci simultáneamente, más de 1.000 loci simultáneamente, más de 5.000 loci simultáneamente, más de 10.000 loci simultáneamente, más de 50.000 loci simultáneamente, y más de 100.000 loci simultáneamente. Los experimentos han demostrado que se pueden evaluar simultáneamente hasta 10.000 loci distintos incluidos e incluso más, en una sola reacción, con una eficacia y especificidad suficientemente buenas para hacer diagnósticos de aneuploidía prenatal no invasivos y/o determinaciones de número de copia con alta precisión. Los ensayos se pueden combinar en una sola reacción con la totalidad de una muestra como una muestra de ADN libre de células aislada del plasma materno, una fracción de esta, o un derivado todavía más procesado de la muestra de ADN libre de células. La muestra (por ejemplo, ADN libre de células o derivado) también se puede dividir en múltiples reacciones multiplexadas paralelas. La división de la muestra óptima y multiplexada está determinada por la compensación de distintas especificaciones de rendimiento. Debido a la cantidad limitada de material, la división de la muestra en múltiples fracciones puede introducir ruido en el muestreo, tiempo de manipulación, y aumentar la posibilidad de error. Por el contrario, un multiplexado más alto puede dar como resultado una mayor cantidad de amplificación falsa y una mayor desigualdad en la amplificación lo que, en ambos casos, puede reducir el rendimiento de la prueba.

Dos consideraciones cruciales relacionadas en la aplicación de los métodos descritos en el presente documento son la cantidad limitada de la muestra original (por ejemplo, plasma) y el número de moléculas originales en ese material del que se obtiene la frecuencia alélica u otras mediciones. Si el número de moléculas originales cae por debajo de un determinado nivel, el ruido del muestreo aleatorio se vuelve significativo, y puede afectar a la precisión de la prueba. De forma típica, se pueden obtener datos con la calidad suficiente para hacer diagnósticos de aneuploidía prenatal no invasiva si se hacen mediciones en una muestra que comprende el equivalente de 500-1000 moléculas originales por locus diana. Hay varias maneras de aumentar el número de mediciones distintas, por ejemplo, aumentar el volumen de la muestra. Cada manipulación aplicada a la muestra también da como resultado potencialmente pérdidas de material. Es esencial para caracterizar las pérdidas incurridas por varias manipulaciones y evitar o, según sea necesario, mejorar el rendimiento de determinadas manipulaciones evitar las pérdidas que podrían degradar el rendimiento de la prueba.

En una realización, es posible mitigar las pérdidas potenciales en posteriores pasos amplificando la totalidad o una fracción de la muestra original (por ejemplo, la muestra de ADN libre de células). Están disponibles varios métodos para amplificar todo el material genético en una muestra, aumentando la cantidad disponible para procedimientos corriente abajo. En una realización, los fragmentos de ADN de la PCR mediada por enlace (LM-PCR) se amplifican por PCR después de la ligadura de o bien un adaptador distinto, dos adaptadores distintos, o muchos adaptadores distintos. En una realización, la amplificación de desplazamiento múltiple (MDA) pi-29 polimerasa se utiliza para amplificar todo el ADN isotérmicamente. En DOP-PCR y variaciones, el cebado aleatorio se utiliza para amplificar el ADN del material original. Cada método tiene determinadas características como uniformidad de amplificación a través de todas las regiones representadas del genoma, eficiencia de captura y amplificación del ADN original, y rendimiento de la amplificación como una función de la longitud del fragmento.

En una realización LM-PCR se puede utilizar con un adaptador heteroduplexado individual que tiene una tirosina de cebado 3. El adaptador heteroduplexado permite el uso de una molécula de adaptador individual que se puede convertir a dos secuencias distintas en extremos de cebado 5 y de cebado 3 del fragmento de ADN original durante la primera ronda de PCR. En una realización, es posible fraccionar la biblioteca amplificada mediante separaciones de tamaño, o productos como AMPURE, TASS u otros métodos similares. Antes de la ligadura, el ADN de la muestra puede haber finalizado repentinamente y, a continuación, una base de adenosina individual se añade al extremo de cebado 3. Antes de la ligadura, el ADN se puede escindir utilizando una enzima de restricción o cualquier otro método de escisión. Durante la ligadura la adenosina de cebado 3 de los fragmentos de muestra y la proyección de tirosina de cebado 3 complementaria del adaptador pueden mejorar la ligadura de un modo eficaz. El paso de extensión de la amplificación por PCR se puede limitar a partir de un punto de vista temporal para reducir la amplificación de fragmentos con una

longitud superior a unos 200 bp, unos 300 bp, unos 400 bp, unos 500 bp o unos 1.000 bp. Puesto que el ADN más largo que se encuentra en el plasma materno es casi exclusivamente materno, esto puede dar como resultado el enriquecimiento del ADN fetal en un 10-50% y la mejora del rendimiento de la prueba. Se ejecutaron una serie de reacciones utilizando condiciones especificadas por los kits disponibles comercialmente; dieron como resultado una

5 ligadura satisfactoria de menos de un 10% de moléculas de ADN de muestra. Una serie de optimizaciones de las condiciones de la reacción para esto mejoraron la ligadura hasta aproximadamente un 70%.

Mini-PCR

El siguiente método de mini-PCR es deseable para muestras que contienen

10 ácidos nucleicos cortos, ácidos nucleicos digeridos, o ácidos nucleicos fragmentados, como ADN libre de células. Los resultados del diseño del ensayo de PCR tradicional dan como resultado pérdidas significativas de moléculas fetales distintas, pero las pérdidas se pueden reducir en gran medida diseñando ensayos de PCR muy cortos, denominados ensayos de mini-PCR. El ADN libre de células fetal en suero materno está altamente fragmentado y los tamaños de fragmento están distribuidos aproximadamente de un modo gaussiano con una media de 160 bp, una desviación

15 estándar de 15 bp, un tamaño mínimo de unos 100 bp, y un tamaño máximo de unos 220 bp. La distribución de las posiciones de inicio y final del fragmento con respecto a los polimorfismos focalizados, aunque no es necesariamente aleatoria, varía ampliamente entre dianas individuales y entre todas las dianas colectivamente y el punto polimórfico de un locus diana determinado puede ocupar cualquier posición desde el inicio hasta el final entre los diversos fragmentos que se originan de ese locus. Cabe señalar que el término Mini-PCR también puede referirse a una PCR normal sin restricciones o limitaciones adicionales.

20 Durante la PCR, la amplificación sólo se producirá de una plantilla de fragmentos de ADN que comprenden ambos puntos del cebador directo e inverso. Puesto que los fragmentos de ADN libre de células fetal son cortos, la probabilidad de que ambos puntos del cebador estén presentes, la probabilidad de un fragmento fetal de longitud L que comprende tanto el cebador directo como el inverso es el ratio de la longitud del amplicón con respecto a la longitud del fragmento. En condiciones ideales, los ensayos en los que el amplicón es 45, 50, 55, 60, 65, o 70 bp se amplificarán

25 satisfactoriamente a partir de un 72%, 69%, 66%, 63%, 59%, o 56%, respectivamente, de las plantillas de fragmentos de moléculas disponibles. La longitud de amplicón es la distancia entre los extremos de cebado 5 de los puntos de cebado directo e inverso. La longitud de amplicón que es más corta que la que se utiliza de forma típica para el experto en la técnica puede dar como resultado mediciones más eficaces de los loci polimórficos deseados al requerir únicamente lecturas de secuencia cortas. En algunas realizaciones, una fracción sustancial de los amplicones debería ser inferior a

30 100 pb, inferior a 90 pb, inferior a 80 pb, inferior a 70 pb, inferior a 65 pb, inferior a 60 pb, inferior a 55 pb, inferior a 50 pb, o inferior a 45 pb.

Cabe señalar que en métodos conocidos en la técnica, los ensayos cortos como los que se describen en el presente documento normalmente se evitan ya que no se requieren e imponen unas restricciones considerables al diseño del cebador al limitar la longitud del cebador, las características de reformación térmica y la distancia entre el cebador

35 directo y el inverso.

También cabe señalar que existe el potencial para una amplificación sesgada si el extremo de cebado 3 de cualquiera de los cebadores está dentro de aproximadamente 1-6 bases del punto polimórfico. Esta diferencia de base individual en el punto de unión de polimerasa inicial puede dar como resultado una amplificación preferencial de un alelo, que puede alterar las frecuencias de alelo observadas y degradar el rendimiento. Todas estas restricciones hacen que sea

40 todo un desafío identificar los cebadores que amplificarán un locus determinado satisfactoriamente y, además, diseñar grandes conjuntos de cebadores que sean compatibles en la misma reacción multiplexada. En una realización, el extremo 3' de los cebadores directos e inversos interiores está diseñado para hibridarse a una región del ADN corriente arriba del punto polimórfico, y separada del punto polimórfico por un pequeño número de bases. De forma ideal, el número de bases puede estar entre 6

45 y 10 bases, pero igualmente puede estar entre 4 y 15 bases, entre tres y 20 bases, entre dos y 30 bases, o entre 1 y 60 bases, y conseguir prácticamente el mismo fin.

La PCR multiplexada puede implicar una sola ronda de PCR en la que todas las dianas son amplificadas o puede implicar una ronda de PCR seguida de una o más rondas de PCR anidada o alguna variante de PCR anidada. La PCR anidada se compone de una ronda o rondas posteriores de amplificación por PCR utilizando uno o más cebadores

50 nuevos que se unen internamente, por al menos un par de bases, a los cebadores utilizados en una ronda previa. La PCR anidada reduce el número de dianas de amplificación falsas al amplificar, en reacciones posteriores, sólo aquellos productos de amplificación de la anterior que tienen la secuencia interna correcta. La reducción de dianas de amplificación falsas mejora el número de mediciones útiles que se pueden obtener, especialmente en la secuenciación. La PCR anidada de forma típica implica el diseño de cebadores completamente internos con respecto a los anteriores

55 puntos de unión del cebador, aumentando necesariamente el tamaño del segmento de ADN necesario para la amplificación. Para muestras, por ejemplo, de ADN libre de células de plasma materno, en las que el ADN está altamente fragmentado, el tamaño de ensayo mayor reduce el número de moléculas de ADN libre de células distintas de las que se puede obtener una medición. En una realización, para contrarrestar este efecto, se puede utilizar un planteamiento de anidado parcial donde uno o ambos de los cebadores de la segunda ronda se solapan con los

60 primeros puntos de unión extendiendo internamente algunos números de bases para conseguir una especificidad adicional a la vez que aumentan mínimamente en el tamaño del ensayo total.

En una realización, un conjunto multiplexado de ensayos por PCR está diseñado para amplificar el SNP potencialmente heterocigotou otros loci polimórficos o no polimórficos en uno o más cromosomas y estos ensayos se utilizan en una sola reacción para amplificar ADN. El número de ensayos por PCR puede estar entre 50 y 200 ensayos por PCR, entre 200 y 1.000 ensayos por PCR, entre 1.000 y 5.000 ensayos por PCR, o entre 5.000 y 20.000 ensayos por PCR (de 50 a 200-plex, de 200 a 1.000-plex, de 1.000 a 5.000-plex, de 5.000 a 20.000-plex, más de 20.000-plex respectivamente). En una realización, un conjunto multiplexado de unos 10.000 ensayos por PCR (10.000-plex) está diseñado para amplificar loci de SNP potencialmente heterocigóticos en los cromosomas X, Y, 13, 18, y 21 y 1 o 2 y estos ensayos se utilizan en una sola reacción para amplificar ADN libre de células obtenido a partir de una muestra de plasma material, muestras de vellosidades coriónicas, muestras de amniocentesis, células individuales o un pequeño número de células, otros fluidos corporales o tejidos, cánceres, u otras materia genética. Las frecuencias de SNP de cada locus se pueden determinar mediante secuenciación clónica o cualquier otro método de secuenciación de los amplicones. Los análisis estadísticos de las distribuciones de la frecuencia alélica o ratios de todos los ensayos se pueden utilizar para determinar si la muestra contiene una trisomía de uno o más de los cromosomas incluidos en la prueba. En otra realización las muestras de ADN libre de células original se dividen en dos muestras y se realizan ensayos de 5.000-plex paralelos. En otra realización las muestras de ADN libre de células original se dividen en muestras n y se realizan ensayos (~10.000/n)-plex paralelos donde n está entre 2 y 12, o entre 12 y 24, o entre 24 y 48, o entre 48 y 96. Los datos se recogen y analizan de forma similar a la que ya se ha descrito. Cabe señalar que este método también se puede aplicar a la detección de translocaciones, deleciones, duplicaciones, y otras anomalías cromosómicas.

En una realización, las colas que no tienen homología con el genoma diana también se pueden añadir al extremo de cebado 3 o al extremo de cebado 5 de cualquiera de los cebadores. Estas colas facilitan las posteriores manipulaciones, procedimientos, o mediciones. En una realización, la secuencia de cola puede ser la misma para los cebadores específicos diana directos e inversos. En una realización, se pueden utilizar diferentes colas para los cebadores específicos diana directos e inversos. En una realización, una pluralidad de colas diferentes se puede utilizar para loci o conjuntos de loci diferentes. Algunas colas se pueden compartir entre todos los loci o entre subconjuntos de loci. Por ejemplo, utilizando colas directas e inversas que correspondan a secuencias directas e inversas requeridas por cualquiera de las plataformas de secuenciación actuales se puede permitir la secuenciación directa después de la amplificación. En una realización, las colas se pueden utilizar como puntos de cebado comunes entre todas las dianas amplificadas que se pueden utilizar para añadir otras secuencias útiles. En algunas realizaciones, los cebadores interiores pueden contener una región que está diseñada para que se hibride o bien corriente arriba o corriente abajo del locus focalizado (por ejemplo, un locus polimórfico). En algunas realizaciones, los cebadores pueden contener un código de barras molecular. En algunas realizaciones, el cebador puede contener una secuencia de cebado universal diseñada para permitir la amplificación por PCR.

En una realización, se crea un conjunto de ensayo de PCR de 10.000-plex de modo que los cebadores directo e inverso tengan colas que correspondan a las secuencias directas e inversas requeridas por un instrumento de secuenciación de alto rendimiento como HISEQ, GAIIIX, o MYSEQ comercializado por ILLUMINA. Además, el cebador incluido en las colas de secuenciación es una secuencia adicional que se puede utilizar como un punto de cebado en una PCR posterior para añadir secuencias de código de barras nucleótido a los amplicones, permitiendo la secuenciación multiplexada de múltiples muestras en una sola línea del instrumento de secuenciación de alto rendimiento.

En una realización, se crea un conjunto de ensayo de PCR de 10.000-plex de modo que los cebadores inversos tengan colas que correspondan a las secuencias inversas requeridas por un instrumento de secuenciación de alto rendimiento. Después de la amplificación con el primer ensayo de 10.000-plex, se puede realizar una amplificación por PCR posterior utilizando otro conjunto de 10.000-plex que tenga cebadores directos parcialmente anidados (por ejemplo, 6-bases anidadas) para todas las dianas y un cebador inverso que corresponde a la cola de la secuenciación inversa incluida en la primera ronda. Esta ronda posterior de amplificación parcialmente anidada con solo un cebador específico diana y un cebador universal limita el tamaño requerido del ensayo, reduciendo el ruido del muestreo, aunque reduce en gran medida el número de amplicones falsos.

Las etiquetas de secuenciación se pueden añadir a los adaptadores de unión anexados y/o como parte de sondas para PCR, de modo que la etiqueta es parte del amplicón final. La fracción fetal afecta al rendimiento de la prueba. Hay varias maneras de enriquecer la fracción fetal del ADN que se encuentra en el plasma materno. La fracción fetal se puede aumentar por el método LM-PCR anteriormente descrito y del que también ya se ha hablado mediante una extracción focalizada de fragmentos maternos largos. En una realización, antes de la amplificación por PCR multiplexada de los loci diana, se puede llevar a cabo una reacción por PCR multiplexada adicional para extraer selectivamente fragmentos largos y en su mayor parte maternos correspondientes a los loci focalizados en la PCR multiplexada posterior. Los cebadores adicionales están diseñados para reformar térmicamente un punto a una mayor distancia del polimorfismo que está previsto que esté presente entre fragmentos de ADN fetal libre de células. Estos cebadores se pueden utilizar en una reacción por PCR multiplexada de un ciclo antes de la PCR multiplexada de los loci polimórficos diana. Estos cebadores distales se etiquetan con una molécula o mitad que permite el reconocimiento selectivo de los pedazos de ADN etiquetados. En una realización, estas moléculas de ADN pueden modificarse covalentemente con una molécula biotina que permite la extracción de un ADN de doble cadena formado recientemente que comprende estos cebadores después de un ciclo de PCR. El ADN de doble cadena formado durante esa primera ronda es probable que sea de origen materno. La extracción del material híbrido se puede conseguir mediante el uso de perlas de estreptavidina magnéticas. Hay otros métodos de etiquetado que pueden funcionar igualmente bien. En una realización, se pueden utilizar métodos de selección de tamaño para enriquecer la muestra para hebras de ADN más

cortas; por ejemplo, aquellas que sean inferiores a 800 bp, inferiores a unos 500 bp, o inferiores a unos 300 bp. La amplificación de fragmentos cortos, a continuación, puede continuar como de costumbre.

El método de mini-PCR descrito en esta divulgación permite la amplificación altamente multiplexada y el análisis de cientos a miles o incluso millones de loci en una sola reacción, a partir de una sola muestra. En el mismo, la detección del ADN amplificado se puede multiplexar; de decenas a cientos de muestras se pueden multiplexar en una línea de secuenciación utilizando la PCR de código de barras. Esta detección multiplexada se ha probado satisfactoriamente hasta 49-plex, y es posible un grado de multiplexado mucho más alto. En efecto, esto permite genotipificar cientos de muestras a miles de SNP en una sola secuenciación. Para estas muestras, el método permite la determinación del genotipo y la tasa de heterocigosidad y simultáneamente la determinación del número de copia, ambas pueden ser utilizadas para la detección de la aneuploidía. Este método es especialmente útil en la detección de la aneuploidía de un feto en gestación a partir del ADN flotante libre que se encuentra en el plasma materno. Este método se puede utilizar como parte de un método para determinar el sexo de un feto, y/o predecir la paternidad del feto. Se puede utilizar como parte de un método para la dosificación de mutaciones. Este método se puede utilizar para cualquier cantidad de ADN o ARN, y las regiones focalizadas pueden ser SNP, otras regiones polimórficas, regiones no polimórficas, y combinaciones de estas.

En algunas realizaciones, se puede utilizar la amplificación por PCR universal mediada por unión de ADN. La amplificación por PCR universal mediada por unión se puede utilizar para amplificar ADN del plasma que, a continuación, se puede dividir en múltiples reacciones paralelas. También se puede utilizar para amplificar preferentemente fragmentos cortos, por tanto, enriqueciendo la fracción fetal. En algunas realizaciones la adición de etiquetas a los fragmentos mediante unión puede permitir la detección de fragmentos más cortos, el uso de porciones específicas de secuencia diana más cortas de los cebadores y/o la reformación térmica a temperaturas más altas que reduce las reacciones no especificadas.

Los métodos descritos en la presente memoria se pueden utilizar para una serie de propósitos donde hay un conjunto diana de ADN que se mezcla con una cantidad de ADN contaminante. En algunas realizaciones, el ADN diana y el ADN contaminante pueden ser de individuos que están genéticamente relacionados. Por ejemplo, se pueden detectar anomalías genéticas en un feto (diana) a partir de plasma materno que contiene ADN fetal (diana) y también ADN materno (contaminante); las anomalías incluyen anomalías en todo el cromosoma (por ejemplo, aneuploidía), anomalías parciales en el cromosoma (por ejemplo, deleciones, duplicaciones, inversiones, translocaciones), polimorfismos polinucleótidos (por ejemplo, STR), polimorfismos de un solo nucleótido, y/u otras anomalías genéticas o diferencias. En algunas realizaciones, el ADN contaminante y la diana pueden ser del mismo individuo, pero donde el ADN contaminante y el diana son diferentes por una o más mutaciones, por ejemplo, en el caso de cáncer, (véase por ejemplo, H. Mamon et al. Preferential Amplification of Apoptotic ADN from Plasma: Potential for Enhancing Detección of Minor ADN Alterations in Circulating ADN. *Clinical Chemistry* 54:9 (2008)). En algunas realizaciones, el ADN se puede encontrar en un sobrenadante (apoptótico) del cultivo celular. En algunas realizaciones, es posible inducir la apoptosis en muestras biológicas (por ejemplo, sangre) para la posterior preparación, amplificación y/o secuenciación de la biblioteca. Un número de flujos de trabajo y protocolos de apoyo para conseguir este objetivo se presentan en otra parte de esta divulgación.

En algunas realizaciones, el ADN diana puede originarse a partir de células individuales, de muestras de ADN que se componen de menos de una copia del genoma diana, de cantidades bajas de ADN, de ADN de origen mixto (por ejemplo, plasma del embarazo: ADN placentario y materno; plasma de pacientes con cáncer y tumores: una mezcla entre ADN sano y con cáncer, trasplante, etc), de otros fluidos corporales, de cultivos celulares, de sobrenadantes del cultivo, de muestras forenses de ADN, de muestras antiguas de ADN (por ejemplo, insectos atrapados en ámbar), de otras muestras de ADN, y combinaciones de estas.

En algunas realizaciones, se puede utilizar un tamaño de amplicón corto. Los tamaños de amplicón corto son especialmente adecuados para ADN fragmentado (véase por ejemplo, A. Sikora, et al. Detection of increased amounts of cell-free fetal DNA with short PCR amplicons. *Clin Chem.* 2010 Jan;56(1): 136-8.) El uso de tamaños de amplicón corto puede producir algunas ventajas significativas. Los tamaños de amplicón corto pueden producir una eficacia de la amplificación optimizada. Los tamaños de amplicón corto de forma típica producen productos más cortos, por tanto, hay una menor probabilidad de un cebado no específico. Los productos más cortos se pueden agrupar más densamente en células de flujo de secuenciación, ya que las agrupaciones serán más pequeñas. Cabe señalar que los métodos descritos en el presente documento pueden funcionar igualmente bien para amplicones por PCR más largos. La longitud del amplicón se puede aumentar si es necesario, por ejemplo, cuando se secuencian tramos más largos. Se realizaron experimentos con amplificación focalizada de 146-plex con ensayos con una longitud de 100 bp a 200 bp como una primera fase en un protocolo para PCR anidada en células individuales y en ADN genómico con resultados positivos.

En algunas realizaciones, los métodos descritos en el presente documento se pueden utilizar para amplificar y/o detectar SNP, número de copia, metilación nucleótida, niveles de mRNA, otros tipos de niveles de expresión de ARN, otras características genéticas y/o epigenéticas. Los métodos de mini-PCR descritos en el presente documento se pueden utilizar junto con la secuenciación de nueva generación; se puede utilizar con otros métodos corriente abajo como microarrays, recuento por PCR digital, PCR en tiempo real, análisis por espectrometría de masas, etc.

En alguna realización, los métodos de amplificación por mini-PCR descritos en el presente documento se pueden utilizar como parte de un método para la cuantificación precisa de poblaciones minoritarias. Se puede utilizar para la cuantificación absoluta utilizando calibradores de patrón. Se puede utilizar para la cuantificación alélica menor /mutación

a través de una secuenciación muy profunda, y se puede ejecutar de una manera altamente multiplexada. Se puede utilizar para pruebas de identidad y paternidad estándar de familiares o antepasados, en humanos, animales, plantas u otras criaturas. Se puede utilizar para pruebas forenses. Se puede utilizar para la determinación rápida del genotipo y el análisis del número de copia (CN), en cualquier clase de material, por ejemplo, líquido amniótico y CVS, espermatozoides, producto de concepción (POC). Se puede utilizar para el análisis de una sola célula, como determinación del genotipo en muestras a las que se ha hecho una biopsia de embriones. Se puede utilizar para el análisis rápido del embrión (en menos de un día, uno o dos días de biopsia) mediante secuenciación focalizada utilizando min-PCR.

En algunas realizaciones se puede utilizar para el análisis de tumores: las biopsias de tumores a menudo son una mezcla de células sanas y células con un tumor. El PCR dirigido permite la secuenciación profunda de SNP y loci con casi ninguna secuencia de fondo. Se puede utilizar para el análisis del número de copia y la pérdida de heterocigosidad en el ADN del tumor. Dicho ADN del tumor puede estar presente en muchos fluidos corporales diferentes o tejidos de pacientes con un tumor. Se puede utilizar para la detección de la recurrencia del tumor, y/o el examen del tumor. Se puede utilizar para las pruebas de control de calidad de semillas. Cabe señalar que cualquiera de estos métodos podría utilizarse igualmente para focalizar loci no polimórficos con el fin de realizar una determinación de ploidía.

Algunos ejemplos de la bibliografía que describen alguno de estos métodos fundamentales que subyacen en los métodos divulgados en el presente documento incluyen: (1) Wang HY, Luo M, Tereshchenko IV, Friker DM, Cui X, Li JY, Hu G, Chu Y, Azaro MA, Lin Y, Shen L, Yang Q, Kambouris ME, Gao R, Shih W, Li H. *Genoma Res.* 2005 Feb;15(2):276-83. Department of Molecular Genetics, Microbiology and Immunology/The Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, New Jersey 08903, USA. (2) High-throughput genotyping of single nucleotide polymorphisms with high sensitivity. Li H, Wang HY, Cui X, Luo M, Hu G, Greenawalt DM, Tereshchenko IV, Li JY, Chu Y, Gao R. *Methods Mol Biol.* 2007;396 - PubMed PMID: 18025699. (3) Un método que comprende el multiplexado de un promedio de 9 ensayos para secuenciación se describe en: Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. Varley KE, Mitra RD. *Genoma Res.* 18 de noviembre de 2008 (I 1): 1844-50. Epub 10 de octubre de 2008. Cabe señalar que los métodos divulgados en el presente documento permiten el multiplexado de órdenes de magnitud superiores a las referencias anteriores.

Variantes de PCR focalizada - Anidado

Hay muchos flujos de trabajo posibles cuando se lleva a cabo una PCR; se describen algunos flujos de trabajo típicos de los métodos divulgados en el presente documento. Los pasos descritos en el presente documento no pretenden excluir otros posibles pasos ni implica que cualquiera de los pasos descritos en el presente documento sean necesarios para que el método funcione correctamente. Se conocen una gran cantidad de variaciones de parámetros u otras modificaciones en la bibliografía, y se pueden realizar sin afectar a la esencia de la invención. Un flujo de trabajo generalizado particular se muestra a continuación seguido de un número de posibles variantes. Las variantes de forma típica hacen referencia a posibles reacciones de PCR secundarias, por ejemplo, diferentes tipos de anidado que se pueden realizar (paso 3). Es importante señalar que las variantes se pueden realizar en momentos diferentes, o en un orden distinto al que explícitamente se describe en el presente documento. Los ejemplos ilustrativos que utilizan loci polimórficos se pueden adaptar rápidamente para la amplificación de loci no polimórficos si se desea.

1. El ADN en la muestra puede tener adaptadores de unión, a los que a menudo se hace referencia como etiquetas de biblioteca o etiquetas de adaptador de unión (LTs), anexados, donde los adaptadores de unión contienen una secuencia de cebado universal, seguida de una amplificación universal. En una realización, esto se puede realizar utilizando un protocolo estándar diseñado para crear bibliotecas de secuenciación después de la fragmentación. En una realización, la muestra de ADN puede haber finalizado repentinamente y, a continuación, se puede añadir una A en el extremo 3'. Se puede añadir y unir un adaptador Y con una proyección T. En algunas realizaciones, se pueden utilizar otros extremos pegajosos que no sean una proyección A o T. En algunas realizaciones, se pueden añadir otros adaptadores, por ejemplo, adaptadores de unión de bucle. En algunas realizaciones, los adaptadores pueden tener una etiqueta diseñada para la amplificación por PCR.

2. Amplificación diana específica (STA): La pre-amplificación de cientos a miles a decenas de miles e incluso de cientos de miles de dianas se pueden multiplexar en un volumen de reacción. La STA se ejecuta de forma típica de 10 a 30 ciclos, aunque también se puede ejecutar de 5 a 40 ciclos, de 2 a 50 ciclos, e incluso de 1 a 100 ciclos. Los cebadores pueden tener una cola, por ejemplo, para un flujo de trabajo más sencillo o para evitar la secuenciación de una gran proporción de dímeros. Cabe señalar que de forma típica, los dímeros de ambos cebadores que llevan la misma etiqueta no se amplificarán o secuenciarán eficientemente. En algunas realizaciones, se pueden llevar a cabo entre 1 y 10 ciclos de PCR; en algunas realizaciones, se pueden llevar a cabo entre 10 y 20 ciclos de PCR; en algunas realizaciones, se pueden llevar a cabo entre 20 y 30 ciclos de PCR; en algunas realizaciones, se pueden llevar a cabo entre 30 y 40 ciclos de PCR; en algunas realizaciones se pueden llevar a cabo más de 40 ciclos de PCR. La amplificación puede ser una amplificación lineal. El número de ciclos de PCR se puede optimizar para dar como resultado un perfil de profundidad de lectura (DOR) óptimo. Perfiles DOR diferentes pueden ser deseables para fines distintos. En algunas realizaciones, es deseable una distribución más uniforme de las lecturas entre todos los ensayos; Si el DOR es demasiado pequeño para algunos ensayos, el ruido estocástico también puede ser demasiado alto para los datos para ser demasiado útil, mientras que si la profundidad de lectura es demasiado alta, la utilidad marginal de cada lectura adicional es relativamente pequeña.

Las colas del cebador pueden mejorar la detección de ADN fragmentado de bibliotecas etiquetadas universalmente. Si la etiqueta de biblioteca y las colas del cebador contienen una secuencia homóloga, se puede mejorar la hibridación (por

- ejemplo, la temperatura de fusión (T_M) baja) y los cebadores se pueden extender si sólo una parte de la secuencia diana del cebador está en el fragmento de ADN de la muestra. En algunas realizaciones, se pueden utilizar 13 o más pares de bases específicos diana. En algunas realizaciones, se pueden utilizar de 10 a 12 pares de bases específicos diana. En algunas realizaciones, se pueden utilizar de 8 a 9 pares de bases específicos diana. En algunas realizaciones, se pueden utilizar de 6 a 7 pares de bases específicos diana. En algunas realizaciones, se puede realizar la STA en ADN preamplificado, por ejemplo, MDA, RCA, otras amplificaciones de todo el genoma, o PCR universal mediada por adaptador. En algunas realizaciones, se puede realizar la STA en muestras que están enriquecidas o desprovistas de determinadas secuencias y poblaciones, por ejemplo, por selección de tamaño, captura diana, degradación dirigida.
3. En algunas realizaciones, es posible realizar PCR multiplexadas secundarias o reacciones de extensión del cebador para aumentar la especificidad y reducir los productos no deseados. Por ejemplo, el anidado total, el semi-anidado, el hemi-anidado, y/o la subdivisión en reacciones paralelas de conjuntos de ensayo más pequeños son todo técnicas que se pueden utilizar para aumentar la especificidad. Los experimentos han demostrado que dividir una muestra en tres reacciones de 400-plex dio como resultado un ADN de producto con una mayor especificidad que una reacción de 1.200-plex con exactamente los mismos cebadores. De forma similar, los experimentos han demostrado que dividir una muestra en cuatro reacciones de 2.400-plex ha dado como resultado un ADN de producto con mayor especificidad que una reacción de 9.600-plex con exactamente los mismos cebadores. En una realización, es posible utilizar cebadores específicos diana y específicos de etiqueta de la misma direccionalidad y de una direccionalidad opuesta.
4. En algunas realizaciones, es posible amplificar una muestra de ADN (dilución, purificada u otra) producida por una reacción de STA utilizando cebadores específicos etiquetados y "amplificación universal", es decir, amplificar muchas o todas las dianas preamplificadas y etiquetadas. Los cebadores pueden contener secuencias funcionales adicionales, por ejemplo, códigos de barras, o una secuencia de adaptador completa necesaria para la secuenciación en una plataforma de secuenciación de alto rendimiento.
- Estos métodos se pueden utilizar para el análisis de cualquier muestra de ADN, y son especialmente útiles cuando la muestra de ADN es especialmente pequeña, o cuando es una muestra de ADN donde el ADN se origina a partir de más de un individuo, como en el caso del plasma materno. Estos métodos se pueden utilizar en muestras de ADN como una sola célula o un pequeño número de ellas, ADN genómico, ADN del plasma, bibliotecas de plasma amplificado, bibliotecas sobrenadantes apoptóticas amplificadas, u otras muestras de ADN mixto. En una realización, estos métodos se pueden utilizar en el caso en el que las células con una constitución genética diferente pueden estar presentes en un solo individuo, como con el cáncer o los trasplantes.
- Variantes del protocolo (variantes y/o añadidos al flujo de trabajo que se indica más arriba)
- Mini-PCR multiplexada directa: La amplificación diana específica (STA) de una pluralidad de secuencias diana con cebadores con etiquetas que se muestra en la Figura 1, 101 denota ADN de doble cadena con un locus polimórfico de interés en X. 102 denota el ADN de doble cadena con adaptadores de unión añadidos para amplificación universal. 103 denota el ADN de una sola cadena que ha sido amplificado universalmente con cebadores para PCR hibridados. 104 denota el producto de la PCR final. En algunas realizaciones, la STA se puede realizar en más de 100, más de 200, más de 500, más de 1.000, más de 2.000, más de 5.000, más de 10.000, más de 20.000, más de 100.000 o más de 200.000 dianas. En una reacción posterior, los cebadores específicos etiquetados amplifican todas las secuencias diana y alargan las etiquetas para incluir todas las secuencias necesarias para la secuenciación, incluyendo los índices de la muestra. En una realización, es posible que los cebadores no estén etiquetados o que sólo determinados cebadores puedan estar etiquetados. Se pueden añadir adaptadores de secuenciación mediante la unión de adaptador convencional. En una realización, los cebadores iniciales pueden llevar las etiquetas.
- En una realización, los cebadores están diseñados para que la longitud del ADN amplificado sea inesperadamente corta. La técnica anterior demuestra que los expertos en la técnica de forma típica diseñan 100+ bp amplicones. En una realización, los amplicones pueden estar diseñados para ser inferiores a 80 bp. En una realización, los amplicones pueden estar diseñados para ser inferiores a 70 bp. En una realización, los amplicones pueden estar diseñados para ser inferiores a 60 bp. En una realización, los amplicones pueden estar diseñados para ser inferiores a 50 bp. En una realización, los amplicones pueden estar diseñados para ser inferiores a 45 bp. En una realización, los amplicones pueden estar diseñados para ser inferiores a 40 bp. En una realización, los amplicones pueden estar diseñados para ser inferiores a 35 bp. En una realización, los amplicones pueden estar diseñados para estar entre 40 y 65 bp.
- Se llevó a cabo un experimento utilizando este protocolo utilizando una amplificación de 1200-plex. Se utilizó tanto ADN genómico como plasma del embarazo; un 70% de lecturas de secuencia correlacionadas con secuencias focalizadas. Se proporcionan detalles en otra parte de este documento. La secuenciación de un 1042-plex sin diseño y la selección de ensayos dio como resultado >99% de secuencias que eran productos de dímeros de cebadores.
- PCR secuencial: Después de la STA1 se pueden amplificar en paralelo múltiples alícuotas del producto con conjuntos de complejidad reducida con los mismos cebadores. La primera amplificación puede dar suficiente material para dividir. Este método es especialmente bueno para pequeñas muestras, por ejemplo, aquellas que tienen unos 6-100 pg, unos 100 pg a 1 ng, de unos 1 ng a 10 ng, o de unos 10 ng a 100 ng. El protocolo se llevó a cabo con 1200-plex en tres 400-plexes. La correspondencia de lecturas de secuenciación aumentó de alrededor de un 60 a un 70 % sólo en los 1200-plex hasta más de un 95%.
- Mini-PCR semianidada: (véase la Figura 2) Después de la STA 1 se realiza una segunda STA que comprende un conjunto multiplexado de cebadores directos anidados internos (103 B, 105 b) y uno (o unos pocos) cebadores inversos

- específicos etiquetados (103 A). 101 denota ADN de doble cadena con un locus polimórfico de interés en X. 102 denota el ADN de doble cadena con adaptadores de unión añadidos para amplificación universal. 103 denota el ADN de una sola cadena que ha sido amplificado universalmente con el cebador directo B y el cebador inverso A hibridados. 104 denota el producto de la PCR de 103. 105 denota el producto de 104 con el cebador directo anidado b hibridado, y la etiqueta inversa A ya parte de la molécula de la PCR que se produjo entre 103 y 104. 106 denota el producto de la PCR final. Con este flujo de trabajo normalmente más del 95% de secuencias se corresponden con las dianas previstas. El cebador anidado puede solaparse con la secuencia del cebador directo exterior pero introduce bases de extremo 3' adicionales. En algunas realizaciones es posible utilizar entre una y 20 bases 3' adicionales. Los experimentos han demostrado que utilizar 9 o más bases 3' adicionales en un diseño de 1200-plex funciona bien.
- 10 Mini-PCR completamente anidada: (véase la Figura 3) Después del paso 1 de la STA, es posible realizar una segunda PCR multiplexada (o m.p. paralela Las PCR de complejidad reducida) con dos cebadores anidados que llevan etiquetas (A, a, B, b). 101 denota ADN de doble cadena con un locus polimórfico de interés en X. 102 denota el ADN de doble cadena con adaptadores de unión añadidos para amplificación universal. 103 denota el ADN de una sola cadena que ha sido amplificado universalmente con el cebador directo B y el cebador inverso A hibridados. 104 denota el producto de la PCR de 103. 105 denota el producto de 104 con el cebador directo anidado b y el cebador inverso anidado a hibridados. 106 denota el producto de la PCR final. En algunas realizaciones es posible utilizar dos conjuntos completos de cebadores. Los experimentos en los que se utiliza un protocolo de mini-PCR completamente anidada se utilizaron para realizar una amplificación de 146-plex en una sola célula y en tres células sin el paso 102 de anexar adaptadores de unión universal y amplificar.
- 15 20 Mini-PCR hemi-anidada: (véase la Figura 4) Es posible utilizar ADN diana que tenga un adaptador en los extremos del fragmento. La STA se realiza comprendiendo un conjunto multiplexado de cebadores directos (B) y uno (o unos pocos) cebadores inversos específicos etiquetados (A). Se puede realizar una segunda STA utilizando un cebador inverso específico etiquetado y un cebador directo específico etiquetado universal. 101 denota ADN de doble cadena con un locus polimórfico de interés en X. 102 denota el ADN de doble cadena con adaptadores de unión añadidos para amplificación universal. 103 denota el ADN de una sola cadena que ha sido amplificado universalmente con cebador inverso A hibridado. 104 denota el producto de la PCR de 103 que se amplificó utilizando el cebador inverso A y el cebador de etiqueta del adaptador de unión LT. 105 denota el producto de 104 con el cebador directo B hibridado. 100.0 denota el producto de la PCR final. En este flujo de trabajo, los cebadores directos e inversos específicos diana se utilizan en reacciones separadas, por tanto, reducen la complejidad de la reacción y evitan la formación de dímeros de los cebadores directos e inversos. Cabe señalar que en este ejemplo, los cebadores A y B se pueden considerar como primeros cebadores, y los cebadores 'a' y 'b' se pueden considerar como cebadores interiores. Este método supone una gran mejora sobre la PCR directa ya que es tan bueno como la PCR directa pero evita los dímeros de cebadores. Después de la primera ronda de protocolo hemi-anidado de forma típica se ve un -99% de ADN no focalizado, sin embargo, después de la segunda ronda, de forma típica, hay una gran mejora.
- 25 30 35 Mini-PCR triplemente hemi-anidada: (véase la **Figura 5**) Es posible utilizar ADN diana que tenga un adaptador en los extremos del fragmento. La STA se realiza comprendiendo un conjunto multiplexado de cebadores directos (B) y uno (o unos pocos) cebadores inversos específicos etiquetados (A) y (a). Se puede realizar una segunda STA utilizando un cebador inverso específico etiquetado y un cebador directo específico etiquetado universal. 101 denota ADN de doble cadena con un locus polimórfico de interés en X. 102 denota el ADN de doble cadena con adaptadores de unión añadidos para amplificación universal. 103 denota el ADN de una sola cadena que ha sido amplificado universalmente con cebador inverso A hibridado. 104 denota el producto de la PCR de 103 que se amplificó utilizando el cebador inverso A y el cebador de etiqueta del adaptador de unión LT. 105 denota el producto de 104 con el cebador directo B hibridado. 106 denota el producto de la PCR de 105 que se amplificó utilizando el cebador inverso A y el cebador directo B. 107 denota el producto de 106 con el cebador inverso 'a' hibridado. 108 denota el producto de la PCR final. Cabe señalar que en este ejemplo, los cebadores 'a' y B se pueden considerar como cebadores interiores, y A se puede considerar como un primer cebador. De forma opcional, tanto A como B se pueden considerar como primeros cebadores, y 'a' se puede considerar como un cebador interior. La designación de cebadores inversos y directos se puede cambiar. En este flujo de trabajo, los cebadores directos e inversos específicos diana se utilizan en reacciones separadas, por tanto, reducen la complejidad de la reacción y evitan la formación de dímeros de los cebadores directos e inversos. Este método supone una gran mejora sobre la PCR directa ya que es tan bueno como la PCR directa pero evita los dímeros de cebadores. Después de la primera ronda de protocolo hemi-anidado de forma típica se ve un -99% de ADN no focalizado, sin embargo, después de la segunda ronda, de forma típica, hay una gran mejora.
- 40 45 50 55 Mini-PCR anidada unilateral: (véase la Figura 6) Es posible utilizar ADN diana que tenga un adaptador en los extremos del fragmento. La STA también se puede realizar con un conjunto multiplexado de cebadores directos anidados y utilizando la etiqueta del adaptador de unión como el cebador inverso. A continuación, se puede realizar una segunda STA utilizando un conjunto de cebadores directos anidados y un cebador inverso universal. 101 denota ADN de doble cadena con un locus polimórfico de interés en X. 102 denota el ADN de doble cadena con adaptadores de unión añadidos para amplificación universal. 103 denota el ADN de una sola cadena que ha sido amplificado universalmente con cebador directo A hibridado. 104 denota el producto de la PCR de 103 que se amplificó utilizando el cebador directo A y el cebador inverso de etiqueta del adaptador de unión LT. 105 denota el producto de 104 con el cebador directo anidado a hibridado. 106 denota el producto de la PCR final. Este método puede detectar secuencias diana más cortas que la PCR estándar utilizando cebadores que se solapan en el primer y el segundo STA. El método se realiza de forma típica fuera de una muestra de ADN que ya ha pasado por el paso 1 de la STA indicado anteriormente, el anexo de
- 60

etiquetas universales y la amplificación; los dos cebadores anidados están sólo en un lado, el otro lado utiliza la etiqueta de biblioteca. El método se ha realizado en bibliotecas de sobrenadantes apoptóticos y plasma del embarazo. Con este flujo de trabajo un 60% de las secuencias se correspondieron con las dianas previstas. Cabe señalar que las lecturas que contenían la secuencia de adaptador inverso no se han correspondido, de modo que se espera que este número sea más alto si esas lecturas que contienen la secuencia de adaptador inverso se corresponden.

Mini-PCR unilateral: Es posible utilizar ADN diana que tenga un adaptador en los extremos del fragmento (véase la Figura 7). La STA se puede realizar con un conjunto multiplexado de cebadores directos y un (o unos pocos) cebadores inversos específicos etiquetados. 101 denota ADN de doble cadena con un locus polimórfico de interés en X. 102 denota el ADN de doble cadena con adaptadores de unión añadidos para amplificación universal. 103 denota el ADN de una sola cadena con cebador directo A hibridado. 104 denota el producto de la PCR de 103 que se amplificó utilizando el cebador directo A y el cebador inverso de etiqueta del adaptador de unión LT, y el cual es el producto de la PCR final. Este método puede detectar secuencias diana más cortas que la PCR estándar. Sin embargo, puede ser relativamente inespecífico, ya que sólo se utiliza un cebador específico diana. Este protocolo es efectivamente la mitad de la mini PCR unilateral anidada

Mini-PCR semianidada inversa: Es posible utilizar ADN diana que tenga un adaptador en los extremos del fragmento (véase la **Figura 8**). La STA se puede realizar con un conjunto multiplexado de cebadores directos y un (o unos pocos) cebadores inversos específicos etiquetados. 101 denota ADN de doble cadena con un locus polimórfico de interés en X. 102 denota el ADN de doble cadena con adaptadores de unión añadidos para amplificación universal. 103 denota el ADN de una sola cadena con cebador inverso B hibridado. 104 denota el producto de la PCR de 103 que se amplificó utilizando el cebador inverso B y el cebador directo de etiqueta del adaptador de unión LT. 105 denota el producto de la PCR de 104 con el cebador directo A hibridado y el cebador inverso interior 'b'. 106 denota el producto de la PCR que se ha amplificado de 105 utilizando el cebador directo A y el cebador inverso 'b', y el cual es el producto de la PCR final. Este método puede detectar secuencias diana más cortas que la PCR estándar.

Puede haber más variantes que son simplemente iteraciones o combinaciones de los métodos anteriores como PCR doblemente anidada, donde se utilizan tres conjuntos de cebadores. Otra variante es una mini-PCR unilateral anidada y media, donde la STA también se puede realizar con un conjunto multiplexado de cebadores directos anidados y un (o unos pocos) cebadores inversos específicos etiquetados.

Cabe señalar que en todas estas variantes, la identidad del cebador directo y el cebador inverso se puede intercambiar. Cabe señalar que en algunas realizaciones, la variante anidada se puede también ejecutar sin la preparación de la biblioteca inicial que comprende anexar las etiquetas del adaptador, y un paso de amplificación universal. Cabe señalar que en algunas realizaciones, se pueden incluir rondas adicionales de PCR, con pasos de amplificación y cebadores directos y/o inversos adicionales; estos pasos adicionales pueden ser especialmente útiles si es deseable aumentar todavía más el porcentaje de moléculas de ADN que corresponden a los loci focalizados.

Anidado de flujos de trabajo

Hay muchas maneras de realizar la amplificación, con diferentes grados de anidado, y con diferentes grados de multiplexado. En la Figura 9, se muestra un diagrama de flujo con algunos de los flujos de trabajo posibles. Cabe señalar que el uso de PCR de 10.000-plex sólo está destinado a ser un ejemplo; estos diagramas de flujo funcionarían igualmente bien para otros grados de multiplexado.

Adaptadores de unión de bucle

Al añadir adaptadores de etiquetado universal por ejemplo, con el fin de crear una biblioteca para la secuenciación, hay varias maneras de unir los adaptadores. Una manera es finalizar abruptamente el ADN de muestra, realizando un factor de cola A, y unir con adaptadores que tienen una proyección T. Hay otros métodos de unir adaptadores. Hay también una serie de adaptadores que se pueden unir. Por ejemplo, se puede utilizar un adaptador Y en el que el adaptador se compone de dos hebras de ADN donde una hebra tiene una región con una doble hebra, y en la que el adaptador se compone de una región de un cebador directo, y donde la otra hebra especificada por una región con una doble hebra que es complementaria de la región con una doble hebra en la primera hebra, y una región con un cebador inverso. La región de doble cadena, cuando se recuce, puede contener una proyección T con el propósito de unirse al ADN de doble cadena con una proyección A.

En una realización, el adaptador puede ser un bucle de ADN donde las regiones terminales son complementarias, y donde la región del bucle contiene una región etiquetada del cebador directo (LFT), una región etiquetada del cebador inverso (LRT), y un punto de escisión entre las dos (Véase la Figura 10). 101 se refiere al ADN diana de doble cadena, finalizado repentinamente. 102 se refiere al ADN diana con una cola A.

103 se refiere al adaptador de unión de bucle con proyección T 'T' y el punto de escisión 'Z'. 104 se refiere al ADN diana con adaptadores de unión de bucle anexados. 105 se refiere al ADN diana con los adaptadores de unión anexados escindidos en el punto de escisión. LFT se refiere a la etiqueta directa del adaptador de unión, y LRT se refiere a la etiqueta inversa del adaptador de unión. La región complementaria puede terminar en una proyección T, u otra característica que se pueda utilizar para la unión al ADN diana. El punto de escisión puede ser una serie de uracilos para escisión por UNG, o una secuencia que se puede reconocer y escindir mediante una enzima de restricción u otro método de escisión o simplemente una amplificación básica. Estos adaptadores se pueden utilizar para cualquier preparación de biblioteca, por ejemplo, para la secuenciación. Estos adaptadores se pueden utilizar en combinación con

cualquiera de los otros métodos descritos en el presente documento, por ejemplo, los métodos de amplificación por mini-PCR.

Cebadores etiquetados internamente

Cuando se utiliza la secuenciación para determinar el alelo presente en un locus polimórfico determinado, la lectura de secuencia de forma típica se inicia corriente arriba del punto de unión (a) del cebador y, a continuación, del punto polimórfico (X). Las etiquetas de forma típica se configuran como se muestra en la Figura 11, izquierda. 101 se refiere al ADN diana de una sola cadena con el locus polimórfico de interés 'X', y el cebador 'a' con la etiqueta 'b' anexada. Para evitar la hibridación no específica, el punto de unión del cebador (región de ADN diana complementaria a 'a') de forma típica tiene una longitud de 18 a 30 bp. La etiqueta de secuencia 'b' es, de forma típica, de unos 20 bp; en teoría estos pueden tener cualquier longitud superior a 15 bp, aunque mucha gente utiliza las secuencias del cebador que comercializa la compañía de la plataforma de secuenciación. La distancia 'd' entre 'a' y 'X' puede ser al menos 2 bp para evitar el sesgo alélico. Cuando se realiza la amplificación por PCR multiplexada utilizando los métodos que se divulgan en el presente documento u otros métodos, donde es necesario un cuidadoso diseño de los cebadores para evitar la interacción excesiva del cebador, la ventana de distancia permitida 'd' entre 'a' y 'X' puede variar un poco: de 2 bp a 10 bp, de 2 bp a 20 bp, de 2 bp a 30 bp, o incluso de 2 bp a más de 30 bp. Por tanto, cuando se utiliza la configuración de cebador que se muestra en la Figura 11, izquierda, las lecturas de secuencia deben tener un mínimo de 40 bp para obtener lecturas lo suficientemente largas para medir el locus polimórfico, y dependiendo de las longitudes de 'a' y 'd' las lecturas de secuencia pueden necesitar hasta 60 o 75 bp. Normalmente, cuanto más largas sean las lecturas de secuencia, mayor será el coste y el tiempo de secuenciar un número de lecturas determinado, por tanto, minimizar la longitud de lectura necesaria puede ahorrar tiempo y dinero. Además, puesto que, por término medio, las bases que se han leído al principio de la lectura se leen de un modo más preciso que las que se leen más tarde en la lectura, disminuir la longitud de lectura de secuencia necesaria puede también aumentar la precisión de las mediciones de la región polimórfica.

En una realización, los cebadores etiquetados denominados internamente, el punto de unión del cebador (a) se divide en una pluralidad de segmentos ((a', a'', a'''...), y la etiqueta de secuencia (b) está en un segmento de ADN que está en medio de dos de los puntos de unión del cebador, como se muestra en la Figura 11, 103. Esta configuración permite al secuenciador hacer lecturas de secuencia más cortas. En una realización, a' + a'' debería ser al menos de unos 18 bp, y puede ser tan largo como 30, 40, 50, 60, 80, 100 o más de 100 bp. En una realización, a'' debería ser al menos de unos 6 bp, y en una realización está entre unos 8 y 16 bp. Siendo todos los demás factores iguales, el uso de los cebadores etiquetados internamente puede cortar la longitud de las lecturas de secuencia necesarias al menos 6 bp, tanto como 8 bp, 10 bp, 12 bp, 15 bp,

e incluso hasta 20 o 30 bp. Esto puede dar como resultado una ventaja significativa en cuanto a dinero, tiempo y precisión. Un ejemplo de cebadores etiquetados internamente se muestra en la Figura 12.

Cebadores con región de unión al adaptador de unión

Un problema del ADN fragmentado es que, puesto que su longitud es corta, la posibilidad de que un polimorfismo esté cerca del extremo de una hebra de ADN es más alta que para una hebra larga (por ejemplo, 101, Figura 10). Puesto que la captura por PCR de un polimorfismo requiere un punto de unión del cebador con una longitud adecuada en ambos lados del polimorfismo, se perderán un número significativo de hebras de ADN con el polimorfismo focalizado debido a un solapamiento insuficiente entre el cebador y el punto de unión focalizado. En una realización, el ADN diana 101 puede tener adaptadores de unión anexados 102, y el cebador diana 103 puede tener una región (cr) que es complementaria de la etiqueta del adaptador de unión (lt) anexada corriente arriba de la región de unión designada (a) (véase la Figura 13); por tanto, en los casos donde la región de unión (región de 101 que es complementaria a a) es más corta que los 18 bp que de forma típica son necesarios para la hibridación, la región (cr) en el cebador que es complementaria a la etiqueta de biblioteca puede aumentar la energía de unión hasta un punto donde la PCR puede proseguir. Cabe señalar que cualquier especificidad que se pierda debido a una región de unión más corta se puede compensar con otros cebadores para PCR con regiones de unión diana convenientemente largas. Cabe señalar que esta realización se puede utilizar en combinación con la PCR directa, o cualquiera de los otros métodos descritos en el presente documento, como PCR anidada, semi PCR anidada, hemi PCR anidada, unilateral anidada o semi o hemi PCR anidada, u otros protocolos de PCR.

Cuando se utilizan los datos de secuenciación para determinar ploidía en combinación con un método analítico que implica la comparación de los datos de alelos observados con las distribuciones alélicas previstas para distintas hipótesis, cada lectura adicional de los alelos con una profundidad de lectura baja producirá más información que una lectura de un alelo con una profundidad de lectura alta. Por tanto, en el mejor de los casos, se desearía ver una profundidad de lectura (DOR) uniforme donde cada locus tenga un número similar de lecturas de secuencia representativas. Por tanto, es deseable minimizar la varianza de DOR. En una realización, es posible disminuir el coeficiente de varianza de DOR (este se puede definir como la desviación estándar de la DOR / la DOR media) aumentando los tiempos de reformación térmica. En algunas realizaciones las temperaturas de reformación térmica pueden durar más de 2 minutos, más de 4 minutos, más de diez minutos, más de 30 minutos, y más de una hora, o incluso más. Puesto que la reformación térmica es un proceso de equilibrio, no hay ningún límite para la mejora de varianza de DOR con el aumento de los tiempos de reformación térmica. En una realización, aumentar la concentración del cebador puede disminuir la varianza de DOR.

Ejemplos de métodos de amplificación de todo el genoma

En algunas realizaciones, un método de la presente divulgación puede implicar amplificar ADN, como el uso de la aplicación de todo el genoma para amplificar una muestra de ácido nucleico antes de amplificar sólo los loci diana. La amplificación del ADN, un proceso que transforma una pequeña cantidad de material genético en una gran cantidad de material genético que comprende un conjunto similar de datos genéticos, se puede realizar mediante una amplia variedad de métodos, incluyendo, aunque sin limitarse a ello, la reacción en cadena de polimerasa (PCR). Un método de amplificar ADN es la amplificación de todo el genoma (WGA). Hay una serie de métodos disponibles para WGA: PCR mediada por unión (LM-PCR), PCR de cebador oligonucleótido degenerado (DOP-PCR), y amplificación de desplazamiento múltiple (MDA). En LM-PCR, las secuencias de ADN cortas denominadas adaptadores están unidas a finales abruptos de ADN. Estos adaptadores contienen secuencias de amplificación universal, que se utilizan para amplificar el ADN por PCR. En DOP-PCR, los cebadores aleatorios que también contienen secuencias de amplificación universal se utilizan en una primera ronda de reformación térmica y PCR. A continuación, se utiliza una segunda ronda de PCR para amplificar las secuencias todavía más con las secuencias del cebador universal. MDA utiliza la ϕ -29 polimerasa, que es una enzima altamente procesiva y no específica que replica ADN y que se ha utilizado para el análisis de células individuales. Las principales limitaciones a la amplificación de material a partir de una sola célula son (1) la necesidad de utilizar concentraciones de ADN extremadamente diluidas o un volumen extremadamente pequeño de la mezcla de reacción, y (2) la dificultad de disociar de un modo fiable el ADN de proteínas en todo el genoma. No obstante, la amplificación de todo el genoma de una sola célula se ha utilizado satisfactoriamente para una variedad de aplicaciones durante varios años. Hay otros métodos de amplificar ADN de una muestra de ADN. La amplificación de ADN transforma la muestra de ADN inicial en una muestra de ADN que es similar en el conjunto de secuencias, pero con una cantidad mucho mayor. En algunos casos, la amplificación puede que no sea necesaria.

En algunas realizaciones, el ADN se puede amplificar utilizando una amplificación universal, como WGA o MDA. En algunas realizaciones, el ADN se puede amplificar mediante amplificación focalizada, por ejemplo, utilizando PCR focalizada, o sondas circularizantes. En algunas realizaciones, el ADN se puede enriquecer preferentemente utilizando un método de amplificación focalizada, o un método que dé como resultado la separación total o parcial de ADN deseado del no deseado, como planteamientos de captura por hibridación. En algunas realizaciones, el ADN se puede amplificar utilizando una combinación de un método de amplificación universal y un método de enriquecimiento preferencial. Una descripción más completa de algunos de estos métodos se puede encontrar en otra parte de este documento.

Ejemplos de métodos de enriquecimiento y secuenciación

En una realización, un método descrito en el presente documento utiliza técnicas de enriquecimiento selectivo que preservan las frecuencias alélicas relativas que están presentes en la muestra original de ADN en cada loci diana (por ejemplo, cada locus polimórfico) de un conjunto de loci diana (por ejemplo, loci polimórficos).

Aunque el enriquecimiento es especialmente beneficioso para los métodos para analizar loci polimórficos, estos métodos de enriquecimiento se pueden adaptar rápidamente para loci no polimórficos si se desea. En algunas realizaciones, la técnica de amplificación y/o enriquecimiento selectivo puede implicar una PCR, como una PCR mediada por unión, captura de un fragmento mediante hibridación, sondas de inversión molecular u otras sondas de circularización. En algunas realizaciones, los métodos para la amplificación o el enriquecimiento selectivo pueden implicar el uso de sondas, donde, tras la correcta hibridación con la secuencia diana, el extremo 3' o el extremo 5' de una sonda de nucleótidos está separado del sitio polimórfico del alelo por un pequeño número de nucleótidos. Esta separación reduce la amplificación preferente de un alelo, lo que se denomina sesgo alélico. Esto supone una mejora respecto de los métodos que implican el uso de sondas en las que el extremo 3' y el extremo 5' de una sonda correctamente hibridada se encuentran directamente adyacentes o muy cerca del sitio polimórfico de un alelo. En una realización, las sondas en las que la región de hibridación pueden contener o contienen un sitio polimórfico son excluidas. Los sitios polimórficos del sitio de hibridación pueden causar una hibridación desigual o inhibir directamente la hibridación en algunos alelos, lo que resulta en una amplificación preferente de determinados alelos. Estas realizaciones implican mejoras con respecto a otros métodos que conllevan una amplificación focalizada y/o un enriquecimiento selectivo, en el sentido de que preservan mejor las frecuencias alélicas originales de la muestra en cada locus polimórfico, cuando la muestra es una muestra genómica pura de un único individuo o una mezcla de individuos.

El uso de una técnica para enriquecer una muestra de ADN en un conjunto de loci diana seguida de una secuenciación como parte de un método para la determinación de alelos prenatal no invasiva o determinación de ploidía puede conferir un número de ventajas inesperadas. En algunas realizaciones de la presente divulgación, el método implica la medición de datos genéticos para su uso con un método basado en informática, como PARENTAL SUPPORT™ (PS). El resultado final de algunas de las realizaciones son los datos genéticos factibles de un embrión o un feto. Hay muchos métodos que se pueden utilizar para medir los datos genéticos del individuo y/o los individuos relacionados como parte de los métodos incorporados. En una realización, se describe en el presente documento un método para enriquecer la concentración de un conjunto de alelos focalizados, el método comprende uno o más de los pasos siguientes: amplificación focalizada de material genético, adición de sondas oligonucleótidas específicas de los loci, unión de hebras especificadas de ADN, aislamiento de conjuntos de ADN deseado, extracción de componentes no deseados de una reacción, detección de determinadas secuencias de ADN por hibridación, y detección de la secuencia de una o de una pluralidad de hebras de ADN mediante métodos de secuenciación de ADN. En algunos casos las hebras de ADN pueden referirse a material genético diana, en algunos casos pueden referirse a cebadores, en algunos casos pueden

referirse a secuencias sintetizadas, o una combinación de estas. Estos pasos se pueden llevar a cabo en una serie de órdenes diferentes.

Por ejemplo, un paso de amplificación universal del ADN antes de la amplificación focalizada puede conferir varias ventajas, como la eliminación del riesgo de cuello de botella y la reducción del sesgo alélico. El ADN se puede mezclar con una sonda oligonucleótida que se puede hibridar con dos regiones colindantes de la secuencia diana, una a cada lado. Después de la hibridación, los extremos de la sonda se pueden conectar añadiendo una polimerasa, un medio de unión, y cualquier reactivo necesario para permitir la circularización de la sonda. Después de la circularización, se puede añadir una exonucleasa para digerir el material genético no circularizado, seguido de la detección de la sonda circularizada. El ADN se puede mezclar con cebadores para PCR que se pueden hibridar con dos regiones colindantes de la secuencia diana, una a cada lado. Después de la hibridación, los extremos de la sonda se pueden conectar añadiendo una polimerasa, un medio de unión, y cualquier reactivo necesario para completar la amplificación por PCR. El ADN amplificado y no amplificado se puede focalizar mediante sondas de captura híbridas que focalizan un conjunto de loci; después de la hibridación, la sonda se puede localizar y separar de la mezcla para proporcionar una mezcla de ADN que está enriquecida en secuencias diana.

El uso de un método para focalizar determinados loci seguido de una secuenciación como parte de un método para la determinación de alelos o determinación de ploidía puede conferir un número de ventajas inesperadas. Algunos métodos mediante los cuales se puede focalizar el ADN, o enriquecerse preferentemente, incluyen el uso de sondas circularizantes, sondas invertidas vinculadas (LIPs, MIPS), la captura por métodos de hibridación como SURESELECT, y estrategias de amplificación por PCR mediada por unión o por PCR focalizada.

En algunas realizaciones, un método de la presente divulgación implica la medición de datos genéticos para su uso con un método basado en informática, como PARENTAL SUPPORT™ (PS), que se describe más adelante en el presente documento. PARENTAL SUPPORT™ es un planteamiento basado en informática para manipular datos genéticos, aspectos del cual se describen en el presente documento. El resultado final de algunas de las realizaciones son los datos genéticos factibles de un embrión o un feto seguido de una decisión clínica basada en los datos factibles. Los algoritmos tras el método PS toman los datos genéticos medidos del individuo diana, a menudo un embrión o feto, y los datos genéticos medidos de individuos relacionados, y son capaces de aumentar la precisión con la que el estado genético del individuo diana es conocido. En una realización, los datos genéticos medidos se utilizan en el contexto de realizar determinaciones de ploidía durante la diagnosis genética prenatal. En una realización, los datos genéticos medidos se utilizan en el contexto de realizar determinaciones de ploidía o determinaciones de alelos en embriones durante la fertilización in vitro. Hay muchos métodos que se pueden utilizar para medir los datos genéticos del individuo y/o los individuos relacionados en los contextos mencionados anteriormente. Los diferentes métodos comprenden una serie de pasos, estos pasos a menudo implican la amplificación de material genético, la adición de sondas oligonucleótidas, la unión de hebras especificadas de ADN, el aislamiento de conjuntos de ADN deseado, la extracción de componentes no deseados de una reacción, la detección de determinadas secuencias de ADN por hibridación, la detección de la secuencia de una o de una pluralidad de hebras de ADN mediante métodos de secuenciación de ADN. En algunos casos las hebras de ADN pueden referirse a material genético diana, en algunos casos pueden referirse a cebadores, en algunos casos pueden referirse a secuencias sintetizadas, o una combinación de los mismos. Estos pasos se pueden llevar a cabo en una serie de órdenes diferentes.

Cabe señalar que en teoría es posible focalizar cualquier número de loci en el genoma, desde un loci hasta más de un millón de loci. Si una muestra de ADN se somete a focalización y, a continuación, se secuenciar, el porcentaje de los alelos que son leídos por el secuenciador se enriquecerá con respecto a su abundancia natural en la muestra. El grado de enriquecimiento puede estar entre uno por ciento (o incluso menos) a multiplicarse por diez, multiplicarse por cien, multiplicarse por mil o incluso multiplicarse por muchos millones. En el genoma humano hay aproximadamente 3 mil millones de pares de bases, y nucleótidos, que comprenden aproximadamente 75 millones de loci polimórficos. Cuantos más loci se focalizan, menor es el grado de enriquecimiento posible. Cuanto menor es el número de loci que se focalizan, mayor es el grado de enriquecimiento posible, y mayor la profundidad de lectura que se puede conseguir en estos loci para un número determinado de lecturas de secuencia.

En una realización de la presente divulgación la focalización o preferencial puede centrarse completamente en SNP. En una realización, la focalización o preferencial puede centrarse en un punto polimórfico. Un número de productos de focalización comercial están disponibles para enriquecer exones. Sorprendentemente, focalizar exclusivamente SNPs, o exclusivamente loci polimórficos, es especialmente beneficioso cuando se utiliza un método para el NPD que se basa en distribuciones alélicas. También hay publicados métodos para el NPD utilizando la secuenciación, por ejemplo, la patente USA 7.888.017, implica un análisis de recuento de lecturas donde el recuento de lecturas se centra en contar el número de lecturas que corresponden a un cromosoma determinado, donde las lecturas de secuencia analizadas no se centran en regiones del genoma que son polimórficas. Estos tipos de metodología que no se centran en alelos polimórficos no se beneficiarían tanto de la focalización o enriquecimiento preferente de un conjunto de alelos.

En una realización de la presente divulgación, es posible utilizar un método de focalización que se centra en las SNP para enriquecer una muestra genética en regiones polimórficas del genoma. En una realización, es posible concentrarse en un pequeño número de SNP, por ejemplo, entre 1 y 100 SNP, o un número mayor, por ejemplo, entre 100 y 1.000, entre 1.000 y 10.000, entre 10.000 y 100.000 o más de 100.000 SNP. En una realización, es posible concentrarse en un cromosoma o en un pequeño número de los cromosomas que están correlacionados con nacimientos trisómicos vivos, por ejemplo, los cromosomas 13, 18, 21, X y Y, o alguna combinación de estos.

En una realización, es posible enriquecer los SNP focalizados por un factor pequeño, por ejemplo, entre multiplicarse por 1,01 y multiplicarse por 100, o por un factor mayor, por ejemplo, entre multiplicarse por 100 y multiplicarse por 1.000.000, o incluso multiplicarse por más de 1.000.000. En una realización de la presente divulgación, es posible utilizar un método de focalización para crear una muestra de ADN que está preferentemente enriquecida en regiones polimórficas del genoma. En una realización, es posible utilizar este método para crear una mezcla de ADN con cualquiera de estas características donde la mezcla de ADN contiene ADN materno y también ADN fetal flotante libre. En una realización, es posible utilizar este método para crear una mezcla de ADN que tiene cualquier combinación de estos factores. Por ejemplo, el método descrito en el presente documento se puede utilizar para producir una mezcla de ADN que comprende ADN materno y ADN fetal, y que está preferentemente enriquecido con ADN que corresponde a 200 SNP, todos los cuales están ubicados en el cromosoma 18 o 21, y los cuales están enriquecidos una media multiplicada por 1000. En otro ejemplo, es posible utilizar el método para crear una mezcla de ADN que está preferentemente enriquecida en 10.000 SNP que están todos o la mayoría ubicados en los cromosomas 13, 18, 21, X y Y, y el promedio de enriquecimiento por loci es mayor que multiplicado por 500. Cualquiera de los métodos de focalización descritos en el presente documento se puede utilizar para crear mezclas de ADN que están enriquecidas preferentemente en determinados loci.

En algunas realizaciones, un método de la presente divulgación incluye además la medición del ADN en la fracción mixta utilizando un secuenciador de ADN de alto rendimiento, donde el ADN en la fracción mixta contiene un número desproporcionado de secuencias de uno o más cromosomas, donde uno o más cromosomas se toman del grupo que comprende el cromosoma 13, el cromosoma 18, el cromosoma 21, el cromosoma X, el cromosoma Y y combinaciones de estos.

En este documento se describen tres métodos: PCR multiplexada, captura focalizada por hibridación, y sondas invertidas vinculadas (LIPs), que se pueden utilizar para obtener y analizar mediciones de un número suficiente de loci polimórficos a partir de una muestra de plasma materno para detectar aneuploidía fetal; esto no pretende excluir otros métodos de enriquecimiento selectivo de loci focalizados. Se pueden utilizar también otros métodos sin cambiar la esencia del método. En cada caso, el polimorfismo al que se ha realizado el ensayo puede incluir polimorfismos de un solo nucleótido (SNPs), pequeñas indels, o STRs. Un método preferido implica el uso de SNPs. Cada planteamiento produce datos de frecuencia alélica; los datos de frecuencia alélica para cada locus focalizado y/o las distribuciones de frecuencia alélica conjunta de estos loci se pueden analizar para determinar la ploidía del feto. Cada planteamiento tiene sus propias consideraciones debido al material de origen limitado y al hecho de que el plasma materno se compone de una mezcla de ADN materno y fetal. Este método se puede combinar con otros planteamientos para proporcionar una determinación más precisa. En una realización, este método se puede combinar con un enfoque de recuento de secuencia como, por ejemplo, el que se describe en la patente USA 7.888.017. Los planteamientos descritos podrían también ser utilizados para detectar la paternidad fetal de un modo no invasivo a partir de muestras de plasma materno. Además, cada planteamiento se puede aplicar a otras mezclas de ADN o muestras de ADN puro para detectar la presencia o ausencia de cromosomas aneuploides, genotipificar una gran cantidad de SNP de muestras de ADN degradadas, detectar variaciones de número de copias segmentarias (CNVs), detectar otros estados genotípicos de interés, o alguna combinación de estos.

Medición precisa de las distribuciones alélicas en una muestra

Se pueden utilizar planteamientos de secuenciación actuales para estimar la distribución de alelos en una muestra. Uno de estos métodos implica secuencias de muestreo aleatorias de un conjunto de ADN, denominado secuenciación por fuerza bruta o "shotgun". La proporción de un alelo determinado en los datos de secuenciación es de forma típica muy baja y se puede determinar por simple estadística. El genoma humano contiene aproximadamente 3 mil millones de pares de bases. De modo que, si el método de secuenciación utilizado hace 100 bp lecturas, un alelo determinado se medirá aproximadamente una vez cada 30 millones de lecturas de secuencia.

En una realización, un método de la presente divulgación se utiliza para determinar la presencia o ausencia de dos o más haplotipos diferentes que contienen el mismo conjunto de loci en una muestra de ADN de las distribuciones alélicas medidas de loci de ese cromosoma. Los diferentes haplotipos podrían representar dos cromosomas homólogos diferentes de un individuo, tres cromosomas homólogos diferentes de un individuo trisómico, tres haplotipos homólogos diferentes de una madre y un feto, donde uno de los haplotipos es compartido entre la madre y el feto, tres o cuatro haplotipos de una madre y un feto, donde uno o dos de los haplotipos son compartidos entre la madre y el feto, u otras combinaciones. Los alelos que son polimórficos entre los haplotipos tienden a ser más informativos; sin embargo, cualquier alelo en el que la madre y el padre no sean ambos homocigotos para el mismo alelo producirá información útil a través de las distribuciones de alelos medidas, además de la información disponible de un simple análisis del recuento de las lecturas.

Sin embargo, la secuenciación por fuerza bruta de esta muestra resulta extremadamente ineficiente, dado que resulta en múltiples secuencias para regiones que no son polimórficas entre los diferentes haplotipos de la muestra, o corresponden a cromosomas que no son de interés, y por tanto no revelan ninguna información acerca de la proporción de los haplotipos diana. En el presente documento se describen métodos que se dirigen de forma específica y/o enriquecen preferentemente segmentos de ADN de la muestra que es más probable que sean polimórficos en el genoma, para aumentar la producción de información alélica obtenida a través de la secuenciación. Cabe señalar que para que las distribuciones alélicas medidas en una muestra enriquecida sean realmente representativas de las cantidades reales presentes en el individuo diana, resulta fundamental que el enriquecimiento preferente de un alelo en

comparación con otro alelo de un loci determinado de los segmentos focalizados sea escaso o nulo. Los métodos actuales conocidos en la técnica para focalizar alelos polimórficos están diseñados para garantizar que se detecten al menos algunos de los alelos presentes. Sin embargo, estos métodos no han sido diseñados con el fin de medir distribuciones alélicas no sesgadas de los alelos polimórficos presentes en la mezcla original. No está claro que ningún método concreto de enriquecimiento de la diana sea capaz de producir una muestra enriquecida en la que las distribuciones alélicas medidas representen de forma precisa las distribuciones alélicas presentes en la muestra original no amplificada mejor que cualquier otro método. Aunque cabe esperar múltiples métodos de enriquecimiento, en teoría, para lograr este objetivo, un experto en la técnica sabrá que existe un importante sesgo estocástico y determinístico en la amplificación, la focalización y otros métodos de enriquecimiento preferente actuales. Una realización de un método descrito en el presente documento permite una pluralidad de alelos que se encuentran en una mezcla de ADN que corresponde a un locus determinado en el genoma para que se amplifique, o se enriquezca preferentemente de una manera que el grado de enriquecimiento de cada uno de los alelos es casi el mismo. Otra manera de decirlo es que el método permite aumentar la cantidad relativa de los alelos presente en la mezcla en conjunto, mientras que el ratio entre los alelos que corresponden a cada locus se mantiene básicamente igual que cuando estaban en la mezcla original de ADN. Para algunos métodos documentados, el enriquecimiento preferente de loci puede dar como resultado sesgos alélicos de más del 1%, más del 2%, más del 5% e incluso más del 10%. El enriquecimiento preferente se puede deber a un sesgo de captura cuando se utiliza un planteamiento de captura por hibridación, o a un sesgo de amplificación que puede ser pequeño para cada ciclo, pero puede ser grande cuando se acumula durante 20, 30 o 40 ciclos. A efectos de la presente divulgación, para que el ratio se mantenga básicamente sin cambios es necesario que el ratio de los alelos de la mezcla original dividido por el ratio de los alelos de la mezcla resultante se sitúe entre 0,95 y 1,05, entre 0,98 y 1,02, entre 0,99 y 1,01, entre 0,995 y 1,005, entre 0,998 y 1,002, entre 0,999 y 1,001, o entre 0,9999 y 1,0001. Cabe señalar que el cálculo de los ratios de alelos presentado en el presente documento no se puede utilizar en la determinación del estado de ploidía del individuo diana y que es solo una métrica utilizada para medir el sesgo alélico.

En una realización, una vez que una mezcla se ha enriquecido preferentemente en el conjunto de loci diana, se puede secuenciar utilizando cualquiera de los instrumentos de secuenciación anteriores, actuales o de nueva generación que secuencian una muestra clónica (una muestra generada a partir de una única molécula; algunos ejemplos incluyen ILLUMINA GAIIx, ILLUMINA HiSEQ, LIFE TECHNOLOGIES SOLiD, 5500XL). Los ratios se pueden evaluar mediante secuenciación a través de los alelos específicos de la región diana. Las lecturas de esta secuenciación se pueden analizar y recontar en función del tipo de alelo y de los ratios de los diferentes alelos determinados en consecuencia. Para las variaciones que tienen entre una y unas cuantas bases de longitud, la detección de los alelos se realizará mediante secuenciación y es fundamental que la lectura de la secuenciación abarque el alelo en cuestión para evaluar la composición alélica de esa molécula capturada. El número total de moléculas capturadas sometidas a ensayo para el genotipo se puede aumentar incrementando la longitud de la lectura de secuenciación. La secuenciación completa de todas las moléculas garantizaría la recopilación de la cantidad máxima de datos disponibles en el conjunto enriquecido. Sin embargo, actualmente la secuenciación resulta cara y un método que puede medir distribuciones alélicas utilizando un número menor de lecturas de secuencia tendrá un gran valor. Además, existen limitaciones técnicas para la longitud máxima posible de lectura, así como limitaciones de precisión a medida que aumentan las longitudes de la lectura. Los alelos de máxima tendrán entre una y unas pocas bases de longitud, aunque teóricamente se puede utilizar cualquier alelo más corto que la longitud de la lectura de secuenciación. A pesar de que las variaciones alélicas se presentan en todos los tipos, los ejemplos proporcionados en el presente documento se centran en los SNP o en las variantes que se encuentran a tan solo algunos pares de bases de distancia. Las variantes más largas como las variantes del número de copias segmentarias se pueden detectar mediante la adición de estas variaciones más pequeñas en muchos casos, dado que las recopilaciones completas del SNP interno del segmento están duplicadas. Las variantes más largas que unas cuantas bases, como las STR, requieren una consideración especial y algunos planteamientos de focalización funcionan y otros no.

Existen múltiples planteamientos de focalización que se pueden utilizar para aislar de forma específica y enriquecer uno o una pluralidad de posiciones variables en el genoma. Típicamente, estas se basan en aprovechar la secuencia invariable que flanquea a la secuencia variable. Hay informes de otros relacionados con la focalización en el contexto de la secuenciación donde el sustrato es plasma materno (véase, por ejemplo, Liao et al., Clin. Chem. 2011; 57(1): pp. 92-101). Sin embargo, estos planteamientos utilizan sondas de focalización dirigidas a exones y no se centran en focalizar regiones polimórficas del genoma. En una realización, un método de la presente divulgación implica el uso de sondas de focalización que se centran de forma exclusiva o de forma casi exclusiva en regiones polimórficas. En una realización, un método de la presente divulgación implica el uso de sondas de focalización que se centran de forma exclusiva o de forma casi exclusiva en los SNP. En algunas realizaciones de la presente divulgación, los puntos polimórficos focalizados se componen de al menos un 10% de SNP, al menos un 20% de SNP, al menos un 30% de SNP, al menos un 40% de SNP, al menos un 50% de SNP, al menos un 60% de SNP, al menos un 70% de SNP, al menos un 80% de SNP, al menos un 90% de SNP, al menos un 95% de SNP, al menos un 98% de SNP, al menos un 99% de SNP, al menos un 99,9% de SNP, o exclusivamente SNP.

En una realización, se puede utilizar un método de la presente divulgación para determinar genotipos (composición de bases del ADN en un loci específico) y las proporciones relativas de estos genotipos a partir de una mezcla de moléculas de ADN, donde dichas moléculas de ADN pueden haberse originado a partir de uno o de varios individuos genéticamente distintos. En una realización, se puede utilizar un método de la presente divulgación para determinar los genotipos de un conjunto de loci polimórficos y los ratios relativos de la cantidad de alelos diferentes presentes en esos loci. En una realización, los loci polimórficos pueden estar compuestos completamente por SNP. En una realización, los

loci polimórficos pueden comprender SNP, repeticiones en tándem simples y otros polimorfismos. En una realización, se puede utilizar un método de la presente divulgación para determinar las distribuciones relativas de alelos en un conjunto de loci polimórficos en una mezcla de ADN, donde la mezcla de ADN comprende ADN procedente de la madre y ADN procedente del feto. En una realización, las distribuciones alélicas conjuntas se pueden determinar en una muestra de ADN aislado de sangre de una mujer embarazada. En una realización, las distribuciones alélicas en un conjunto de loci se pueden utilizar para determinar el estado de ploidía de uno o más cromosomas de un feto en gestación.

En una realización, la mezcla de moléculas de ADN se podría obtener de ADN extraído de múltiples células de un individuo. En una realización, la recopilación original de células de las que se obtiene el ADN puede comprender una mezcla de células diploides o haploides del mismo o de diferentes genotipos, si ese individuo presenta mosaicismo (germinal o somático). En una realización, la mezcla de moléculas de ADN también se podría obtener de ADN extraído de células únicas. En una realización, la mezcla de moléculas de ADN también se podría obtener de ADN extraído de una mezcla de dos o más células del mismo individuo o de diferentes individuos. En una realización, la mezcla de moléculas de ADN se podría obtener de ADN aislado de material biológico que ya ha sido liberado por células como el plasma sanguíneo, que se sabe que contienen ADN libre de células. En una realización, este material biológico puede ser una mezcla de ADN de uno o más individuos, como sucede durante el embarazo, donde se ha demostrado que el ADN fetal está presente en la mezcla. En una realización, el material biológico podría proceder de una mezcla de células que se encuentran en la sangre materna, donde algunas de las células son de origen fetal. En una realización, el material biológico podrían ser células de la sangre de una embarazada que han sido enriquecidas en células fetales.

Sondas circularizantes

Algunas realizaciones de la presente divulgación implican el uso de "sondas invertidas enlazadas" (LIP), que han sido anteriormente descritas en la bibliografía, para amplificar los loci diana antes o después de la amplificación utilizando cebadores que no son LIP en los métodos de PCR multiplexada de la invención. LIP es un término genérico que pretende abarcar tecnologías que implican la creación de una molécula circular de ADN, donde las sondas están diseñadas para hibridarse con una región focalizada de ADN a uno de los lados de un alelo focalizado, de forma que la adición de las polimerasas y/o ligasas adecuadas, y las condiciones, tampones y otros reactivos adecuados, completará la región invertida complementaria del ADN del alelo focalizado para crear un bucle circular de ADN que captura la información que se encuentra en el alelo focalizado. Las LIP también se pueden denominar sondas precircularizadas, sondas de precircularización o sondas de circularización. Las LIP pueden ser una molécula de ADN lineal de entre 50 y 500 nucleótidos de longitud, y en una realización de entre 70 y 100 nucleótidos de longitud; en algunas realizaciones, puede ser más larga o más corta de lo que se describe en el presente documento. Otras realizaciones de la presente divulgación implican diferentes encarnaciones de la tecnología de las LIP, tales como sondas candado y sondas de inversión molecular (MIP).

Un método para focalizar ubicaciones específicas para la secuenciación consiste en sintetizar sondas en las que los extremos 3' y 5' de las sondas se reforman térmicamente con ADN diana en ubicaciones adyacentes a cualquiera de los lados de la región focalizada, de manera invertida, de forma que la adición de ADN polimerasa y ADN ligasa resulta en la extensión desde el extremo 3', añadiendo bases a la sonda de cadena única que son complementarias de la molécula diana (relleno de huecos), seguida de la unión del nuevo extremo 3' al extremo 5' de la sonda original, lo que resulta en una molécula de ADN circular que posteriormente se puede aislar del ADN de fondo. Los extremos de la sonda están diseñados para flanquear la región de interés focalizada. Un aspecto de este planteamiento se denomina habitualmente MIPS y se ha utilizado conjuntamente con tecnologías de array para determinar la naturaleza de la secuencia rellena. Una desventaja del uso de MIP en el contexto de la medición de los ratios de alelos es que los pasos de hibridación, circularización y amplificación no se producen a los mismos ratios para los diferentes alelos de los mismos loci. Esto da como resultado unos ratios de alelos medidos que no son representativos de los ratios de alelos reales presentes en la mezcla original.

En una realización, las sondas circularizantes se construyen de tal modo que la región de la sonda que está diseñada para que se hibride corriente arriba del locus polimórfico focalizado y la región de la sonda que está diseñada para que se hibride corriente abajo del locus polimórfico focalizado están conectadas covalentemente a través de la estructura central de ácido no nucleico. Esta estructura central puede ser cualquier molécula biocompatible o combinación de moléculas biocompatibles. Algunos ejemplos de posibles moléculas biocompatibles son poli (etilen glicol), policarbonatos, poliuretanos, polietilenos, polipropilenos, polímeros sulfonados, silicona, celulosa, fluoropolímeros, compuestos acrílicos, copolímeros de bloque estirénicos y otros copolímeros de bloque.

En una realización de la presente divulgación, este planteamiento ha sido modificado para que se pueda utilizar fácilmente para la secuenciación como medio de cuestionar la secuencia rellena. Con el fin de conservar las proporciones alélicas originales de la muestra original, se debe tener en cuenta al menos una consideración clave. Las posiciones variables entre diferentes alelos en la región de relleno de huecos no deben estar demasiado cerca de los puntos de unión de la sonda, dado que puede darse un sesgo de iniciación del ADN polimerasa que dé como resultado un diferencial de las variantes. Otra consideración es que puede haber otras variantes presentes en los puntos de unión de la sonda que están correlacionadas con las variantes de la región de relleno de huecos, lo que puede producir una amplificación desigual de los diferentes alelos. En una realización de la presente divulgación, los extremos 3' y los extremos 5' de la sonda precircularizada están diseñados para hibridarse a bases que se encuentran a una o unas cuantas posiciones de distancia de las posiciones variables (puntos polimórficos) del alelo focalizado. El número de bases entre el punto polimórfico (SNP u otro) y la base a cuyo extremo 3' o 5' se hibridará la sonda precircularizada

puede ser de una base, puede ser de dos bases, puede ser de tres bases, puede ser de cuatro bases, puede ser de cinco bases, puede ser de seis bases, puede ser de siete a diez bases, puede ser de once a quince bases, o puede ser de dieciséis a veinte bases, de veinte a treinta bases, o de treinta a sesenta bases. Los cebadores directos e inversos pueden estar diseñados para hibridarse a un número de bases de distancia del punto polimórfico. Las sondas circularizantes se pueden generar en grandes cantidades con la tecnología de síntesis del ADN actual lo que permite que se generen una gran cantidad de sondas y se agrupen potencialmente, lo que permite la interrogación de muchos loci simultáneamente. Se ha documentado que funciona con más de 300.000 sondas. Dos documentos que exponen un método que implica sondas de circularización que se pueden utilizar para medir los datos genómicos del individuo diana incluyen: Porreca et al., *Nature Methods*, 2007 4(11), pp. 931-936.; y también Turner et al., *Nature Methods*, 2009, 6(5), pp. 315-316. Los métodos descritos en estos documentos se pueden utilizar en combinación con otros métodos descritos en el presente documento. Determinados pasos del método de estos dos documentos se pueden utilizar en combinación con otros pasos de otros métodos que se describen en el presente documento. En algunas realizaciones de los métodos divulgados en el presente documento, el material genético del individuo diana es opcionalmente amplificado, seguido de una hibridación de las sondas precircularizadas, realizando un relleno de huecos para rellenar las bases entre los dos extremos de las sondas hibridadas, uniendo los dos extremos para formar una sonda circularizada, y amplificando la sonda circularizada, por ejemplo, mediante amplificación por círculo rodante. Una vez que se ha capturado la información genética de los alelos diana deseados mediante la circularización de las sondas de oligonucleótidos convenientemente diseñadas, como en el sistema LIP, la secuencia genética de las sondas circularizadas puede ser medida para obtener los datos de la secuencia deseados. En una realización, las sondas de oligonucleótidos convenientemente diseñadas pueden ser circularizadas directamente en el material genético no amplificado del individuo diana y amplificarse posteriormente. Cabe señalar que se pueden utilizar diversos procedimientos de amplificación para amplificar el material genético original, o las LIP circularizadas, incluyendo la amplificación por círculo rodante, la MDA u otros protocolos de amplificación. Se pueden utilizar diferentes métodos para medir la información genética del genoma diana, por ejemplo, utilizando una secuenciación de alto rendimiento, secuenciación de Sanger, otros métodos de secuenciación, captura por hibridación, captura por circularización, PCR multiplexada, otros métodos de hibridación y combinaciones de estos.

Una vez que el material genético del individuo se ha medido utilizando uno de los métodos anteriores o una combinación de estos, se puede utilizar un método informático, como el método PARENTAL SUPPORT™, junto con las mediciones genéticas correspondientes, para la determinación del estado de ploidía de uno o más cromosomas del individuo, y/o el estado genético de uno o un conjunto de alelos, concretamente de aquellos alelos que están correlacionados con una enfermedad o estado genético de interés. Cabe señalar que el uso de LIP ha sido documentado para la captura multiplexada de secuencias genéticas, seguida de la determinación del genotipo mediante secuenciación. Sin embargo, los datos de secuenciación resultantes de una estrategia basada en LIP para la amplificación del material genético que se encuentra en una única célula, una pequeña cantidad de células o ADN extracelular, no han sido utilizados al objeto de determinar el estado de ploidía de un individuo diana.

La aplicación de un método basado en la informática para determinar el estado de ploidía de un individuo a partir de los datos genéticos medidos mediante arrays de hibridación, como el array ILLUMINA INFINIUM o el chip genético AFFYMETRIX, ha sido descrita en documentos a los que se hace referencia en otros apartados del presente documento. Sin embargo, el método descrito en el presente documento muestra mejoras respecto de los métodos anteriormente descritos en la bibliografía. Por ejemplo, el planteamiento basado en LIP seguido de una secuenciación de alto rendimiento proporciona sorprendentemente unos mejores datos genotípicos debido a que el planteamiento tiene una mayor capacidad de multiplexado, una mejor especificidad de captura, una mejor uniformidad y un menor sesgo alélico. La mayor capacidad de multiplexado permite focalizar un mayor número de alelos, dando unos resultados más precisos.

La mejora de la uniformidad permite la medición de un mayor número de los alelos focalizados, dando unos resultados más precisos. Los bajos índices de sesgo alélico permiten unas tasas más reducidas de errores en la determinación, dando unos resultados más precisos. Unos resultados más precisos suponen una mejora de los resultados clínicos y una mejor atención médica.

Es importante señalar que las LIP se pueden utilizar como método para focalizar loci específicos en una muestra de ADN para la determinación del genotipo a través de métodos distintos de la secuenciación. Por ejemplo, se pueden utilizar LIP con el ADN diana para determinar el genotipo utilizando arrays de SNP u otros microarrays basados en ADN o ARN.

PCR mediada por unión

La PCR mediada por unión se puede utilizar para amplificar los loci diana antes o después de la amplificación por PCR utilizando cebadores que no están unidos. La PCR mediada por unión es un método de PCR utilizado para enriquecer preferentemente una muestra de ADN amplificando uno o una pluralidad de loci en una mezcla de ADN, donde el método consiste en lo siguiente: obtener un conjunto de pares de cebadores, donde cada cebador del par contiene una secuencia específica diana y una secuencia no diana, donde la secuencia específica diana está preferiblemente diseñada para hibridarse con una región diana, una corriente arriba y una corriente abajo del punto polimórfico, y que puede estar separada del punto polimórfico por 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11-20, 21-30, 31-40, 41-50, 51-100, o más de 100; la polimerización del ADN desde el extremo 3' del cebador corriente arriba para rellenar la región de cadena simple entre este y el extremo 5' de cebado del cebador corriente abajo con nucleótidos complementarios de la molécula

diana; la unión de la última base polimerizada del cebador corriente arriba a la base del extremo 5' de cebado adyacente del cebador corriente abajo; y amplificación solo de las moléculas polimerizadas y unidas utilizando las secuencias no diana contenidas en el extremo 5' de cebado del cebador corriente arriba y el extremo 3' de cebado del cebador corriente abajo. Los pares de cebadores para las distintas dianas pueden estar mezclados en la misma reacción. Las secuencias no diana sirven como secuencias universales de forma que todos los pares de cebadores que han sido polimerizados y unidos con éxito se pueden amplificar con un único par de cebadores de amplificación.

Captura mediante hibridación

En algunas realizaciones, un método de la presente divulgación puede implicar utilizar cualquiera de los siguientes métodos de captura mediante hibridación además de utilizar PCR multiplexada para amplificar los loci diana. El enriquecimiento preferente de un conjunto específico de secuencias en un genoma diana se puede realizar de múltiples maneras. En otro apartado de este documento se recoge una descripción de cómo se pueden utilizar las LIP para focalizar un conjunto específico de secuencias, pero en todas esas aplicaciones se pueden utilizar otros métodos de focalización y/o enriquecimiento preferente igualmente bien para los mismos fines. Un ejemplo de otro método de focalización es la captura mediante un planteamiento de hibridación. Algunos ejemplos de tecnologías de captura mediante hibridación disponibles en el mercado incluyen AGILENT's SURE SELECT Y TruSeq de ILLUMINA. En la captura mediante hibridación, se permite que se hibride un conjunto de oligonucleótidos que es complementario o en gran medida complementario de las secuencias focalizadas deseadas con una mezcla de ADN y, a continuación, se separa físicamente de la muestra. Una vez que las secuencias deseadas se han hibridado con los oligonucleótidos de focalización, el efecto de separar físicamente los oligonucleótidos de focalización incluye también la eliminación de las secuencias focalizadas. Una vez que se han retirado los oligonucleótidos hibridados, pueden ser calentados por encima de su temperatura de fusión y amplificados. Algunas maneras de eliminar físicamente los oligonucleótidos de focalización consisten en enlazar covalentemente los oligonucleótidos de focalización con un soporte sólido, por ejemplo, una perla magnética o un chip. Otra forma de eliminar físicamente los oligonucleótidos de focalización consiste en enlazarlos covalentemente con una fracción molecular con una fuerte afinidad por otra fracción molecular. Un ejemplo de este par de moléculas es biotina y estreptavidina, como el que se utiliza en SURE SELECT. De esta forma, las secuencias focalizadas se podrían unir covalentemente a una molécula de biotina y, después de la hibridación, utilizar un soporte sólido con estreptavidina para hacer bajar los oligonucleótidos biotinilados, a los que se hibridan las secuencias focalizadas.

La captura híbrida implica la hibridación de sondas que son complementarias de las dianas de interés de las moléculas diana. Las sondas de captura híbrida fueron originalmente desarrolladas para focalizar y enriquecer fracciones grandes del genoma con una uniformidad relativa entre dianas. En esa aplicación, era importante que todas las dianas se amplificasen con una uniformidad suficiente para que todas las regiones se pudiesen detectar mediante secuenciación, aunque no se prestó atención a conservar la proporción de alelos de la muestra original. Tras la captura, los alelos presentes en la muestra se pueden determinar mediante secuenciación directa de las moléculas capturadas. Estas lecturas de secuenciación se pueden analizar y recontar en función del tipo de alelos. Sin embargo, utilizando la tecnología actual, las distribuciones alélicas medidas de las secuencias capturadas típicamente no son representativas de las distribuciones alélicas originales.

En una realización, la detección de los alelos se realiza mediante secuenciación. A fin de capturar la identidad alélica del punto polimórfico, es esencial que la lectura de secuenciación abarque el alelo en cuestión para evaluar la composición alélica de esa molécula capturada. Dado que las moléculas capturadas suelen tener una longitud variable tras la secuenciación no se puede garantizar que solapen las posiciones variables a menos que se secuencie la molécula entera. Sin embargo, las consideraciones de costes y las limitaciones técnicas asociadas a la máxima longitud posible y a la precisión de las lecturas de secuenciación hacen que la secuenciación de la molécula entera resulte inviable. En una realización, la longitud de la lectura se puede aumentar desde unas 30 hasta unas 50 o unas 70 bases y puede aumentar en gran medida el número de lecturas que solapan las posiciones variables dentro de las secuencias focalizadas.

Otra forma de aumentar el número de lecturas que cuestionan la posición de interés consiste en disminuir la longitud de la sonda, de forma que no resulte en un sesgo en los alelos enriquecidos subyacentes. La longitud de la sonda sintetizada debería ser lo suficientemente larga para que dos sondas diseñadas para que se hibriden a dos alelos diferentes hallados en un locus se hibriden con casi igual afinidad a los diversos alelos en la muestra original. En la actualidad, los métodos conocidos en la técnica describen sondas que tienen típicamente más de 120 bases de largo. En una realización actual, si el alelo tiene entre una y unas pocas bases entonces las sondas de captura pueden tener menos de unas 110 bases, menos de unas 100 bases, menos de unas 90 bases, menos de unas 80 bases, menos de unas 70 bases, menos de unas 60 bases, menos de unas 50 bases, menos de unas 40 bases, menos de unas 30 bases y menos de unas 25 bases, y esto es suficiente para garantizar un enriquecimiento homogéneo de todos los alelos. Cuando la mezcla de ADN que se va a enriquecer utilizando la tecnología de captura híbrida es una mezcla que comprende ADN flotante libre aislado de una muestra de sangre, por ejemplo, de sangre materna, la longitud media del ADN es bastante corta, típicamente menos de 200 bases. El uso de sondas más cortas resulta en una mayor probabilidad de que las sondas de captura híbrida capturen los fragmentos de ADN deseados. Las variaciones más largas pueden requerir sondas más largas. En una realización, las variaciones de interés tienen entre una (un SNP) y unas pocas bases de longitud. En una realización, las regiones focalizadas del genoma se pueden enriquecer preferentemente utilizando sondas de captura híbrida donde las sondas de captura híbrida tienen una longitud inferior a 90 bases y pueden tener menos de 80 bases, menos de 70 bases, menos de 60 bases, menos de 50 bases, menos de

40 bases, menos de 30 bases o menos de 25 bases. En una realización, para aumentar la probabilidad de que el alelo deseado sea secuenciado, la longitud de la sonda diseñada para hibridarse con las regiones que flanquean la ubicación del alelo polimórfico se puede disminuir desde más de 90 bases hasta unas 80 bases, o hasta unas 70 bases, o hasta unas 60 bases, o hasta unas 50 bases, o hasta unas 40 bases, o hasta unas 30 bases o hasta unas 25 bases.

5 Existe un solapamiento mínimo entre la sonda sintetizada y la molécula diana para permitir la captura. Esta sonda sintetizada se puede hacer tan corta como sea posible, mientras siga siendo más larga que este solapamiento mínimo requerido. El efecto de utilizar una longitud de sonda más corta para focalizar una región polimórfica es que habrá más moléculas que solapan la región alélica diana. El estado de fragmentación de las moléculas de ADN originales también afecta al número de lecturas que solaparán los alelos focalizados. Algunas muestras de ADN, como muestras de
10 plasma, ya están fragmentadas debido a procesos biológicos que se producen in vivo. Sin embargo, las muestras con fragmentos más largos pueden beneficiarse de una fragmentación previa a la secuenciación de la preparación y el enriquecimiento de la biblioteca. Cuando tanto las sondas como los fragmentos son cortos (unas 60-80 pares de bases), se puede conseguir una especificidad máxima, dado que un número relativamente pequeño de lecturas de secuencia no solapan la región crítica de interés.

15 En una realización, las condiciones de hibridación se pueden ajustar para maximizar la uniformidad en la captura de diferentes alelos presentes en la muestra original. En una realización, las temperaturas de hibridación se reducen para minimizar las diferencias en el sesgo de hibridación entre alelos. Los métodos conocidos en la técnica evitan utilizar temperaturas inferiores para la hibridación, porque la bajada de la temperatura tiene el efecto de aumentar la hibridación de las sondas con dianas no deseadas. Sin embargo, cuando el objetivo consiste en preservar los ratios de alelos con la
20 máxima fidelidad, el planteamiento de utilizar bajas temperaturas de hibridación proporciona unos ratios de alelos óptimamente precisos, a pesar del hecho de que la técnica actual no recomienda este planteamiento. La temperatura de hibridación también se puede aumentar para exigir un mayor solapamiento entre la diana y la sonda sintetizada, de forma que solo se capturan las dianas con un solapamiento sustancial de la región focalizada. En algunas realizaciones de la presente divulgación, la temperatura de hibridación se reduce desde la temperatura de hibridación normal hasta
25 unos 40°C, hasta unos 45°C, hasta unos 50°C, hasta unos 55°C, hasta unas 60°C, hasta unos 65°C o hasta unos 70°C.

En una realización, las sondas de captura híbrida pueden estar diseñadas de forma que la región de la sonda de captura con el ADN que sea complementario al ADN que se encuentra en las regiones que flanquean el alelo polimórfico no se encuentra inmediatamente adyacente al punto polimórfico. En vez de esto, la sonda de captura puede estar diseñada de forma que la región de la sonda de captura que está diseñada para hibridarse con el ADN que
30 flanquea el punto polimórfico de la diana está separado de la parte de la sonda de captura que se pondrá en contacto conforme a las fuerzas de van der Waals con el punto polimórfico por una pequeña distancia que tiene una longitud equivalente a una o un pequeño número de bases. En una realización, la sonda de captura híbrida está diseñada para hibridarse con una región que flanquea el alelo polimórfico pero no lo cruza; esto se puede denominar sonda de captura de flanco. La longitud de la sonda de captura de flanco es posible que tenga menos de unas 120 bases, menos de
35 unas 110 bases, menos de unas 100 bases, menos de unas 90 bases, y puede tener menos de unas 80 bases, menos de unas 70 bases, menos de unas 60 bases, menos de unas 50 bases, menos de unas 40 bases, menos de unas 30 bases o menos de unas 25 bases. La región del genoma que la sonda de captura de flanco focaliza se puede separar mediante el locus polimórfico por 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11-20, o más de 20 pares de bases.

Descripción de una prueba de detección de una enfermedad basada en la captura focalizada utilizando una captura de
40 secuencia focalizada. Captura de secuencia focalizada personalizada, como las que ofrecen actualmente AGILENT (SURE SELECT), ROCHE-NIMBLEGEN, o ILLUMINA. Las sondas de captura podrían diseñarse de forma personalizada para garantizar la captura de diversos tipos de mutaciones. Para las mutaciones puntuales, una o más sondas que solapan la mutación puntual deberían ser suficientes para capturar y secuenciar la mutación.

Para las pequeñas inserciones o deleciones, una o más sondas que solapan la mutación pueden ser suficientes para
45 capturar y secuenciar los fragmentos que comprenden la mutación. La hibridación puede ser menos eficaz por lo que respecta a la eficacia de captura con limitación de sonda, típicamente diseñada para la secuencia del genoma de referencia. Para garantizar la captura de fragmentos que comprenden la mutación, se podrían diseñar dos sondas, una correspondientes al alelo normal y otra correspondiente al alelo mutante. Una sonda más larga puede mejorar la hibridación. Múltiples sondas con solapamiento pueden mejorar la captura. Por último, la colocación de una sonda
50 inmediatamente adyacente a la mutación, pero sin solapamiento, puede permitir una eficacia de captura relativamente similar de los alelos normales y mutantes.

En el caso de las repeticiones en tándem simples (STR), es poco probable que una sonda que solapa estos puntos altamente variables capture bien el fragmento. Para mejorar la captura se podría colocar una sonda adyacente al punto variable, aunque sin solapamiento. A continuación, se podría secuenciar el fragmento con normalidad para revelar la
55 longitud y la composición de la STR.

En el caso de grandes deleciones, puede funcionar el uso de una serie de sondas de solapamiento, un planteamiento habitual que se utiliza actualmente en los sistemas de captura de exones. Sin embargo, con este planteamiento puede ser difícil determinar si un individuo es heterocigoto o no. La focalización y evaluación de los SNP dentro de la región capturada podrían potencialmente revelar la pérdida de heterocigosidad en la región, indicando que el individuo es
60 portador. En una realización, es posible colocar sondas sin solapamiento o sondas únicas en la región potencialmente suprimida y utilizar el número de fragmentos capturados como medida de la heterocigosidad. En el caso en el que un individuo lleva una gran deleción, se espera que la mitad del número de fragmentos esté disponible para la captura en

relación a un locus de referencia (diploide) no suprimido. Por consiguiente, el número de lecturas obtenidas de las regiones suprimidas debería ser aproximadamente la mitad del obtenido de un locus diploide normal. La agrupación y la determinación de la media de la profundidad de las lecturas de secuenciación de múltiples sondas únicas en una región potencialmente suprimida pueden mejorar la señal y la certeza del diagnóstico. También se puede combinar los dos planteamientos, la focalización de SNP para identificar la pérdida de heterocigosidad y la utilización de múltiples sondas únicas para obtener una medida cuantitativa de la cantidad de fragmentos subyacentes de ese locus. Cualquiera de estas estrategias o ambas se pueden combinar con otras estrategias para obtener mejor el mismo fin.

Si durante las pruebas del ADN libre de células, la detección de un feto varón, indicado por la presencia de los fragmentos del cromosoma Y, capturados y secuenciados en la misma prueba, y una mutación dominante vinculada a X donde la madre y el padre no se ven afectados, o una mutación dominante donde la madre no se ve afectada, indicaría un mayor riesgo para el feto. La detección de dos alelos recesivos mutantes dentro del mismo gen de una madre no afectada implicaría que el feto ha heredado un alelo mutante del padre y potencialmente un segundo alelo mutante de la madre. En todos los casos, se puede indicar la realización de pruebas de seguimiento mediante amniocentesis o muestreo de las vellosidades coriónicas.

Una prueba de detección de una enfermedad basada en la captura focalizada se podría combinar con una prueba de diagnóstico prenatal no invasiva basada en la captura focalizada para determinar la aneuploidía.

Existen diversas formas de reducir la variabilidad de la profundidad de lectura (DOR): por ejemplo, se podrían aumentar las concentraciones del cebador, se podrían utilizar sondas de amplificación focalizadas más largas, o se podrían ejecutar más ciclos STA (como más de 25, más de 30, más de 35 o incluso más de 40).

Ejemplos de métodos de determinación del número de moléculas de ADN en una muestra

En el presente documento se describe un método para determinar el número de moléculas de ADN en una muestra generando una molécula exclusivamente identificada para cada molécula de ADN original de la muestra durante la primera ronda de amplificación del ADN. En el presente documento se describe un procedimiento para conseguir este fin seguido de un método de secuenciación de una única molécula o clónico.

El planteamiento conlleva la focalización de uno o más loci específicos y la generación de una copia etiquetada de las moléculas originales de tal manera que la mayoría o todas las moléculas etiquetadas de cada locus focalizado tendrán una etiqueta única y se pueden distinguir entre sí tras la secuenciación de este código de barras utilizando la secuenciación clónica o de una única molécula. Cada código de barras único secuenciado representa una única molécula en la muestra original. Simultáneamente, los datos de secuenciación se utilizan para determinar el locus del que se origina la molécula. Utilizando esta información se puede determinar el número de moléculas únicas en la muestra original para cada locus.

Este método se puede utilizar para cualquier aplicación en la que se requiere la evaluación cuantitativa del número de moléculas de una muestra original. Además, el número de moléculas únicas de una o más dianas se puede relacionar con el número de moléculas únicas de una o más dianas diferentes para determinar el número de copias, la distribución de los alelos o el ratio alélico relativos. Alternativamente, se puede establecer un modelo del número de copias detectadas de las diversas dianas mediante una distribución, con el fin de identificar el número de copias más probable de las dianas originales. Las aplicaciones incluyen, entre otras, la detección de inserciones y deleciones como las que se encuentran en los portadores de la distrofia muscular de Duchenne; la cuantificación de deleciones o duplicaciones de segmentos de cromosomas como las que se observan en variantes del número de copias; el número de copias del cromosoma de las muestras de individuos nacidos; el número de copias del cromosoma de las muestras de individuos no nacidos como embriones o fetos.

El método se puede combinar con la evaluación simultánea de variaciones contenidas en la secuencia focalizada. Esto se puede utilizar para determinar el número de moléculas que representan cada alelo en la muestra original. Este método del número de copias se puede combinar con la evaluación de los SNP u otras variaciones de secuencia para determinar el número de copias del cromosoma de los individuos nacidos y no nacidos; la discriminación y cuantificación de las copias de los loci que tienen variaciones de secuencias cortas, pero en las que la PCR puede amplificar a partir de múltiples regiones diana como en la detección de un portador de la atrofia muscular espinal; determinación del número de copias de diferentes fuentes de moléculas de muestras que se componen de mezclas de diferentes individuos como en la detección de la aneuploidía fetal a partir del ADN flotante libre obtenido del plasma materno.

En una realización, el método ya que pertenece a un locus diana individual puede comprender uno o más de los siguientes pasos: (1) El diseño de un par de oligómeros estándar para la amplificación por PCR de un locus específico. (2) La adición, durante la síntesis, de una secuencia de bases especificadas sin complementariedad o con una complementariedad mínima al locus diana o genoma al extremo 5' de uno de los oligómeros específicos diana. Esta secuencia, denominada cola, es una secuencia conocida, que se utilizará para la amplificación posterior, seguida por una secuencia de nucleótidos aleatorios. Estos nucleótidos aleatorios comprenden la región aleatoria. La región aleatoria comprende una secuencia generada aleatoriamente de ácidos nucleicos que probabilísticamente difieren entre cada molécula de la sonda. Por consiguiente, tras la síntesis, el grupo de oligómeros con cola se compondrá de una colección de oligómeros que comienzan con una secuencia conocida seguida de una secuencia desconocida que difiere entre moléculas, seguida de la secuencia específica diana. (3) Realización de una ronda de amplificación (desnaturalización, reformación térmica, ampliación) utilizando solo el oligómero con cola. (4) Adición de exonucleasa a la reacción, detención efectiva de la reacción por PCR e incubación de la reacción a la temperatura apropiada para

eliminar los oligómeros de cadena simple directos que no se han reformado térmicamente con respecto a la plantilla y ampliar para formar un producto de doble cadena. (5) Incubación de la reacción a una temperatura elevada para desnaturalizar la exonucleasa y eliminar su actividad. (6) Adición a la reacción de un nuevo oligonucleótido que es complementario a la cola del oligómero utilizado en la primera reacción junto con el otro oligómero específico diana para permitir la amplificación por PCR del producto generado en la primera ronda de la PCR. (7) Continuar la amplificación para generar suficiente producto para la secuenciación clónica corriente abajo. (8) Medición del producto de la PCR amplificado a través de múltiples métodos, por ejemplo, secuenciación clónica, a un número suficiente de bases para prolongar la secuencia.

En una realización, un método de la presente divulgación implica la focalización de múltiples loci en paralelo o de otro modo. Se pueden generar cebadores para diferentes loci diana independientemente y mezclados para crear grupos para la PCR multiplexada. En una realización, las muestras originales se pueden dividir en subconjuntos y se pueden focalizar diferentes loci en cada subconjunto antes de su recombinación y secuenciación. En una realización, el paso de etiquetado y un número de ciclos de amplificación se pueden realizar antes de subdividir el conjunto, con el fin de garantizar un etiquetado eficaz de todas las dianas antes de la división y mejorar la amplificación posterior continuando la amplificación utilizando conjuntos más pequeños de cebadores en grupos subdivididos.

Un ejemplo de una aplicación donde esta tecnología sería especialmente útil es en el diagnóstico de la aneuploidía prenatal no invasiva donde se puede utilizar el ratio de alelos en un locus determinado o una distribución de alelos en un número de loci para ayudar a determinar el número de copias de un cromosoma presente en un feto. En este contexto, resulta recomendable amplificar el ADN presente en la muestra inicial al tiempo que se mantienen las cantidades relativas de los diversos alelos. En algunas circunstancias, especialmente en los casos en los que hay una cantidad muy pequeña de ADN, por ejemplo, menos de 5000 copias del genoma, menos de 1000 copias del genoma, menos de 500 copias del genoma y menos de 100 copias del genoma, se puede producir un cuello de botella. Esto ocurre cuando hay un pequeño número de copias de cualquier alelo dado en la muestra inicial y los sesgos de amplificación pueden dar como resultado el conjunto amplificado de ADN que tiene unos ratios significativamente diferentes de los alelos que se encuentran en la mezcla inicial de ADN. Aplicando un conjunto de códigos de barras únicos o prácticamente únicos a cada cadena de ADN antes de la amplificación por PCR estándar, se pueden excluir $n-1$ copias de ADN de un conjunto de n moléculas idénticas del ADN secuenciado que procedía de la misma molécula original.

Por ejemplo, supongamos un SNP heterocigoto en el genoma de un individuo y una mezcla de ADN del individuo donde 10 moléculas de cada alelo se encuentran presentes en la muestra original de ADN. Después de la amplificación puede haber 100.000 moléculas de ADN que corresponden a ese locus. Debido a los procesos estocásticos, el ratio de ADN podría ser cualquiera entre 1:2 y 2:1; sin embargo, dado que cada una de las moléculas originales se ha etiquetado con una etiqueta única, se podría determinar que el ADN del grupo amplificado procedía exactamente de 10 moléculas de ADN de cada alelo. Por tanto, este método proporcionaría una medición más precisa de las cantidades relativas de cada alelo que un método que no utilice este planteamiento. Para los métodos en los que resulta deseable que la cantidad relativa de sesgo alélico sea minimizada, este método proporcionará datos más precisos.

La asociación del fragmento secuenciado al locus diana se puede lograr de varias maneras. En una realización, una secuencia de longitud suficiente se obtiene del fragmento focalizado para que abarque el código de barras de la molécula, así como un número suficiente de bases únicas que corresponden a la secuencia diana para permitir la identificación sin ambigüedad del locus diana. En otra realización, el cebador con el código de barras molecular que contiene el código de barras molecular generado aleatoriamente puede también contener un código de barras específico del locus (código de barras del locus) que identifica la diana a la que se va a asociar. Este código de barras del locus sería idéntico entre todos los cebadores con código de barras molecular para cada diana individual y, por tanto, todos los amplicones resultantes, pero diferente de todas las otras dianas. En una realización, el método de etiquetado que se describe en el presente documento se puede combinar con un protocolo de anidado unilateral.

En una realización, el diseño y la generación de cebadores del código de barras molecular se puede reducir en la práctica a lo siguiente: los cebadores de códigos de barras moleculares se pueden componer de una secuencia que no es complementaria de la secuencia diana, seguida de una región de código de barras molecular aleatoria, seguida de una secuencia específica diana. La secuencia 5' del código de barras molecular se puede utilizar para la amplificación por PCR de la subsecuencia y se puede componer de secuencias útiles para la conversión del amplicón en una biblioteca para la secuenciación. La secuencia del código de barras molecular aleatoria se podría generar de múltiples maneras. El método preferible sintetiza el cebador de etiquetado de la molécula de tal forma que incluye las cuatro bases para la reacción durante la síntesis de la región del código de barras. Todas o varias combinaciones de bases se pueden especificar utilizando los códigos de ambigüedad de ADN IUPAC. De esta manera la colección sintetizada de moléculas contendrá una mezcla aleatoria de secuencias en la región del código de barras molecular. La longitud de la región del código de barras determinará cuántos cebadores contendrán códigos de barras únicos. El número de secuencias únicas está relacionado con la longitud de la región del código de barras como N^L donde N es el número de bases, típicamente 4, y L es la longitud del código de barras. Un código de barras de cinco bases puede producir hasta 1024 secuencias únicas; un código de barras de ocho bases puede producir 65536 secuencias únicas. En una realización, el ADN se puede medir mediante un método de secuenciación, donde los datos de secuencia representan la secuencia de una única molécula. Esto puede incluir métodos en los que las moléculas únicas son secuenciadas directamente o métodos en los que las moléculas únicas son amplificadas para formar clones detectables mediante el instrumento de la secuencia, pero que aun así representan moléculas únicas denominadas en el presente documento secuenciación clónica.

Ejemplos de métodos y reactivos para la cuantificación de los productos de la amplificación

La cuantificación de secuencias específicas de ácido nucleico de interés se realiza de forma típica mediante técnicas de PCR en tiempo real cuantitativas como TAQMAN (LIFE TECHNOLOGIES), sondas INVADER (THIRD WAVE TECHNOLOGIES), y similares. Dichas técnicas sufren de numerosas deficiencias como una capacidad limitada de conseguir el análisis simultáneo de múltiples secuencias en paralelo (multiplexación) y la capacidad de proporcionar datos cuantitativos precisos para sólo un estrecho rango de ciclos de amplificación posibles (por ejemplo, cuando el logaritmo de la cantidad de producción por la amplificación por PCR frente al número de ciclos está en un rango lineal). Las técnicas de secuenciación del ADN, especialmente las técnicas de secuenciación de nueva generación de alto rendimiento (a menudo denominadas técnicas de secuenciación masivamente paralelas) como las empleadas en MYSEQ (ILLUMINA), HISEQ (ILLUMINA), ION TORRENT (LIFE TECHNOLOGIES), GENOME ANALYZER ILX (ILLUMINA), GS FLEX+ (ROCHE 454) etc., se pueden utilizar para mediciones cuantitativas del número de copias de secuencia de interés presente en la muestra, proporcionando así información cuantitativa sobre los materiales de inicio, por ejemplo, número de copia o niveles de transcripción. Los secuenciadores genéticos de alto rendimiento son susceptibles de utilizar la codificación por medio de barras (es decir, etiquetado de la muestra con secuencias de ácido nucleico distintivas) para identificar muestras específicas de individuos permitiendo así el análisis simultáneo de múltiples

muestras en una sola ejecución del secuenciador de ADN. El número de veces que una región determinada del genoma en una preparación de la biblioteca (u otra preparación nucleica de interés) se secuencia (número de lecturas) será proporcional al número de copias de esa secuencia en el genoma de interés (o nivel de expresión en el caso de cDNA que contiene preparaciones). Sin embargo, la preparación y secuenciación de bibliotecas genéticas (y preparaciones derivadas de genoma similares) puede presentar numerosos sesgos que interfieren con la obtención de una lectura cuantitativa precisa para la secuencia de ácido nucleico de interés. Por ejemplo, secuencias de ácido nucleico diferentes se pueden amplificar con eficacias diferentes durante los pasos de amplificación nucleica que se producen durante la preparación de la biblioteca genética o la preparación de la muestra.

El problema con la eficacia de amplificación diferenciales se puede mitigar utilizando determinadas realizaciones de la invención sujeto. La invención objeto incluye distintas métodos y composiciones que están relacionados con el uso de estándares para su inclusión en procesos de amplificación que se pueden utilizar para mejorar la precisión de cuantificación. La invención se utiliza, entre otras áreas, en la detección de aneuploidía en un feto mediante el análisis del ADN fetal flotante libre en la sangre materna, como se describe en el presente documento y como se describe, entre otros lugares, en la Patente USA nº. 8.008.018; la Patente USA nº. 7.332.277; Solicitud publicada PCT WO 2012/078792A2; y la Solicitud publicada PCT WO 2011/146632 A1. Las realizaciones de la invención también se utilizan en la detección de aneuploidía en embriones generados in vitro. Entre las aneuploidías comercialmente significativas que se pueden detectar se incluye la aneuploidía de los cromosomas humanos 13, 18, 21, 20 X e Y. Las realizaciones de la divulgación se pueden utilizar con ácidos nucleicos humanos o no humanos, y se puede aplicar a ácidos nucleicos derivados de animales y plantas. Las realizaciones de la invención también se pueden utilizar para detectar y/o cuantificar alelos para otros trastornos genéticos caracterizados por deleciones o inserciones. Los alelos que contienen deleciones se pueden detectar en presuntos portadores del alelo de interés.

Una realización de la invención sujeto incluye estándares que están presentes en una cantidad conocida (relativa o absoluta). Por ejemplo, considere una biblioteca genética realizada a partir de un origen genético que es diploide para el cromosoma 8 (que contiene el locus A) y triploide para el cromosoma 21 (que contiene el locus B). Se puede producir una biblioteca genética de esta muestra que contendrá secuencias en cantidades que son una función del número de los cromosomas presentes en la muestra, por ejemplo, 200 copias del locus A y 300 copias del locus B. Sin embargo, si el locus A amplifica de un modo mucho más eficaz que el locus B, después de la PCR puede haber 60.000 copias del amplicón A y 30.000 copias del amplicón B, oscureciendo así el verdadero número de copia cromosómica de la muestra genómica inicial cuando se realiza el análisis mediante secuenciación del ADN de alto rendimiento (u otras técnicas de detección de ácido nucleico cuantitativas). Para mitigar este problema se emplea una secuencia estándar para el locus A, en la que la secuencia estándar se amplifica con básicamente la misma eficacia que el locus A. De forma similar, se crea una secuencia estándar para el locus B, donde la secuencia estándar se amplifica con básicamente la misma eficacia que el locus B. Una secuencia estándar de locus A y una secuencia estándar para el locus B se añaden a la mezcla antes de la PCR (u otras técnicas de amplificación). Estas secuencias estándar están presentes en cantidades conocidas, o bien cantidades relativas o cantidades absolutas. De este modo, si se añadiera una mezcla 1:1 de secuencia estándar A y secuencia estándar B (antes de la amplificación) a la mezcla del ejemplo anterior, se producirían 3000 copias del amplicón A estándar y se producirían 1000 copias del amplicón B estándar, que mostrarían que el locus A se ha amplificado 3 veces más eficientemente que el locus B, bajo la misma serie de condiciones.

En varias realizaciones una o más regiones seleccionadas de un genoma que contiene un SNP (u otro polimorfismo) de interés se puede amplificar específicamente y secuenciar posteriormente. Esta amplificación específica diana puede producirse durante la formación de una biblioteca genética para la secuenciación. La biblioteca puede contener numerosas regiones focalizadas para la amplificación. En algunas realizaciones al menos 10; 100; 500; 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 regiones de interés. Ejemplos de dichas bibliotecas se describen en el presente documento y se pueden encontrar en la patente USA con el número 13/300.235 (publicada como US2012/0270212), presentada el 18 de noviembre de 2011.

Muchas técnicas de secuenciación del ADN de alto rendimiento requieren la modificación del material de inicio genético, por ejemplo, la unión de puntos de cebado universal y/o códigos de barras, para formar bibliotecas con el fin de facilitar la amplificación clónica de fragmentos pequeños de ácido nucleico antes de realizar las reacciones de secuenciación posteriores. En algunas realizaciones, se añaden una o más secuencias estándar durante la formación de la biblioteca genética o se añaden a un componente precursor de una biblioteca genética antes de la amplificación de la biblioteca. Las secuencias estándar se pueden seleccionar para imitar (pero aun así pudiendo distinguirse en base a una secuencia base nucleótida) fragmentos genómicos diana para prepararse para la secuenciación mediante una técnica de secuenciación genética de alto rendimiento. En una realización, la secuencia estándar puede ser idéntica al fragmento genómico diana exceptuando uno, dos, tres, de cuatro a diez, o de once a veinte nucleótidos. En algunas realizaciones, cuando la secuencia genética diana contiene un SNP, la secuencia estándar puede ser idéntica al SNP exceptuando el nucleótido en la base polimórfica, que se puede elegir que sea uno de los cuatro nucleótidos que no se observa en esa ubicación de forma natural. Las secuencias estándar se pueden utilizar en un análisis altamente multiplexado de múltiples loci diana (como loci polimórficos). Se pueden añadir secuencias estándar durante el proceso de formación de la biblioteca (antes de la amplificación) en cantidades conocidas (relativas o absolutas) con el fin de proporcionar una métrica estándar para obtener una mayor precisión en la determinación de la cantidad de secuencia diana de interés en la muestra de análisis. La combinación del conocimiento de las cantidades conocidas de las secuencias estándar utilizadas conjuntamente con el conocimiento de la formación del nivel de ploidía de la biblioteca para la secuenciación formado a partir de un genoma del nivel de ploidía caracterizado anteriormente, por ejemplo, conocido por ser diploide para todos los cromosomas autosómicos, se puede utilizar para calibrar las propiedades de amplificación de cada secuencia estándar con respecto a su secuencia diana correspondiente y explicar las variaciones entre lotes de mezclas que comprenden múltiples secuencias estándar. Dado que a menudo es necesario analizar simultáneamente un gran número de loci, es útil producir una mezcla que comprenda un amplio conjunto de secuencias estándar. Las realizaciones de la divulgación incluyen mezclas que comprenden múltiples secuencias estándar. Idealmente, la cantidad de cada secuencia estándar en la mezcla se conoce con una alta precisión. Sin embargo, es extremadamente difícil conseguir este ideal ya que en la práctica hay una cantidad significativa de variación en la cantidad de cada secuencia estándar en la mezcla, especialmente para mezclas que comprenden un gran número de oligonucleótidos sintéticos diferentes. Esta variación tiene numerosas fuentes, por ejemplo, las variaciones en las eficacias de la reacción de síntesis oligonucleotídica in vitro entre lote, inexactitudes en la medición del volumen, variaciones en el pipeteo. Además, esta variación puede producirse entre lotes diferentes de aquellos que teóricamente contienen el mismo conjunto exacto de secuencias estándar en las mismas cantidades exactas. Por tanto, es interesante calibrar cada lote de secuencias estándar independientemente. Los lotes de secuencias estándar se pueden calibrar frente a los genomas de referencia de la composición cromosómica conocida. La agrupación de lotes de secuencias estándar se puede calibrar mediante la secuenciación del lote de secuencias estándar con pasos de amplificación mínimos o sin pasos de amplificación en el protocolo de secuenciación. Las realizaciones de la divulgación incluyen mezclas calibradas de secuencias estándar diferentes. Otras realizaciones de la divulgación incluyen métodos de calibración de mezclas de secuencias estándar diferentes y mezclas calibradas de secuencias estándar diferentes realizadas por los métodos sujeto.

Distintas realizaciones de las mezclas sujeto de secuencias estándar y de los métodos para utilizarlas pueden comprender al menos 10; 100; 500; 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 o más secuencias estándar, así como varias cantidades intermedias. El número de secuencias estándar puede ser el mismo que el número de secuencias diana seleccionado para el análisis durante la generación de una biblioteca focalizada para la secuenciación del ADN. Sin embargo, en algunas realizaciones, puede ser beneficioso utilizar un número inferior de secuencias estándar que el número de regiones focalizadas en la biblioteca que se está construyendo. Puede ser beneficioso utilizar el número más bajo para evitar hacer frente a los límites de la capacidad de secuenciación del secuenciador de ADN de alto rendimiento que se está utilizando. El número de secuencias estándar puede ser 50% o inferior al número de regiones focalizadas, 40% o inferior al número de regiones focalizadas, puede ser un 30% o inferior al número de regiones focalizadas, 20% o inferior al número de regiones focalizadas, puede ser un 10% o inferior al número de regiones focalizadas, 5% o inferior al número de regiones focalizadas, 1% o inferior al número de regiones focalizadas, así como distintos valores intermedios. Por ejemplo, si se crea una biblioteca genética utilizando 15.000 pares de cebadores focalizados a SNP específicos que contienen loci, una mezcla adecuada que contiene 1500 secuencias estándar correspondientes a 1500 de los 15.000 loci focalizados se pueden añadir antes del paso de amplificación de las construcciones de la biblioteca.

La cantidad de secuencias estándar que se añaden durante la construcción de la biblioteca puede variar considerablemente entre diferentes realizaciones. En algunas realizaciones, la cantidad de cada secuencia estándar puede ser aproximadamente la misma que la cantidad prevista de la secuencia diana presente en la muestra de material genómico que se utiliza para la preparación de la biblioteca. En otras realizaciones, la cantidad de cada secuencia estándar puede ser superior o inferior que la cantidad prevista de la secuencia diana presente en la muestra de material genómico que se utiliza para la preparación de la biblioteca. Mientras que las cantidades relativas iniciales de la secuencia diana y la secuencia estándar no son críticas para la función del método, es preferible que la cantidad esté dentro del rango 100 veces superior que 100 veces inferior a la cantidad de la secuencia diana presente en la muestra de material genómico que se utiliza para la preparación de la biblioteca. Cantidades excesivas de estándar pueden utilizar demasiada capacidad de secuenciación del secuenciador de ADN en una ejecución determinada del instrumento. Si se utiliza una cantidad demasiado baja de secuencias estándar se producirán datos insuficientes para ayudar en el análisis de la variación en la eficiencia de la amplificación.

Las secuencias estándar se pueden seleccionar para que sean muy similares en la secuencia base nucleótida a las regiones amplificadas de interés; preferiblemente la secuencia estándar tiene los mismos puntos exactos de unión del cebador que la región genómica analizada, es decir, la "secuencia diana." La secuencia estándar se debe poder distinguir de la secuencia diana correspondiente en un locus determinado. Por comodidad, esta región que se puede distinguir de la secuencia estándar se la denominará "secuencia de marcador." En algunas realizaciones, la región de la secuencia de marcador de las secuencias diana contiene la región polimórfica, por ejemplo, un SNP, y se puede flanquear a ambos lados por regiones de unión del cebador. La secuencia estándar se puede seleccionar para que se corresponda estrechamente con el contenido de GC de la secuencia diana correspondiente. En algunas realizaciones, las regiones de unión del cebador de la secuencia estándar están flanqueadas por puntos de cebado universal. Estos puntos de cebado universal se seleccionan para que coincidan con los puntos de cebado universal que se utilizan en una biblioteca genómica para análisis. En otras realizaciones, las secuencias estándar no tienen puntos de cebado universal y los puntos de cebado universal se añaden durante la creación de una biblioteca. Las secuencias estándar de forma típica se proporcionan en forma de una sola cadena. Una secuencia estándar se define con respecto a una secuencia diana correspondiente y los reactivos específicos de la secuencia que se utilizan para amplificar la secuencia diana. En algunas realizaciones, la secuencia diana contiene el polimorfismo de interés, por ejemplo, un SNP, una delección o inserción, presente en la muestra de ácido nucleico para análisis. La secuencia estándar es un polinucleótido sintético que es similar en la secuencia base nucleótida a la secuencia diana, pero que, sin embargo, se puede distinguir de la secuencia diana en virtud de al menos una diferencia base nucleótida, proporcionando así un mecanismo para distinguir las secuencias de amplicones derivadas de la secuencia estándar de las secuencias de amplicones derivadas de la secuencia diana. Las secuencias estándar se seleccionan de modo que tengan básicamente las mismas propiedades de amplificación que la secuencia diana correspondiente cuando se amplifican con el mismo conjunto de reactivos de amplificación, por ejemplo, cebadores para PCR. En algunas realizaciones, las secuencias estándar pueden tener los mismos puntos de unión de secuencia del cebador que las secuencias diana correspondientes. En otras realizaciones, las secuencias estándar pueden tener puntos de unión de secuencia del cebador diferentes que las secuencias diana correspondientes. En algunas realizaciones, las secuencias estándar se pueden seleccionar para producir amplicones que tienen la misma longitud que la longitud de los amplicones producidos a partir de las secuencias diana correspondientes. En otras realizaciones, las secuencias estándar se pueden seleccionar para producir amplicones que tienen longitudes ligeramente diferentes que la longitud de los amplicones producidos a partir de las secuencias diana correspondientes.

Después de que se han completado las reacciones de amplificación, la biblioteca se secuenciará en un secuenciador de ADN de alto rendimiento donde la molécula individual se amplifica clónicamente y se secuenciará. Se cuenta el número de lecturas de secuencia para cada alelo de la secuencia diana, también se cuenta el número de lecturas de secuencia para la secuencia estándar que corresponde a la secuencia diana. El proceso también se lleva a cabo para al menos otro par de secuencias diana y las secuencias estándar correspondientes. Considere, por ejemplo, que se producen lecturas del locus A, X_{A1} para el alelo 1 del locus A; se producen lecturas X_{A2} para el alelo 2 del locus A, y se producen lecturas X_{AC} para la secuencia estándar A. El ratio de (X_{A1} más X_{A2}) a X_{AC} se determina para cada locus de interés. Como se ha descrito anteriormente, el proceso se puede realizar en un genoma de referencia, por ejemplo, un genoma que se conoce que es diploide para todos los cromosomas. El proceso se puede repetir muchas veces para proporcionar un gran número de valores de lectura con el fin de determinar un número promedio de lecturas y la desviación estándar en el número de lecturas. El proceso se realiza con una mezcla que comprende un gran número de secuencias estándar diferentes que corresponden a loci diferentes. Al asumir que (1) X_{A1} más X_{A2} corresponde al número conocido de cromosoma, por ejemplo, 2 para el genoma femenino humano normal y (2) las secuencias estándar tienen propiedades de amplificación (y detección) similares a las de sus loci naturales correspondientes, se pueden determinar las cantidades relativas de las secuencias estándar diferentes en la mezcla estándar multiplexada. La mezcla de la secuencia estándar multiplexada calibrada puede utilizarse, a continuación, para ajustarse para la variabilidad en la eficacia de la amplificación entre los loci diferentes en una reacción de amplificación multiplexada.

Otras realizaciones de la divulgación incluyen métodos y composiciones para medir el número de copia de genes específicos de interés, incluyendo duplicaciones y genes mutantes caracterizados por grandes delecciones que interferirían con la cuantificación por secuenciación. La secuenciación tendría problemas para detectar alelos con estas delecciones. Las secuencias estándar incluidas en el proceso de amplificación se pueden utilizar para reducir este problema.

En una realización de la divulgación la secuencia diana para análisis es un gen que tiene un tipo salvaje (es decir, funcional) y una forma mutante caracterizada por una delección. Un ejemplo de dichos genes es SMN1, un alelo que tiene una delección responsable de la enfermedad genética SMA (atrofia muscular espinal). Es interesante detectar un individuo portador de la forma mutante del gen mediante técnicas de secuenciación genéticas de alto rendimiento. La aplicación de dichas técnicas a la detección de mutaciones de delección puede ser problemática, entre otras razones, por la ausencia de secuencias observadas en la secuenciación (en oposición a detectar una simple mutación puntual o SNP). Dichas realizaciones emplean (1) un par de cebadores de amplificación específicos para el gen de interés, donde en los cebadores de amplificación amplificarán el gen de interés (o una porción de este) y no amplificarán significativamente el alelo mutante, (2) una secuencia estándar correspondiente al alelo de tipo salvaje del gen de interés (es decir, una secuencia diana), pero que se diferencia en al menos una base nucleótida detectable, (3) un par de cebadores de amplificación específicos para una segunda secuencia diana que sirva como secuencia de referencia, y (4) una secuencia estándar que corresponda a la secuencia de referencia.

En una realización de la divulgación se proporciona un método para medir el número de copias del gen de interés, en donde el gen de interés tiene un alelo promedio que comprende una delección. El método puede emplear un reactivo de amplificación específico para el gen de interés, por ejemplo, cebadores para PCR, que son específicos para el gen de interés al amplificar al menos una porción del gen de interés, o todo el gen de interés, o una región adyacente al gen de interés, aunque no amplificando la delección que comprende el alelo del gen de interés. De forma adicional, el método sujeto emplea una secuencia estándar que corresponde al gen de interés, donde la secuencia estándar difiere en al menos una base nucleótida del gen de interés (de modo que la secuencia de la secuencia estándar se puede distinguir rápidamente del gen de interés que se produce naturalmente). De forma típica, la secuencia estándar contendrá los mismos puntos de unión del cebador que el gen de interés con el fin de minimizar cualquier discriminación de la amplificación entre el gen de interés y la secuencia estándar que corresponde al gen de interés. La reacción también comprenderá reactivos de amplificación específicos para una secuencia de referencia. La secuencia de referencia es una secuencia de un número de copia conocido (o al menos que se presupone que es conocido) en el genoma que se va a analizar. La reacción además comprende una secuencia estándar que corresponde a la secuencia de referencia. De forma típica, la secuencia estándar que corresponde a la secuencia de referencia contendrá los mismos puntos de unión del cebador que la secuencia de referencia con el fin de minimizar cualquier discriminación de la amplificación entre la secuencia de referencia y la secuencia estándar que corresponde a la secuencia de referencia.

Ejemplos de muestras de ácido nucleico

En algunas realizaciones, la muestra genética puede estar preparada y/o purificada. Hay una serie de procedimientos estándar conocidos en la técnica para llegar a tal fin. En algunas realizaciones, la muestra puede ser centrifugada para separar diversas capas. En algunas realizaciones, el ADN puede ser aislado utilizando filtración. En algunas realizaciones, la preparación de ADN puede implicar amplificación, separación, purificación mediante cromatografía, separación líquido-líquido, aislamiento, enriquecimiento preferente, amplificación preferente, amplificación focalizada o cualquiera de una serie de otras técnicas conocidas en la técnica o descritas en el presente documento.

En algunas realizaciones, se podría utilizar un método divulgado en el presente documento en situaciones en las que hay una cantidad muy reducida de ADN presente, como en la fertilización in vitro o en situaciones forenses, cuando se dispone de una o muy pocas células (normalmente menos de 10 células, menos de 20 células o menos de 40 células). En estas realizaciones, un método divulgado en el presente documento sirve para realizar determinaciones del estado de ploidía a partir de una pequeña cantidad de ADN que no está contaminado con otro ADN, pero en las que la determinación del estado de ploidía resulta muy difícil dada la reducida cantidad de ADN. En algunas realizaciones, un método divulgado en el presente documento se podría utilizar en situaciones en las que el ADN diana está contaminado con ADN de otro individuo, por ejemplo, en la sangre materna en el contexto del diagnóstico prenatal, las pruebas de paternidad o productos de pruebas de concepción. Algunas otras situaciones en las que estos métodos resultarían particularmente ventajosos serían en las pruebas del cáncer cuando solamente hay una o un pequeño número de células presente entre una mayor cantidad de células normales. Las mediciones genéticas utilizadas como parte de estos métodos se podrían realizar en cualquier muestra que comprenda ADN o ARN, por ejemplo, a título meramente enunciativo: sangre, plasma, fluidos corporales, orina, cabello, lágrimas, saliva, tejido, piel, uñas, blastómeros, embriones, líquido amniótico, muestras de vello coriónico, heces, bilis, linfa, moco cervical, semen u otras células o materiales que comprenden ácidos nucleicos. En una realización, un método divulgado en el presente documento se podría realizar con métodos de detección de ácido nucleico, como la secuenciación, microarrays, qPCR, PCR digital u otros métodos utilizados para medir ácidos nucleicos. Si por alguna razón se considerase deseable, los ratios de las probabilidades del recuento de alelos en un locus se podrían calcular y los ratios de alelos se podrían utilizar para determinar el estado de ploidía en combinación con algunos de los métodos descritos en el presente documento, siempre que los métodos sean compatibles. En algunas realizaciones, un método divulgado en el presente documento implica el cálculo, en un ordenador, de ratios de alelos en la pluralidad de loci polimórficos de las mediciones de ADN realizadas con las muestras procesadas. En algunas realizaciones, un método divulgado en el presente documento implica el cálculo, en un ordenador, de ratios de alelos en la pluralidad de loci polimórficos de las mediciones de ADN realizadas con las muestras procesadas junto con cualquier combinación de otras mejoras descritas en esta divulgación.

En algunas realizaciones, este método se puede utilizar para determinar el genotipo de una sola célula, un pequeño número de células, de dos a cinco células, de seis a diez células, de diez a veinte células, de veinte a cincuenta células, de cincuenta a cien células, de cien a mil células, o una pequeña cantidad de ADN extracelular, por ejemplo, de uno a diez picogramos, de diez a cien pictogramas, de cien pictogramas a un nanogramo, de uno a diez nanogramos, de diez a cien nanogramos o de cien nanogramos a un microgramo.

Ejemplos de estudios de expresión de ARN

Los métodos de PCR multiplexada de la invención se pueden utilizar para aumentar el número de loci diana que se pueden evaluar durante los experimentos de perfilado de expresión del gen. Por ejemplo, los niveles de expresión de miles de genes se pueden supervisar simultáneamente para determinar si una persona tiene una secuencia (como un polimorfismo u otra mutación) asociada a una enfermedad (como cáncer) o un mayor riesgo de una enfermedad. Estos métodos se pueden utilizar para identificar secuencias (como polimorfismos u otras mutaciones) asociados con un mayor o menor riesgo para una enfermedad como cáncer comparando la expresión del gen (como la expresión de alelos mRNA determinados) en muestras de pacientes con y sin la enfermedad. Además, se puede determinar el efecto de tratamientos, enfermedades, o etapas de desarrollo específicos en la expresión del gen. De forma similar, estos métodos se pueden utilizar para identificar genes cuya expresión cambia en respuesta a patógenos u otros organismos.

mediante la comparación de la expresión del gen en tejidos o células infectados y no infectados. En estos métodos el número de lecturas de secuenciación se pueden ajustar en base a la frecuencia de los polimorfismos que se están analizando de modo que se realizan suficientes lecturas para que se detecten los polimorfismos si están presentes.

En algunas realizaciones, una muestra que contiene ARN (como mRNA) se amplifica utilizando una transcriptasa inversa (RT) y el ADN resultante (como, por ejemplo, cDNA) se amplifica, a continuación, utilizando un ADN polimerasa (PCR). Las etapas de RT y PCR se pueden llevar a cabo de forma secuencial en el mismo volumen de reacción o por separado. Cualquiera de las bibliotecas de cebadores de la invención se puede utilizar en este método de reacción en cadena de polimerasa de transcripción inversa (RT-PCR). En varias realizaciones, se realiza la transcripción inversa utilizando oligo-dT, cebadores aleatorios, una mezcla de oligo-dT y cebadores aleatorios, o cebadores específicos a los loci diana. Para evitar la amplificación de ADN genómico contaminante, los cebadores para RT-PCR pueden estar diseñados de modo que parte de un cebador se hibride con el extremo 3' de un exón y la otra parte del cebador se hibride con el extremo 5' del exón adyacente. Dichos cebadores se reforman térmicamente con el cDNA sintetizado a partir de mRNAs ensamblados, pero no a ADN genómico. Para detectar la amplificación del ADN contaminante, se pueden diseñar pares de cebador RT-PCR para flanquear una región que contiene al menos un intrón. Los productos amplificados de cDNA (que no son intrones) son más pequeños que los que están amplificados a partir de ADN genómico (que contiene intrones). La diferencia de tamaño en los productos se utiliza para detectar la presencia de ADN contaminante. En algunas realizaciones cuando sólo se conoce la secuencia mRNA, se eligen puntos de reformación térmica del cebador que están separados por al menos 300-400 pares de bases ya que es probable que los fragmentos de este tamaño de ADN eucariótico contengan zonas de unión exón-intrón. De forma alternativa, la muestra se puede tratar con DNase para degradar el ADN contaminante.

Ejemplos de métodos para pruebas de paternidad

Los métodos de PCR multiplexada de la invención se pueden utilizar para mejorar la precisión de las pruebas de paternidad ya que muchos loci diana se pueden analizar inmediatamente (véase, por ejemplo, la publicación USA con el número 2012/0122701, presentada el 22 de diciembre de 2011. Por ejemplo, el método de PCR multiplexada puede permitir analizar miles de loci polimórficos (como SNPs) para ser utilizados en el algoritmo PARENTAL SUPPORT que se describe en el presente documento para determinar si un presunto padre es el padre biológico de un feto. En algunas realizaciones, el método consiste en (i) amplificar simultáneamente una pluralidad de loci polimórficos que incluye al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes en material genético del presunto padre para producir un primer conjunto de productos amplificados; (ii) amplificar simultáneamente la pluralidad de loci polimórficos correspondiente en una muestra mezclada de ADN que se origina de una muestra de sangre de una madre embarazada para producir un segundo conjunto de productos amplificados; en donde la muestra mezclada de ADN comprende ADN fetal y ADN materno; (iii) determinar en un ordenador la probabilidad de que el presunto padre sea el padre biológico del feto utilizando mediciones genotípicas en base al primer y segundo conjuntos de productos amplificados; y (iv) establecer si el presunto padre es el padre biológico del feto utilizando la probabilidad determinada de que el presunto padre es el padre biológico del feto. En varias realizaciones, el método incluye además amplificar simultáneamente la pluralidad de loci polimórficos correspondiente en material genético de la madre para producir un tercer conjunto de productos amplificados; en donde la probabilidad de que el presunto padre sea el padre biológico del feto se determina utilizando mediciones genotípicas en base al primer, segundo y tercer conjuntos de productos amplificados.

Ejemplos de métodos para la selección y caracterización del embrión

Los métodos de PCR multiplexada de la invención se pueden utilizar para mejorar la selección de los embriones para la fertilización in vitro permitiendo analizar miles de loci diana inmediatamente (véase, por ejemplo, la pub. USA número 2011/0092763, presentada el 27 de mayo de 2008, presentada el 22 de diciembre de 2011). Por ejemplo, el método de PCR multiplexada puede permitir analizar miles de loci polimórficos (como SNPs) para ser utilizados en el algoritmo PARENTAL SUPPORT que se describe en el presente documento para seleccionar un embrión de un conjunto de embriones para la fertilización in vitro. En algunas realizaciones, la divulgación proporciona métodos de estimar las probabilidades relativas de que cada embrión de un conjunto de embriones se desarrollará como se desea. En algunas realizaciones, el método consiste en poner en contacto una muestra de cada embrión con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes para producir una mezcla de reacción para cada embrión, en donde cada una de las muestras se deriva de una o más células de un embrión. En algunas realizaciones, cada mezcla de reacción se somete a condiciones de reacción de extensión del cebador para producir productos amplificados. En algunas realizaciones, el método incluye determinar en un ordenador una o más características de al menos una célula de cada embrión en base a los productos amplificados; y estimar en un ordenador las probabilidades relativas de que cada embrión se desarrollará según se desee, en base a una o más características de al menos una célula para cada embrión. En algunas realizaciones, el método incluye el uso de un método basado en informática para determinar al menos una característica, como el algoritmo PARENTAL SUPPORT que se describe en el presente documento. En algunas realizaciones, la característica incluye un estado de ploidía.

En algunas realizaciones, la característica se selecciona del grupo que se compone de aneuploide, euploide, mosaicismo, nulisomía, monosomía, disomía uniparental, trisomía, tetrasomía, un tipo de aneuploidía, trisomía por error en la copia no emparejada, trisomía por error en la copia emparejada, origen materno de la aneuploidía, origen paterno de la aneuploidía, una presencia o ausencia de un gen vinculado a una enfermedad, una identidad cromosómica de

5 cualquier cromosoma aneuploide, una condición genética anómala, una delección o duplicación, una probabilidad de una característica, y combinaciones de estas. La característica puede estar asociada con un cromosoma tomado del grupo compuesto por el cromosoma uno, cromosoma dos, cromosoma tres, cromosoma cuatro, cromosoma cinco, cromosoma seis, cromosoma siete, cromosoma ocho, cromosoma nueve, cromosoma diez, cromosoma once, cromosoma doce, cromosoma trece, cromosoma catorce, cromosoma quince, cromosoma dieciséis, cromosoma diecisiete, cromosoma dieciocho, cromosoma diecinueve, cromosoma veinte, cromosoma veintiuno, cromosoma veintidós, cromosoma X o cromosoma Y, y combinaciones de estos.

Ejemplos de métodos de diagnóstico prenatal

10 Los métodos de PCR multiplexada de la presente invención se pueden utilizar para mejorar los métodos de diagnóstico prenatal, como la determinación del estado de ploidía de los cromosomas fetales. Dado el gran número de loci diana que se pueden amplificar simultáneamente, se pueden realizar determinaciones más precisas.

15 En una realización, la presente divulgación proporciona métodos ex vivo para determinar el estado de ploidía en un cromosoma de un feto en gestación a partir de los datos genotípicos medidos de una muestra mezclada de ADN (es decir, ADN de la madre del feto y ADN del feto) y opcionalmente de los datos genotípicos medidos de una muestra de material genético de la madre y posiblemente también del padre, donde la determinación se realiza utilizando un modelo de distribución conjunto para crear un conjunto de distribuciones de alelos previstas para diferentes estados de ploidía fetal posibles dados los datos genotípicos parentales, y comparando las distribuciones alélicas previstas con las distribuciones alélicas reales medidas en la muestra mezclada, y seleccionando el estado de ploidía cuyo patrón de distribución alélica prevista más se aproxime al patrón de distribución alélica observada. En una realización, la muestra

20 mezclada se obtiene de sangre materna, plasma o suero materno. En una realización, la muestra mezclada de ADN puede ser enriquecida preferentemente en un loci diana (por ejemplo, una pluralidad de loci polimórficos). En una realización, el enriquecimiento preferente se realiza de forma que se minimiza el sesgo alélico. En una realización, la presente divulgación se refiere a una composición de ADN que ha sido preferentemente enriquecido en una pluralidad de loci de forma que el sesgo alélico es reducido. En una realización, la distribución o distribuciones alélicas se miden secuenciando el ADN de la muestra mezclada. En una realización, el modelo de distribución conjunto asume que los alelos se distribuirán de forma binomial. En una realización, la serie de distribuciones alélicas conjuntas previstas se crean para loci genéticamente vinculados, al tiempo que se tienen en cuenta las frecuencias de recombinación existentes de diversas fuentes, por ejemplo, utilizando datos del International HapMap Consortium.

30 En una realización, la presente divulgación proporciona métodos para el diagnóstico prenatal no invasivo (NPD), en particular determinando el estado de aneuploidía de un feto mediante la observación de las mediciones alélicas en una pluralidad de loci polimórficos de los datos genotípicos medidos en las mezclas de ADN, donde determinadas mediciones alélicas son indicativas de un feto aneuploide, mientras que otras mediciones alélicas son indicativas de un feto euploide. En una realización, los datos genotípicos se miden secuenciando mezclas de ADN obtenidas de plasma materno. En una realización, la muestra de ADN puede ser preferiblemente enriquecida con moléculas de ADN que corresponden a la pluralidad de loci cuyas distribuciones alélicas se están calculando. En una realización, una muestra de ADN que comprende solo o prácticamente solo material genético de la madre y posiblemente también una muestra de ADN que comprende solo o prácticamente solo material genético del padre se someten a medición. En una realización, las mediciones genéticas de uno o los dos progenitores junto con la fracción fetal estimada se utilizan para crear una pluralidad de distribuciones alélicas previstas correspondientes a diferentes estados genéticos subyacentes posibles del feto; las distribuciones alélicas previstas se pueden denominar hipótesis. En una realización, los datos genéticos maternos no se determinan midiendo el material genético que pertenece exclusiva o casi exclusivamente a la madre, sino que se estiman a partir de las mediciones genéticas realizadas con el plasma materno que comprende una mezcla de ADN materno y fetal. En algunas realizaciones, las hipótesis pueden comprender la ploidía del feto en uno o más cromosomas, donde los segmentos de dichos cromosomas del feto fueron heredados de los progenitores, y combinaciones de estos. En algunas realizaciones, el estado de ploidía del feto se determina comparando las mediciones alélicas observadas con las diferentes hipótesis, donde al menos algunas de las hipótesis corresponden a diferentes estados de ploidía, y seleccionando el estado de ploidía que corresponde a la hipótesis que es más probable que sea cierta dadas las mediciones alélicas observadas. En una realización, este método implica el uso de datos de mediciones alélicas de algunos o todos los SNP medidos, con independencia de que los loci sean homocigóticos o heterocigóticos, y por tanto no implica el uso de alelos en loci que son únicamente heterocigóticos. Este método puede no resultar apropiado para situaciones en las que los datos genéticos pertenecen solo a un locus polimórfico. Este método resulta particularmente ventajoso cuando los datos genéticos comprenden datos para más de 10 loci polimórficos de un cromosoma diana o más de 20 loci polimórficos. Este método resulta especialmente ventajoso cuando los datos genéticos comprenden datos para más de 50 loci polimórficos para un cromosoma diana, más de 100 loci polimórficos o más de 200 loci polimórficos para un cromosoma diana. En algunas realizaciones, los datos genéticos pueden comprender datos para más de 500 loci polimórficos de un cromosoma diana, más de 1000 loci polimórficos, más de 2000 loci polimórficos o más de 5000 loci polimórficos para un cromosoma diana.

60 En una realización, un método divulgado en el presente documento produce una medición cuantitativa del número de observaciones independientes de cada alelo en un locus polimórfico. Esto contrasta con la mayoría de los métodos, como la PCR cualitativa o de microarrays, que proporcionan información sobre el ratio de dos alelos, pero no cuantifican el número de observaciones independientes de ninguno de los alelos. Con los métodos que proporcionan información cuantitativa sobre el número de observaciones independientes, en los cálculos de la ploidía solo se utiliza el ratio, mientras que la información cuantitativa en sí no resulta útil. Para ilustrar la importancia de retener información sobre el

número de observaciones independientes, tendremos en cuenta el locus de la muestra con dos alelos, A y B. En un primer experimento se observan 20 alelos A y 20 alelos B, y en un segundo experimento se observan 200 alelos A y 200 alelos B. En ambos experimentos el ratio $(A/(A+B))$ es igual a 0,5; sin embargo, el segundo experimento proporciona más información que el primero acerca de la certidumbre de la frecuencia del alelo A o B. Algunos métodos de otros implican determinar la media o sumar los ratios de alelos (ratios del canal) (es decir, x_i/y_i) del alelo individual y analizar este ratio, o bien comparándolo con un cromosoma de referencia o utilizando una regla que pertenece a cómo se espera que este ratio se comporte en situaciones concretas. Estos métodos no implican ninguna ponderación de los alelos y asumen que se puede garantizar aproximadamente la misma cantidad de producto de la PCR para cada alelo y que todos los alelos se deberían comportar del mismo modo. Este método presenta una serie de desventajas y, lo que es más importante, impide el uso de una serie de mejoras que se describen en la presente divulgación.

En una realización, un método divulgado en el presente documento establece un modelo explícitamente de las distribuciones de la frecuencia alélica prevista en la disomía, así como una pluralidad de distribuciones de la frecuencia alélica que se pueden esperar en caso de trisomía resultante de la ausencia de disyunción durante la meiosis I, la ausencia de disyunción durante la meiosis II, y/o la ausencia de disyunción durante la mitosis temprana en el desarrollo fetal. Para ilustrar por qué esto es importante, podemos poner como ejemplo un caso en el que no se produjeron cruces: la ausencia de disyunción durante la meiosis I resultaría en una trisomía en la que dos homólogos diferentes se heredaron de un progenitor; por el contrario, la ausencia de disyunción durante la meiosis II o durante la mitosis temprana en el desarrollo fetal resultaría en dos copias del mismo homólogo de un progenitor. Cada escenario daría como resultado unas frecuencias alélicas previstas diferentes en cada locus polimórfico y también en todos los loci considerados conjuntamente, debido a la unión genética. Los cruces, que provocan el intercambio de material genético entre homólogos, hacen que el patrón de herencia sea más complejo; en una realización, el método instantáneo tiene esto en cuenta utilizando información sobre la tasa de recombinación además de la distancia física entre loci. En una realización, para permitir una distinción mejorada entre la ausencia de disyunción de la meiosis I y la ausencia de disyunción de la meiosis II o la ausencia de disyunción de la mitosis, el método instantáneo incorpora en el modelo una probabilidad creciente de cruce dado que la distancia desde el centrómero aumenta. La ausencia de disyunción de la meiosis II y la mitosis se puede distinguir por el hecho de que la ausencia de disyunción mitótica típicamente resulta en copias idénticas o prácticamente idénticas de un homólogo, mientras que los dos homólogos presentes tras un evento de ausencia de disyunción de la meiosis II a menudo difieren debido a uno o más cruces durante la gametogénesis.

En algunas realizaciones, un método divulgado en el presente documento implica la comparación de las mediciones alélicas observadas con las hipótesis teóricas correspondientes a una posible aneuploidía genética fetal, y no implica un paso de cuantificación del ratio de alelos en un locus heterocigótico. Cuando el número de loci es inferior a unos 20, la determinación de la ploidía realizada utilizando un método que comprende la cuantificación de un ratio de alelos en un locus heterocigoto y una determinación de la ploidía realizada utilizando un método consistente en comparar las mediciones alélicas observadas con las hipótesis de distribución alélica teóricas correspondientes a posibles estados genéticos fetales pueden dar un resultado similar. Sin embargo, cuando el número de loci es superior a 50, es probable que estos dos métodos proporcionen resultados muy diferentes; cuando el número de loci es superior a 400, superior a 1000, o superior a 2000, es muy probable que estos dos métodos den resultados con unas diferencias cada vez más significativas. Estas diferencias se deben al hecho de que un método que comprende la cuantificación de un ratio de alelos en un locus heterocigoto sin medir la magnitud de cada alelo independientemente y agregar o determinar la media de los ratios impide el uso de técnicas como el uso de un modelo de distribución conjunto, realizar un análisis de enlaces, utilizar un modelo de distribución binomial, y/u otras técnicas estadísticas avanzadas, mientras que el uso de un método que comprende la comparación de las mediciones alélicas observadas con hipótesis de distribución alélica teóricas correspondientes a posibles estados genéticos fetales permite utilizar estas técnicas que pueden aumentar de forma significativa la precisión de la determinación.

En una realización, un método divulgado en el presente documento implica la determinación de si la distribución de mediciones alélicas observadas es indicativa de un feto euploide o aneuploide utilizando un modelo de distribución conjunto. El uso de un modelo de distribución conjunto es diferente y representa una mejora significativa respecto de los métodos que determinan las tasas de heterocigosidad tratando los loci polimórficos de forma independiente, y las determinaciones resultantes son significativamente más precisas. Sin ánimo de vincularse a ninguna teoría concreta, se cree que una razón por la que tienen una mayor precisión es que el modelo de distribución conjunto tiene en cuenta el enlace entre SNP, y la probabilidad de que se hayan producido cruces durante la meiosis que dio lugar a los gametos que formaron el embrión que creció hasta convertirse en feto. El propósito de utilizar el concepto de enlace a la hora de crear la distribución prevista de las mediciones alélicas para una o más hipótesis es que permite la creación de distribuciones de las mediciones alélicas previstas que se corresponden con la realidad considerablemente mejor que cuando no se utiliza el enlace. Por ejemplo, imaginemos que hay dos SNP, 1 y 2, cercanos entre sí, y la madre es A en el SNP 1 y A en el SNP 2 de un homólogo, y B en SNP1 y B en SNP 2 del homólogo dos. Si el padre es A en los dos SNP de los dos homólogos, y se mide una B para el SNP1 del feto, esto indica que el homólogo dos ha sido heredado por el feto y, por tanto, tiene muchas más probabilidades de que B esté presente en el feto en SNP 2. Un modelo que tiene en cuenta el enlace prevería esto, mientras que un modelo que no lo tiene en cuenta no.

Alternativamente, si una madre era AB en SNAP 1 y AB en SNP 2 cercano, entonces se podrían utilizar dos hipótesis correspondientes a la trisomía materna en ese lugar, una que implicaría un error en la copia coincidente (ausencia de disyunción en la meiosis II o mitosis en el desarrollo fetal temprano) y una que implicaría un error en la copia no coincidente (ausencia de disyunción en la meiosis I). En el caso de la trisomía por error en la copia coincidente, si el feto

heredase AA de la madre en SNP 1, entonces el feto tendría muchas más posibilidades de heredar AA o BB de la madre en SNP 2, pero no AB. En el caso del error en la copia no coincidente, el feto heredaría AB de la madre en los dos SNP. Las hipótesis de distribución alélica realizadas con un método de determinación del estado de ploidía que tiene en cuenta el enlace haría estas predicciones y, por tanto, se corresponde con las mediciones alélicas reales en una medida mucho mayor que en el caso de un método de determinación del estado de ploidía que no tiene en cuenta el enlace. Cabe señalar que el planteamiento del enlace no resulta posible cuando se utiliza un método que depende del cálculo de los ratios de alelos y en la suma de dichos ratios de alelos.

Una razón por la que se cree que las determinaciones del estado de ploidía que utilizan un método que comprende la comparación de las mediciones alélicas observadas con las hipótesis teóricas correspondientes a posibles estados genéticos fetales ofrecen una mayor precisión es que cuando se utiliza la secuenciación para medir los alelos este método puede averiguar más información, con respecto a otros métodos, de los datos de los alelos cuando el número total de lecturas es bajo; por ejemplo, un método que depende del cálculo y la suma de los ratios de alelos produciría un ruido estocástico desproporcionadamente ponderado. Por ejemplo, imaginemos un caso que implicase la medición de los alelos utilizando secuenciación y en el que hubiese un conjunto de loci en el que solo se detectasen cinco lecturas de secuencias para cada locus. En una realización, para cada uno de los alelos, los datos se pueden comparar con las hipótesis de distribución alélica, y ponderarse en función del número de lecturas de secuencias; por tanto, los datos de estas mediciones serían convenientemente ponderados e incorporados a la determinación total. Esto contrasta con un método que implica la cuantificación de un ratio de alelos en un locus heterocigótico, dado que este método solo podría calcular ratios de 0%, 20%, 40%, 60%, 80% o 100% como ratios de alelos posibles; ninguno de estos puede acercarse a los ratios de alelos previstos. En este último caso, los ratios de alelos calculados tendrían que ser descartados debido a lecturas insuficientes o tendrían una ponderación desproporcionada e introducirían un ruido estocástico en la determinación, reduciendo así su precisión. En una realización, las mediciones alélicas individuales pueden ser tratadas como mediciones independientes, donde la relación entre las mediciones realizadas en alelos del mismo locus no es diferente de la relación entre las mediciones realizadas en alelos de diferentes loci.

En una realización, un método divulgado en el presente documento implica la determinación de si la distribución de las mediciones alélicas observadas es indicativa de un feto euploide o aneuploide sin comparar ninguna métrica con las mediciones alélicas observadas en un cromosoma de referencia que se espera que sea disómico (denominado método RC). Esto supone una mejora significativa respecto de otros métodos, como los métodos que utilizan una secuenciación por fuerza bruta que detectan la aneuploidía evaluando la proporción de fragmentos secuenciados aleatoriamente de un cromosoma sospechoso con respecto a uno o más cromosomas de referencia que se suponen disómicos. Este método RC produce resultados incorrectos si el cromosoma de referencia supuestamente disómico no lo es en realidad. Esto puede ocurrir en casos en los que la aneuploidía es más sustancial que la trisomía de un único cromosoma o cuando el feto es triploide y todos los autosomas son trisómicos. En el caso de un feto triploide femenino (69, XXX) en efecto no existen cromosomas disómicos en absoluto. El método descrito en el presente documento no requiere un cromosoma de referencia y podría identificar correctamente cromosomas trisómicos en fetos triploides femeninos. Para cada cromosoma, hipótesis, fracción del niño y nivel de ruido, se puede adaptar un modelo de distribución conjunto, sin datos de un cromosoma de referencia, sin una estimación total de la fracción del niño y sin una hipótesis de referencia fija.

En una realización, un método divulgado en el presente documento demuestra cómo la observación de distribuciones alélicas en loci polimórficos se puede utilizar para determinar el estado de ploidía de un feto con una mayor precisión que los métodos existentes en la técnica. En una realización, el método utiliza la secuenciación focalizada para obtener genotipos materno-fetales mixtos y opcionalmente genotipos de la madre y/o del padre en una pluralidad de SNP para establecer primero las diversas distribuciones de la frecuencia alélica previstas bajo las diferentes hipótesis y, a continuación, observar la información alélica cuantitativa obtenida sobre la mezcla materno-fetal y evaluar qué hipótesis se ajusta mejor a los datos, donde el estado genético correspondiente a la hipótesis que más se ajusta a los datos se denomina el estado genético correcto. En una realización, un método divulgado en el presente documento también utiliza el grado de adecuación para generar una certeza en que el estado genético determinado es el estado genético correcto. En una realización, un método divulgado en el presente documento implica el uso de algoritmos que analizan la distribución de alelos encontrados para los loci que tienen distintos contextos parentales y la comparación de las distribuciones alélicas observadas con las distribuciones alélicas previstas para los diferentes estados de ploidía para los diferentes contextos parentales (diferentes patrones genotípicos parentales). Esto difiere y supone una mejora respecto de los métodos que no permiten la estimación del número de casos independientes de cada alelo en cada locus en una muestra materno-fetal mezclada. En una realización, un método divulgado en el presente documento implica la determinación de si la distribución de mediciones alélicas observadas es indicativa de un feto euploide o aneuploide utilizando distribuciones alélicas observadas medidas en loci donde la madre es heterocigótica. Esto difiere y supone una mejora respecto de los métodos que no utilizan las distribuciones alélicas observadas en loci donde la madre es heterocigótica porque, en los casos en los que el ADN no es preferentemente enriquecido o es preferentemente enriquecido para loci que se sabe que son altamente informativos para ese individuo diana concreto, permite el uso de aproximadamente el doble de datos de medición genética de un conjunto de datos de secuencia en la determinación de la ploidía, lo que da como resultado una determinación más precisa.

En una realización, un método divulgado en el presente documento utiliza un modelo de distribución conjunto que asume que las frecuencias alélicas en cada locus son multinomiales (y por tanto binomiales cuando los SNP son bialélicos). En algunas realizaciones el modelo de distribución conjunto utiliza distribuciones beta-binomiales. Cuando el uso de una técnica de medición, como la secuenciación, proporciona una medida cuantitativa para cada alelo presente

en cada locus, el modelo binomial se puede aplicar a cada locus y se puede averiguar el grado de las frecuencias alélicas subyacentes y la certeza en dicha frecuencia. Con los métodos conocidos en la técnica que generan determinaciones del estado de ploidía a partir de ratios de alelos o los métodos en los que la información cuantitativa de los alelos es desechada, no es posible averiguar la certidumbre del ratio observado. El método instantáneo es diferente y supone una mejora respecto de los métodos que calculan ratios de alelos y suman estos ratios para realizar una determinación del estado de ploidía, puesto que cualquier método que implica el cálculo de un ratio de alelos en un locus determinado y la suma de dichos ratios necesariamente asume que las intensidades o los recuentos medidos que son indicativos de la cantidad de ADN de cualquier alelo o locus concreto se distribuirán de forma gaussiana. El método divulgado en el presente documento no implica el cálculo de ratios de alelos. En algunas realizaciones, un método divulgado en el presente documento puede implicar la incorporación del número de observaciones de cada alelo de una pluralidad de loci en un modelo. En algunas realizaciones, un método divulgado en el presente documento puede implicar el cálculo de las propias distribuciones previstas, permitiendo el uso de un modelo de distribución binomial conjunto que puede ser más preciso que cualquier modelo que asuma una distribución gaussiana de las mediciones alélicas. La probabilidad de que el modelo de distribución binomial sea significativamente más preciso que la distribución gaussiana se incrementa a medida que aumenta el número de loci. Por ejemplo, cuando se analizan menos de 20 loci, la probabilidad de que el modelo de distribución binomial sea significativamente mejor es baja. Sin embargo, cuando se utilizan más de 100, o especialmente más de 400, o especialmente más de 1000, o especialmente más de 2000 loci, el modelo de distribución binomial tiene una probabilidad muy elevada de ser significativamente más preciso que el modelo de distribución gaussiana, lo que resulta en una determinación del estado de ploidía más preciso. La probabilidad de que el modelo de distribución binomial sea significativamente más preciso que la distribución gaussiana también se incrementa a medida que aumenta el número de observaciones en cada locus. Por ejemplo, cuando se observan menos de 10 secuencias distintas en cada locus, la probabilidad de que el modelo de distribución binomial sea significativamente mejor es baja. Sin embargo, cuando se utilizan más de 50 lecturas de secuencias, o especialmente más de 100 lecturas de secuencias, o especialmente más de 200 lecturas de secuencias, o especialmente más de 300 lecturas de secuencias, el modelo de distribución binomial tiene una probabilidad muy elevada de ser significativamente más preciso que el modelo de distribución gaussiana, lo que da como resultado una determinación del estado de ploidía más precisa.

En una realización, un método divulgado en el presente documento utiliza la secuenciación para medir el número de casos de cada alelo en cada locus de una muestra de ADN. Cada lectura de secuenciación se puede correlacionar con un locus concreto y tratarse como una lectura de secuencia binaria; alternatively, la probabilidad de la identidad de la lectura y/o correlación se puede incorporar como parte de la lectura de la secuencia, resultando en una lectura de la secuencia probabilística que es el número entero o la fracción probable de lecturas de la secuencia que se correlaciona con unos determinados loci. Utilizando recuentos binarios o probabilidad de recuentos se puede utilizar una distribución binomial para cada conjunto de mediciones, lo que permite calcular un intervalo de certeza alrededor del número de recuentos. Esta capacidad de utilizar la distribución binomial permite realizar estimaciones de la ploidía más precisas y calcular intervalos de certeza más precisos. Esto difiere y supone una mejora respecto de otros métodos que utilizan intensidades para medir la cantidad de un alelo presente, como, por ejemplo, los métodos que utilizan microarrays o los métodos que realizan mediciones utilizando lectores de fluorescencia para medir la intensidad de

ADN con etiquetas fluorescentes en bandas electroforéticas.

En una realización, un método divulgado en el presente documento utiliza aspectos del conjunto de datos presentes para determinar parámetros para la distribución de la frecuencia alélica estimada para ese conjunto de datos. Esto supone una mejora respecto a los métodos que utilizan conjuntos de datos de formación o conjuntos de datos previos para establecer parámetros para determinar las distribuciones de la frecuencia alélica prevista presente o los ratios de alelos posiblemente previstos. Esto se debe a que hay diferentes conjuntos de condiciones implicados en la recogida y medición de cada muestra genética y, por tanto, un método que utiliza datos del conjunto de datos instantáneo para determinar los parámetros para el modelo de distribución conjunto que se utilizará en la determinación del estado de ploidía para esa muestra tenderá a ser más preciso.

En una realización, un método divulgado en el presente documento implica la determinación de si la distribución de mediciones alélicas observadas es indicativa de un feto euploide o aneuploide utilizando una técnica de probabilidad máxima. El uso de la técnica de la probabilidad máxima es diferente y supone una mejora significativa respecto de los métodos que utilizan la técnica de rechazo por hipótesis única en el sentido de que las determinaciones resultantes se realizarán con una precisión significativamente mayor. Un motivo de ello es que las técnicas de rechazo por hipótesis única establecen umbrales de corte basados en una única distribución de la medición y no en dos, lo que significa que normalmente los umbrales no son óptimos. Otra razón es que la técnica de probabilidad máxima permite la optimización del umbral de corte para cada muestra individual en lugar de determinar un umbral de corte que se utilizará para todas las muestras con independencia de las características concretas de cada muestra individual. Otra razón es que el uso de la técnica de probabilidad máxima permite el cálculo de una certeza para cada determinación del estado de ploidía. La capacidad para realizar un cálculo de certeza para cada determinación permite a un médico saber qué determinaciones resultan precisas y cuáles son más probables que sean erróneas. En algunas realizaciones, se puede combinar una amplia variedad de métodos con una técnica de estimación de la probabilidad máxima para mejorar la precisión de las determinaciones del estado de ploidía. En una realización, la técnica de probabilidad máxima se puede utilizar en combinación con el método descrito en la Patente USA 7.888.017. En una realización, la técnica de la probabilidad máxima se puede utilizar en combinación con el método de utilizar la amplificación por PCR focalizada para amplificar el

ADN en la muestra mezclada seguida por la secuenciación y el análisis utilizando un método de recuento de lecturas como el utilizado por TANDEM DIAGNOSCS, presentado en el Congreso Internacional de Genética Humana celebrado en Montreal en octubre de 2011. En una realización, un método divulgado en el presente documento implica la estimación de la fracción fetal de ADN en la muestra mezclada y la utilización de dicha estimación para calcular tanto la determinación del estado de ploidía como la certeza de la determinación del estado de ploidía. Cabe señalar que esto es diferente y distinto de los métodos que utilizan la fracción fetal estimada como filtro para determinar la fracción fetal suficiente, seguida de una determinación del estado de ploidía realizada utilizando la técnica de rechazo por hipótesis única que no tiene en cuenta la fracción fetal ni produce un cálculo de la certeza para la determinación.

En una realización, un método divulgado en el presente documento tiene en cuenta la tendencia de los datos a ser ruidosos y a contener errores asignando una probabilidad a cada medición. El uso de técnicas de probabilidad máxima para seleccionar la hipótesis correcta del conjunto de hipótesis que se realizaron utilizando los datos de medición con estimaciones probabilísticas asignadas hace que sea más probable que se descuenten las mediciones incorrectas y que se utilicen las mediciones correctas en los cálculos que conducen a la determinación de la ploidía. Para ser más precisos, este método reduce de forma sistemática la influencia de los datos que son medidos incorrectamente sobre la determinación del estado de ploidía. Esto representa una mejora respecto de los métodos en los que se asume que todos los datos son igualmente correctos o de los métodos en los que los datos periféricos son excluidos de forma arbitraria de los cálculos que conducen a la determinación del estado de ploidía. Los métodos existentes que utilizan mediciones del ratio del canal afirman ampliar el método a los SNP múltiples estableciendo una media de los ratios del canal de los SNP individuales. El hecho de no ponderar los SNP individuales por la varianza de la medición prevista basada en la calidad del SNP y en la profundidad observada de la lectura reduce la precisión de la estadística resultante, lo que provoca una reducción significativa de la precisión de la determinación del estado de ploidía, en particular en los casos límite.

En una realización, un método divulgado en el presente documento no presupone el conocimiento de cuáles SNP u otros loci polimórficos son heterocigotos en el feto. Este método permite realizar una determinación del estado de ploidía en los casos en los que no se dispone de información genotípica paterna. Esto supone una mejora respecto de los métodos en los que es necesario conocer cuáles de los SNP son heterocigotos de antemano para seleccionar convenientemente los loci diana o para interpretar las mediciones genéticas realizadas sobre la muestra de ADN fetal/materno mezclada.

Los métodos descritos en el presente documento resultan particularmente ventajosos cuando se utilizan con muestras en las que hay una pequeña cantidad de ADN disponible o cuando el porcentaje de ADN fetal es bajo. Esto se debe a la correspondiente tasa de pérdida de alelos más elevada que se produce cuando solo hay una pequeña cantidad de ADN disponible y/o la correspondiente tasa de pérdida de alelos fetales más elevada que se produce cuando el porcentaje de ADN fetal es bajo en una muestra mezclada de ADN fetal y materno. Una tasa de pérdida de alelos elevada, lo que significa que un importante porcentaje de los alelos no se han medido para el individuo diana, resulta en unos cálculos de las fracciones fetales poco precisos y en unas determinaciones del estado de ploidía poco precisas. Dado que los métodos divulgados en el presente documento pueden utilizar un modelo de distribución conjunto que tiene en cuenta el enlace en los patrones de herencia entre SNP, se pueden realizar determinaciones del estado de ploidía significativamente más precisas. Los métodos descritos en el presente documento permiten realizar una determinación precisa del estado de ploidía cuando el porcentaje de moléculas de ADN que son fetales en la mezcla es inferior al 40%, inferior al 30%, inferior al 20%, inferior al 10%, inferior al 8% e incluso inferior al 6%.

En una realización, resulta posible determinar el estado de ploidía de un individuo basándose en las mediciones, cuando el ADN de dicho individuo está mezclado con ADN de un individuo relacionado. En una realización, la mezcla de ADN es el ADN flotante libre que se encuentra en el plasma materno, que puede incluir ADN de la madre, con un cariotipo conocido y un genotipo conocido, y que puede estar mezclado con ADN del feto, con cariotipo desconocido y genotipo desconocido. Resulta posible utilizar la información genotípica conocida de uno o los dos progenitores para predecir una pluralidad de potenciales estados genéticos del ADN de la muestra mezclada para diferentes estados de ploidía, diferentes contribuciones cromosómicas de cada uno de los progenitores al feto y, opcionalmente, diferentes fracciones de ADN fetal en la mezcla. Cada composición potencial se puede denominar una hipótesis. El estado de ploidía del feto se puede determinar a continuación analizando las mediciones reales y determinando qué composiciones potenciales resultan más probables dados los datos observados.

En otra parte de este documento se recoge una exposición más detallada de los anteriores puntos.

Diagnóstico prenatal no invasivo (NPD)

El proceso de diagnóstico prenatal no invasivo implica una serie de pasos. Algunos de los pasos pueden incluir: (1) obtención de material genético del feto; (2) enriquecimiento del material genético del feto que puede estar en una muestra mezclada, ex vivo; (3) amplificación del material genético, ex vivo; (4) enriquecimiento preferente de loci específicos en el material genético, ex vivo; (5) medición del material genético, ex vivo; y (6) análisis de los datos genotípicos, en un ordenador y ex vivo. Los métodos para reducir la práctica de estos seis pasos y otros relevantes se describen en el presente documento. Al menos algunos de los pasos del método no se aplican directamente sobre el organismo. En una realización, la presente divulgación se refiere a métodos de tratamiento y diagnóstico aplicados al tejido y otros materiales biológicos aislados y separados del organismo. Al menos algunos de los pasos del método se ejecutan en un ordenador.

Algunas realizaciones de la presente divulgación permiten a un médico determinar el estado genético de un feto que se está gestando en una madre de forma no invasiva, de modo que la salud del bebé no se pone en peligro por la recogida del material genético del feto y que la madre no se tiene que someter a un procedimiento invasivo. Por otra parte, en determinados aspectos, la presente divulgación permite determinar el estado genético fetal con una elevada precisión, una precisión significativamente mayor que, por ejemplo, las pruebas basadas en el análisis del suero materno no invasivas, como la prueba triple, que se utilizan generalmente en la atención prenatal.

La alta precisión de los métodos divulgados en el presente documento es resultado de un planteamiento informático para el análisis de los datos del genotipo, tal y como se describe en el presente documento. Los avances tecnológicos recientes han permitido medir grandes cantidades de información genética a partir de una muestra genética utilizando métodos como la secuenciación de alto rendimiento y los arrays para la determinación del genotipo. Los métodos divulgados en el presente documento permiten a un médico sacar mayor partido de las grandes cantidades de datos disponibles y realizar un diagnóstico más preciso del estado genético fetal. Los detalles de una serie de realizaciones se proporcionan a continuación. Las diferentes realizaciones pueden implicar diferentes combinaciones de los mencionados pasos. Se pueden utilizar de forma intercambiable diversas combinaciones de las diferentes realizaciones de los distintos pasos.

En una realización, se toma una muestra de sangre de una madre embarazada, y el ADN flotante libre en el plasma de la sangre de la madre, que contiene una mezcla de ADN de origen materno y ADN de origen fetal, que es aislado y utilizado para determinar el estado de ploidía del feto. En una realización, un método divulgado en el presente documento implica el enriquecimiento preferente de las secuencias de ADN en una mezcla de ADN que corresponden a los alelos polimórficos de forma que los ratios de alelos y/o las distribuciones de alelos se mantienen fundamentalmente constantes tras el enriquecimiento. En una realización, un método divulgado en el presente documento implica la amplificación basada en la PCR focalizada de alta eficiencia de forma que un porcentaje muy elevado de las moléculas resultantes corresponden a los loci diana. En una realización, un método divulgado en el presente documento implica la secuenciación de una mezcla de ADN que contiene tanto ADN de origen materno como ADN de origen fetal. En una realización, un método divulgado en el presente documento implica el uso de distribuciones de alelos medidas para determinar el estado de ploidía de un feto que se está gestando en una madre. En una realización, un método divulgado en el presente documento implica la comunicación del estado de ploidía determinado a un médico. En una realización, un método divulgado en el presente documento implica emprender una acción clínica, por ejemplo, realizar pruebas de seguimiento invasivas como una amniocentesis o muestreo de vello coriónico, prepararse para el nacimiento de un individuo trisómico o la terminación electiva de un feto trisómico.

Esta aplicación hace referencia a la Solicitud de patente USA con el número de serie 11/603.406, presentada el 28 de noviembre de 2006 (Publicación USA con el número: 20070184467); La Solicitud de patente USA con el número de serie 12/076.348, presentada el 17 de marzo de 2008 (Publicación USA con el número: 20080243398); Solicitud PCT con el número de serie PCT/US09/52730, presentada el 4 de agosto de 2009 (Publicación PCT con el número: W0/2010/017214); Solicitud PCT con el número de serie PCT/US10/050824, presentada el 30 de septiembre de 2010 (Publicación PCT con el número: W0/2011/041485), Solicitud de patente USA con el número de serie 13/110.685, presentada el 18 de mayo de 2011 (Publicación USA con el número: 20110288780), y la Solicitud PCT con el número de serie PCT/12/58578, presentada el 3 de octubre de 2012 (Publicación PCT con el número: W0/2013/052557). Parte del vocabulario utilizado en este documento puede tener sus antecedentes en estas referencias. Parte de los conceptos descritos en el presente documento se pueden entender mejor a la luz de los conceptos que se encuentran en estas referencias.

Análisis de sangre materna que contiene ADN fetal flotante libre

Los métodos empleados en el presente documento se pueden utilizar para ayudar a determinar el genotipo de un niño, feto u otro individuo diana cuando el material genético de la diana se encuentra en presencia de una cantidad de otro material genético. En algunas realizaciones, el genotipo se puede referir al estado de ploidía de uno o una pluralidad de cromosomas, se puede referir a uno o una pluralidad de alelos vinculados a una enfermedad o alguna combinación de estos. En la presente divulgación, el debate se centra en determinar el estado genético de un feto cuando el ADN fetal se encuentra en la sangre materna, pero este ejemplo no pretende limitar los posibles contextos en los que se puede aplicar este método. Por otra parte, el método puede resultar aplicable en casos en los que la cantidad de ADN diana se encuentra en cualquier proporción con el ADN no diana; por ejemplo, el ADN diana podría representar cualquier proporción entre el 0,000001 y el 99,999999% del ADN presente. Asimismo, el ADN no diana no tiene por qué ser necesariamente de un individuo, o incluso de un individuo relacionado, siempre que los datos genéticos de alguno o todos los individuos no diana relevantes sean conocidos. En una realización, un método divulgado en el presente documento se puede utilizar para determinar los datos genotípicos de un feto a partir de la sangre materna que contiene ADN fetal. También se puede utilizar en un caso en el que hay múltiples fetos en el útero de una mujer embarazada o cuando pueda haber otro ADN contaminante en la muestra, por ejemplo, de otros hermanos ya nacidos.

Esta técnica puede hacer uso del fenómeno de las células de la sangre fetal consiguiendo acceder a la circulación materna a través del vello de la placenta. Normalmente solo un pequeño número de células fetales entra en la circulación materna de esta forma (no suficiente para producir una prueba de Kleihauer-Betke para la hemorragia fetal-materna). Las células fetales se pueden clasificar y analizar a través de diversas técnicas para buscar secuencias de ADN concretas, pero sin los riesgos que implican inherentemente los procedimientos invasivos. Esta técnica también puede hacer uso del fenómeno del ADN fetal flotante libre obteniendo acceso a la circulación materna a través de la

liberación de ADN tras la apoptosis de tejido placentario cuando el tejido placentario en cuestión contiene ADN del mismo genotipo que el feto. Se ha demostrado que el ADN flotante libre que se encuentra en el plasma materno contiene ADN fetal en proporciones de hasta el 30-40% de ADN fetal.

En una realización, se puede extraer sangre a una mujer embarazada. La investigación ha demostrado que la sangre materna puede contener una pequeña cantidad de ADN flotante libre del feto, además de ADN flotante libre de origen materno. Por otra parte, también puede haber células sanguíneas fetales enucleadas que contienen ADN de origen fetal, además de múltiples células sanguíneas de origen materno, que típicamente no contienen ADN nuclear. Existen métodos conocidos en la técnica para aislar ADN fetal o crear fracciones enriquecidas en ADN fetal. Por ejemplo, se ha demostrado que la cromatografía crea determinadas fracciones enriquecidas en ADN fetal.

Una vez que se dispone de la muestra de sangre, plasma u otro fluido materno, extraído de manera relativamente no invasiva y que contiene una cantidad de ADN fetal, sea celular o flotante libre, enriquecido en su proporción respecto del ADN materno o en su proporción original, se puede determinar el genotipo del ADN encontrado en dicha muestra. En algunas realizaciones, la sangre se puede extraer utilizando una aguja para extraer sangre de una vena, por ejemplo, la vena basilica. El método descrito en el presente documento se puede utilizar para determinar los datos genotípicos del feto. Por ejemplo, se pueden utilizar para determinar el estado de ploidía de uno o más cromosomas, se pueden utilizar para determinar la identidad de uno o un conjunto de SNP, incluyendo inserciones, deleciones y translocaciones. Se pueden utilizar para determinar uno o más haplotipos, incluyendo el progenitor de origen de una o más características genotípicas.

Cabe señalar que este método funcionará con cualquiera de los ácidos nucleicos que se pueden utilizar para cualesquiera métodos de secuenciación o determinación del genotipo, como la plataforma ILLUMINA INFINIUM ARRAY, AFFYMETRIX GENECHIP, ILLUMINA GENOME ANALYZER, o LIFE TECHNOLOGIES' SOLID SYSTEM. Este incluye ADN flotante libre extraído del plasma o amplificaciones (por ejemplo, amplificación de todo el genoma, PCR) de este; ADN genómico de otros tipos de célula (por ejemplo, linfocitos humanos de toda la sangre) o amplificaciones de este. Para la preparación del ADN, también funcionará cualquier extracción o método de purificación que genere ADN genómico adecuado para una de estas plataformas. Este método podría funcionar igualmente bien con muestras de ARN. En una realización, el almacenamiento de muestras se puede realizar de forma que se minimice la degradación (por ejemplo, mediante congelación a unos -20 °C o a una temperatura inferior).

Algunas realizaciones se pueden utilizar en combinación con el método PARENTAL SUPPORT™ (PS), cuyas realizaciones se describen en la Solicitud USA con el número 11/603.406 (Publicación USA con el número: 20070184467), Solicitud USA con el número 12/076.348 (Publicación USA con el número: 20080243398), Solicitud USA 13/110.685 (Publicación USA con el número: 20110288780), Solicitud PCT PCT/US09/52730 (Publicación PCT con el número: W0/2010/017214), y Solicitud PCT con el número PCT/US10/050824 (Publicación PCT con el número: WO/2011/041485). PARENTAL SUPPORT™ es un planteamiento basado en la informática que se puede utilizar para analizar datos genéticos. En algunas realizaciones, los métodos divulgados en el presente documento se pueden considerar parte del método PARENTAL SUPPORT™. En algunas realizaciones, el método PARENTAL SUPPORT™ es un conjunto de métodos que pueden ser utilizados para determinar los datos genéticos del individuo diana, con alta precisión, de una célula o un número reducido de células de ese individuo, o una mezcla de ADN compuesto por ADN del individuo diana y ADN de uno o una pluralidad de otros individuos, específicamente para determinar alelos vinculados a una enfermedad, otros alelos de interés, y/o el estado de ploidía de uno o una pluralidad de cromosomas en el individuo diana. PARENTAL SUPPORT™ se puede referir a cualquiera de estos métodos. PARENTAL SUPPORT™ es un ejemplo de método basado en la informática. Ejemplos de realizaciones del método PARENTAL SUPPORT™ se muestran en las Figuras 29-31G y se describen en el Experimento 19.

El método PARENTAL SUPPORT™ utiliza datos genéticos parentales conocidos, es decir, datos genéticos haplotípicos y/o diploides de la madre y/o el padre, junto con el conocimiento del mecanismo de la meiosis y la medición imperfecta del ADN diana, y posiblemente de uno o más individuos relacionados, junto con frecuencias de cruce basadas en la población, para reconstruir, in silico, el genotipo de una pluralidad de alelos, y/o el estado de ploidía de un embrión o de cualquier célula o células diana, y el ADN diana en la ubicación de loci clave con un alto grado de certeza. El método PARENTAL SUPPORT™ puede reconstruir no solamente polimorfismos de un solo nucleótido (SNP) que se midieron deficientemente, sino también inserciones y deleciones, y SNP o regiones completas de ADN que no se midieron en absoluto. Además, el método PARENTAL SUPPORT™ puede medir múltiples loci vinculados a una enfermedad, y también realizar un cribado de aneuploidía, partiendo de una sola célula. En algunas realizaciones, el método PARENTAL SUPPORT™ puede ser utilizado para caracterizar una o más células de embriones a los que se ha realizado una biopsia durante un ciclo de IVF para determinar la condición genética de la o las células.

El método PARENTALSUPPORT™ permite limpiar datos genéticos ruidosos. Esto puede hacerse deduciendo los correctos alelos genéticos en el genoma diana (embrión) utilizando el genotipo de individuos relacionados (padres) como referencia. PARENTAL SUPPORT™ puede ser especialmente relevante cuando se dispone solamente de una pequeña cantidad de material genético (por ejemplo, PGD) y donde las mediciones directas de los genotipos son inherentemente ruidosas, debido a las limitadas cantidades de material genético. PARENTAL SUPPORT™ puede ser especialmente relevante cuando solamente una pequeña fracción del material genético disponible procede del individuo diana (por ejemplo, NPD) y donde las mediciones directas de los genotipos son inherentemente ruidosas, debido a la señal del ADN contaminante de otro individuo. El método PARENTAL SUPPORT™ puede reconstruir secuencias de alelos diploides ordenadas con alta precisión en el embrión, junto con el número de copias de segmentos de

cromosomas, aunque las mediciones diploides convencionales no ordenadas pueden caracterizarse por elevadas tasas de pérdidas de alelos, inclusiones, sesgos de amplificación variable y otros errores. El método puede emplear un modelo genético subyacente, y un modelo subyacente de medición de error. El modelo genético puede determinar probabilidades de alelos en cada SNP y probabilidades de cruce entre SNP. Las probabilidades de los alelos pueden ser modeladas en cada SNP en base a datos obtenidos de los padres, y modelar probabilidades de cruce entre SNP en base a datos obtenidos de la base de datos HapMap, desarrollada por el International HapMap Project. Con el modelo genético subyacente y el modelo de error de medición, puede utilizarse una estimación máxima a posteriori (MAP), con modificaciones para eficiencia informática, para calcular los valores de alelos ordenados correctos en cada SNP en el embrión.

En algunos casos las técnicas anteriormente mencionadas pueden determinar el genotipo de un individuo con una cantidad muy pequeña de ADN procedente de ese individuo. Este podría ser ADN de una o de un pequeño número de células o podría proceder de la pequeña cantidad de ADN fetal que se encuentra en la sangre materna.

Hipótesis

En el contexto de esta divulgación, una hipótesis se refiere a un posible estado genético. Puede referirse a un posible estado de ploidía. Puede referirse a un posible estado alélico. Un conjunto de hipótesis se puede referir a un conjunto de posibles estados genéticos, un conjunto de posibles estados de ploidía o combinaciones de estos. En algunas realizaciones, se puede diseñar un conjunto de hipótesis de forma que una de las hipótesis del conjunto corresponda al estado genético real de un individuo determinado. En algunas realizaciones, se puede diseñar un conjunto de hipótesis de forma que todo posible estado genético pueda ser descrito por lo menos por una hipótesis del conjunto. En algunas realizaciones de la presente divulgación, un aspecto del método consiste en determinar qué hipótesis corresponde al estado genético real del individuo en cuestión.

En otra realización de la presente divulgación, un paso incluye la creación de una hipótesis. En algunas realizaciones puede ser una hipótesis del número de copias. En algunas realizaciones puede incluir una hipótesis sobre qué segmentos de un cromosoma de cada uno de los individuos relacionados corresponde genéticamente a qué segmentos, de haberlos, de los otros individuos relacionados. Crear una hipótesis puede referirse al hecho de establecer los límites de las variables, de forma que la totalidad del conjunto de posibles estados genéticos que están siendo considerados estén comprendidos en esas variables.

Una "hipótesis de número de copias", denominada también una "hipótesis de ploidía", o una "hipótesis de estado de ploidía", puede referirse a una hipótesis relacionada con un posible estado de ploidía para un cromosoma determinado, un tipo de cromosoma, o sección de un cromosoma, en el individuo diana. Puede referirse también al estado de ploidía en más de uno de los tipos de cromosomas del individuo.

Un conjunto de hipótesis de número de copias puede referirse a un conjunto de hipótesis donde cada hipótesis corresponde a un posible estado de ploidía distinto en un individuo. Un conjunto de hipótesis se puede referir a un conjunto de posibles estados de ploidía, un conjunto de posibles contribuciones de haplotipos parentales, un conjunto de posibles porcentajes de ADN fetal en la muestra mezclada o combinaciones de estos.

Un individuo normal contiene uno de cada tipo de cromosoma de cada progenitor. No obstante, debido a errores en meiosis y mitosis, es posible que un individuo tenga 0, 1, 2, o más de un tipo de cromosoma determinado de cada progenitor. En la práctica, es poco frecuente ver más de dos de un cromosoma determinado de un progenitor. En esta divulgación, algunas realizaciones solo consideran las hipótesis posibles en las que 0, 1, o 2 copias de un cromosoma determinado proceden de un progenitor; es una extensión trivial considerar más o menos copias posibles procedentes de un progenitor. En algunas realizaciones, para un cromosoma determinado hay nueve posibles hipótesis: las tres hipótesis posibles referentes a 0, 1, o 2 cromosomas de origen materno, multiplicado por las tres hipótesis posibles sobre 0, 1, o 2 cromosomas de origen paterno. Consideremos que (m, f) se refiere a la hipótesis en la que m es el número de un cromosoma determinado heredado de la madre, y f es el número de un cromosoma determinado heredado del padre. En consecuencia, las nueve hipótesis son (0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), y (2,2). Estas también se pueden escribir como H_{00} , H_{01} , H_{02} , H_{10} , H_{11} , H_{12} , H_{20} , H_{21} , y H_{22} . Las distintas hipótesis corresponden a diferentes estados de ploidía. Por ejemplo, (1,1) se refiere a un cromosoma disómico normal; (2,1) se refiere a una trisomía materna, y (0,1) se refiere a una monosomía paterna. En algunas realizaciones, el caso en el que dos cromosomas son heredados de un progenitor y un cromosoma del otro puede diferenciarse además en dos casos: uno en el que los dos cromosomas son idénticos (error de copias emparejadas), y uno en el que los dos cromosomas son homólogos pero no idénticos (error de copias no emparejadas). En estas realizaciones, hay dieciséis hipótesis posibles. Cabe señalar que se pueden utilizar otros conjuntos de hipótesis y un número diferente de hipótesis.

En algunas realizaciones de la presente divulgación, la hipótesis de la ploidía puede referirse a una hipótesis sobre qué cromosoma de otros individuos relacionados corresponde a un cromosoma hallado en el genoma del individuo diana. En algunas realizaciones, una clave del método es el hecho de que cabe esperar que individuos relacionados compartan bloques haplotípicos, y utilizando datos genéticos medidos de individuos relacionados, junto con el conocimiento de qué bloques haplotípicos coinciden entre el individuo diana y el individuo relacionado, es posible inferir los datos genéticos correctos de un individuo diana con mayor certeza que utilizando solamente las mediciones genéticas del individuo diana. Como tal, en algunas realizaciones la hipótesis de ploidía puede referirse no solamente al número de cromosomas, sino también a qué cromosomas en individuos relacionados son idénticos, o casi idénticos, a uno o más cromosomas del individuo diana.

Una vez se ha definido el conjunto de hipótesis, cuando los algoritmos operan sobre los datos genéticos de entrada, pueden dar como resultado una probabilidad estadística determinada para cada una de las hipótesis consideradas. Las probabilidades de las diversas hipótesis pueden determinarse calculando matemáticamente, para cada una de las distintas hipótesis, el valor de la probabilidad, como lo indican una o más de las técnicas especializadas, los algoritmos, y/o los métodos descritos en otra parte de esta divulgación, utilizando como entrada los datos genéticos pertinentes.

Una vez calculadas las probabilidades de las distintas hipótesis, como se haya determinado por diversas técnicas, se pueden combinar. Esto puede implicar multiplicar para cada hipótesis las probabilidades determinadas mediante cada técnica. El producto de las probabilidades de las hipótesis puede ser normalizado. Hay que advertir que una hipótesis de ploidía se refiere a un posible estado de ploidía de un cromosoma.

El proceso de “combinación de probabilidades”, denominado también “hipótesis combinadas”, o combinar los resultados de técnicas especializadas, es un concepto que debe resultar familiar al experto en la técnica del álgebra lineal. Una posible forma de combinar probabilidades es como sigue: cuando se utiliza una técnica especializada para evaluar un conjunto de hipótesis en un conjunto determinado de datos genéticos, el resultado del método es un conjunto de probabilidades asociadas, de forma uno-a-uno, a cada hipótesis del conjunto de hipótesis. Cuando un conjunto de probabilidades que han sido determinadas por una primera técnica especializada, cada una de las cuales está asociada a una de las hipótesis del conjunto, se combina con un conjunto de probabilidades determinadas por una segunda técnica especializada, cada una de las cuales va asociada con el mismo conjunto de hipótesis, los dos conjuntos de probabilidades se multiplican. Esto significa que, para cada hipótesis del conjunto, las dos probabilidades asociadas a esa hipótesis, determinada por los dos métodos especializados, se multiplican juntas, y el producto correspondiente es el resultado de probabilidades. Este proceso puede ampliarse a cualquier número de técnicas especializadas. Si se utiliza solamente una técnica especializada, las probabilidades de salida son las mismas que las de entrada. Si se utilizan más de dos técnicas especializadas, las probabilidades pertinentes pueden multiplicarse al mismo tiempo. Los productos pueden normalizarse de forma que las probabilidades de las hipótesis en el conjunto de hipótesis sumen 100%.

En algunas realizaciones, si las probabilidades combinadas de una hipótesis determinada son mayores que las probabilidades combinadas de cualquiera de las otras hipótesis, puede considerarse que esa hipótesis se determina como la más probable. En algunas realizaciones, se puede determinar una hipótesis como la más probable, y el estado de ploidía, u otro estado genético puede ser determinado si la probabilidad normalizada es superior a un umbral. En una realización, esto puede significar que el número y la identidad de los cromosomas asociados a esa hipótesis pueden ser determinados como el estado de ploidía. En una realización, esto puede significar que la identidad de los alelos que están asociados a esa hipótesis puede ser determinada como el estado alélico. En algunas realizaciones el umbral puede situarse entre el 50% y aproximadamente el 80%. En algunas realizaciones el umbral puede situarse entre el 80% y aproximadamente el 90%. En algunas realizaciones el umbral puede situarse entre el 90% y aproximadamente el 95%. En algunas realizaciones el umbral puede situarse entre el 95% y aproximadamente el 99%. En algunas realizaciones el umbral puede situarse por encima de aproximadamente el 99,9%.

Contextos parentales

El contexto parental se refiere al estado genético de un alelo determinado, en cada uno de los dos cromosomas relevantes para cada uno de los dos progenitores de la diana. Hay que advertir que en una realización, el contexto parental no se refiere al estado alélico de la diana, sino al estado alélico de los padres. El contexto parental de un SNP determinado puede consistir en cuatro pares de bases, dos paternos y dos maternos; pueden ser iguales o distintos entre sí. Esto viene expresado típicamente como “ $m_1n_2|f_1f_2$ ” donde m_1 y m_2 son el estado genético del SNP concreto en los dos cromosomas maternos, y f_1 y f_2 son el estado genético de dicho SNP en los dos cromosomas paternos. En algunas realizaciones, el contexto parental puede venir expresado como “ $f_1f_2|m_1m_2$.” Hay que señalar que los subíndices “1” y “2” se refieren al genotipo, en ese alelo determinado, del primer y el segundo cromosoma; véase también que la elección de qué cromosoma se etiqueta como “1” y cuál como “2” es arbitraria.

Hay que señalar que, en esta divulgación, A y B se utilizan frecuentemente para representar de forma genérica identidades de pares de bases; A o B podrían representar igualmente bien a C (citocina), G (guanina), A (adenina) o T (timina). Por ejemplo, si en un alelo basado en SNP determinado, el genotipo materno fue T en ese SNP de un cromosoma, y G en ese SNP del cromosoma homólogo, y el genotipo paterno en ese alelo es G en ese SNP en ambos cromosomas homólogos, se podría decir que el alelo del individuo diana tiene el contexto parental de AB|BB; también se podría decir que el alelo tiene el contexto parental de AB|AA. Tenga en cuenta que, en teoría, cualquiera de los cuatro nucleótidos posibles podría darse en un alelo determinado, y así es posible, por ejemplo, que la madre tenga un genotipo de AT, y el padre tenga un genotipo de GC en un alelo determinado. No obstante, datos empíricos indican que en la mayoría de los casos solo dos de los cuatro posibles pares de bases se observan en un alelo determinado. Resulta posible, por ejemplo, cuando se utilizan repeticiones en tándem únicas, tener más de dos, más de cuatro e incluso más de diez contextos parentales. En esta divulgación, en la discusión se supone que se observarán solamente dos posibles pares de bases en un alelo determinado, aunque las realizaciones divulgadas en el presente documento se podrían modificar para tener en cuenta los casos en los que este supuesto no se sostiene.

Un “contexto parental” puede referirse a un conjunto o subconjunto de SNP diana que tienen el mismo contexto parental. Por ejemplo, si hubiera que medir 1000 alelos en un cromosoma determinado en un individuo diana, el contexto AA|BB podría referirse al conjunto de todos los alelos en el grupo de 1000 alelos donde el genotipo de la

madre de la diana era homocigoto, y el genotipo del padre de la diana es homocigoto, pero donde el genotipo materno y el genotipo paterno son distintos en ese locus. Si los datos parentales no están por fases, y por tanto $AB = BA$, hay nueve contextos parentales posibles: $AA|AA$, $AA|AB$, $AA|BB$, $AB|AA$, $AB|AB$, $AB|BB$, $BB|AA$, $BB|AB$, y $BB|BB$. Si los datos parentales están ajustados por fases, y por tanto $AB \neq BA$, hay dieciséis posibles contextos parentales distintos:

5 $AA|AA$, $AA|AB$, $AA|BA$, $AA|BB$, $AB|AA$, $AB|AB$, $AB|BA$, $AB|BB$, $BA|AA$, $BA|AB$, $BA|BA$, $BA|BB$, $BB|AA$, $BB|AB$, $BB|BA$, y $BB|BB$. Cada alelo SNP de un cromosoma, excluyendo algunos SNP en los cromosomas sexuales, tiene uno de esos contextos parentales. El conjunto de SNP donde el contexto parental en un progenitor es heterocigoto puede ser denominado el contexto heterocigoto.

Uso de los contextos parentales en NPD

10 El diagnóstico prenatal no invasivo es una técnica importante que se puede utilizar para determinar el estado genético de un feto a partir del material genético que se obtiene de manera no invasiva, por ejemplo, de la sangre extraída a la madre embarazada. La sangre se podría separar y el plasma aislarse, para después aislar el ADN del plasma. La selección del tamaño se podría utilizar para aislar el ADN de la longitud apropiada. El ADN se puede enriquecer preferentemente en un conjunto de loci. Este ADN se puede entonces medir a través de varios medios, como mediante

15 hibridación con un array para la determinación del genotipo y midiendo la fluorescencia o secuenciación en un secuenciador de alto rendimiento.

Cuando se utiliza la secuenciación para determinar el estado de ploidía de un feto en el contexto del diagnóstico prenatal no invasivo, existen varias formas de utilizar los datos de la secuencia. La forma más habitual en que se podrían utilizar los datos de la secuencia consiste simplemente en contar el número de lecturas que corresponden a un

20 determinado cromosoma. Por ejemplo, imagine que está intentando determinar el estado de ploidía del cromosoma 21 del feto. Imagine también que el ADN de la muestra se compone de un 10% de ADN de origen fetal y un 90% de ADN de origen materno. En este caso, se podría analizar el número medio de lecturas en un cromosoma que cabe esperar que sea disómico, por ejemplo, el cromosoma 3, y compararlo con el número de lecturas del cromosoma 21, donde las lecturas se ajustan para el número de pares de bases de ese cromosoma que forman parte de una secuencia única. Si

25 el feto fuese euploide, cabría esperar que la cantidad de ADN por unidad de genoma fuese aproximadamente la misma en todas las ubicaciones (con sujeción a variaciones estocásticas). Por otra parte, si el feto fuese trisómico en el cromosoma 21, entonces cabría esperar que hubiese algo más de ADN por unidad genética del cromosoma 21 que en las demás ubicaciones del genoma.

Específicamente cabría esperar que hubiese aproximadamente un 5% más de ADN del cromosoma 21 en la mezcla.

30 Cuando se utiliza la secuenciación para medir el ADN, cabría esperar aproximadamente un 5% más de lecturas únicamente susceptibles de correlación con el cromosoma 21 por segmento único que de los otros cromosomas. Se podría utilizar la observación de una cantidad de ADN de un cromosoma concreto que supere un determinado umbral, cuando está ajustada para el número de secuencias que son únicamente susceptibles de correlación con ese cromosoma, como base para el diagnóstico de la aneuploidía. Otro método que se podría utilizar para detectar la

35 aneuploidía es similar al anterior, salvo por el hecho de que se podrían tener en cuenta los contextos parentales.

A la hora de plantearse los alelos a focalizar, se podría considerar la probabilidad de que algunos contextos parentales ofrezcan más información que otros. Por ejemplo, $AA|BB$ y el contexto simétrico $BB|AA$ son los contextos más informativos, porque se sabe que el feto lleva un alelo que es diferente de la madre. Por razones de simetría, tanto los contextos $AA|BB$ como $BB|AA$ se pueden denominar $AA|BB$. Otro conjunto de contextos parentales informativos son

40 $AA|AB$ y $BB|AB$, porque en estos casos el feto tiene un 50% de posibilidades de llevar un alelo que la madre no tiene. Por razones de simetría, tanto los contextos $AA|AB$ como $BB|AB$ se pueden denominar $AA|AB$. Un tercer conjunto de contextos parentales informativos son $AB|AA$ y $AB|BB$, porque en estos casos el feto lleva un alelo paterno conocido, y el alelo también se encuentra presente en el genoma materno. Por razones de simetría, tanto los contextos $AB|AA$ como $AB|BB$ se pueden denominar $AB|AA$. Un cuarto contexto parental es $AB|AB$ cuando el feto tiene un estado alélico

45 desconocido y, con independencia del estado alélico, es uno en el que la madre tiene los mismos alelos. El quinto contexto parental es $AA|AA$ donde la madre y el padre son heterocigotos

Diferentes implementaciones de las realizaciones divulgadas en el presente documento

En el presente documento se divulgan métodos para determinar el estado de ploidía de un individuo diana. El individuo diana puede ser un blastómero, un embrión o un feto. En algunas realizaciones de la presente divulgación, un método

50 para determinar el estado de ploidía de uno o más cromosomas en un individuo diana puede incluir cualquiera de los pasos descritos en este documento y combinaciones de estos.

En algunas realizaciones, la fuente del material genético que se va a utilizar para determinar el estado genético del feto puede ser células fetales, como glóbulos rojos fetales nucleados, aislados de la sangre materna. El método puede implicar la obtención de una muestra de sangre de la madre embarazada. El método puede implicar el aislamiento de un

55 glóbulo rojo fetal utilizando técnicas visuales, basadas en la idea de que una determinada combinación de colores está exclusivamente asociada con un glóbulo rojo nucleado, y una combinación de colores similar no está asociada a ninguna otra célula presente en la sangre materna. La combinación de colores asociada con los glóbulos rojos nucleados puede incluir el color rojo de la hemoglobina alrededor del núcleo, color que se puede diferenciar mediante tñido, y el color del material nuclear que puede ser tñido, por ejemplo, de azul. Al aislar las células de la sangre

60 materna y extenderlas sobre un portaobjetos para después identificar aquellos puntos en los que se aprecia tanto el rojo (de la hemoglobina) como el azul (del material nuclear), resulta posible identificar la ubicación de los glóbulos rojos

nucleados. A continuación, se pueden extraer estos glóbulos rojos nucleados utilizando un micromanipulador, utilizando técnicas de determinación del genotipo y/o secuenciación para medir aspectos del genotipo del material genético que contienen esas células.

En una realización, se pueden teñir los glóbulos rojos nucleados con un tinte que solo emite fluorescencia en presencia de hemoglobina fetal y no de hemoglobina materna, para de este modo eliminar la ambigüedad entre los glóbulos rojos nucleados obtenidos de la madre o del feto. Algunas realizaciones de la presente divulgación pueden implicar el teñido u otro tipo de marcado del material nuclear. Algunas realizaciones de la presente divulgación pueden implicar específicamente el marcado del material nuclear fetal utilizando anticuerpos específicos para las células fetales.

Hay muchas otras maneras de aislar las células fetales de la sangre materna o el ADN fetal de la sangre materna, o de enriquecer muestras de material genético fetal en presencia de material genético materno. Algunos de los métodos se recogen aquí, aunque no se deberá considerar una lista exhaustiva. Algunas técnicas apropiadas se enumeran aquí por conveniencia: uso de anticuerpos etiquetados con fluorescencia u otro marcado, cromatografía de exclusión por tamaño, etiquetas de afinidad etiquetadas magnéticamente o de otro modo, diferencias epigenéticas, como metilación diferencial entre las células maternas y fetales en alelos específicos, centrifugación en gradiente de densidad seguida por reducción de CD45/14 y selección CD71-positiva de las células CD45/14 negativas, gradientes de Percoll sencillos o dobles con diferentes osmolaridades, o método de lectina específico de galactosa.

En una realización de la presente divulgación, el individuo diana es un feto y las diferentes mediciones del genotipo se realizan en una pluralidad de muestras de ADN del feto. En algunas realizaciones de la presente divulgación, las muestras de ADN fetal proceden de células fetales aisladas donde las células fetales pueden estar mezcladas con células maternas. En algunas realizaciones de la presente divulgación, las muestras de ADN fetal proceden de ADN fetal flotante libre, donde el ADN fetal puede estar mezclado con ADN materno flotante libre. En algunas realizaciones, las muestras de ADN fetal se pueden obtener de plasma materno o sangre materna que contiene una mezcla de ADN materno y ADN fetal. En algunas realizaciones, el ADN fetal puede estar mezclado con ADN materno en ratios materno:fetales que oscilan entre 99,9:0,1% a 99:1%; 99:1% a 90:10%; 90:10% a 80:20%; 80:20% a 70:30%; 70:30% a 50:50%; 50:50% a 10:90%; o 10:90% a 1:99%; 1:99% a 0,1:99,9%.

Los datos genéticos del individuo diana y/o del individuo relacionado se pueden transformar de un estado molecular a un estado electrónico, midiendo el material genético adecuado utilizando herramientas o técnicas seleccionadas de un grupo, incluyendo, entre otras, microarrays para la determinación del genotipo y secuenciación de alto rendimiento. Algunos métodos de secuenciación de alto rendimiento incluyen la secuenciación de ADN de Sanger, la pirosecuenciación, la plataforma ILLUMINA SOLEXA, GENOME ANALYZER de ILLUMINA, o la plataforma de secuenciación 454 de APPLIED BIOSYSTEM, la plataforma TRUE SINGLE MOLECULE SEQUENCING de HELICOS, el método de secuenciación por microscopio de electrones de HALCYON MOLECULAR, o cualquier otro método de secuenciación. Todos estos métodos transforman físicamente los datos genéticos almacenados en una muestra de ADN en un conjunto de datos genéticos que se almacenan típicamente en un dispositivo de memoria que va a ser procesado.

Los datos genéticos de un individuo relevante se pueden medir analizando sustancias tomadas de un grupo, incluyendo, entre otras: el tejido diploide bruto del individuo, una o más células diploides del individuo, una o más células haploides del individuo, uno o más blastómeros del individuo diana, material genético extracelular encontrado en el individuo, material genético extracelular del individuo encontrado en la sangre materna, células del individuo encontradas en la sangre materna, uno o más embriones creados de un gameto o gametos del individuo relacionado, uno o más blastómeros tomados de este embrión, material genético extracelular encontrado en el individuo relacionado, material genético que se sabe que procede del individuo relacionado y combinaciones de estos.

En algunas realizaciones, se puede crear un conjunto de al menos una hipótesis del estado de ploidía para cada uno de los tipos de cromosomas de interés del individuo diana. Cada una de las hipótesis del estado de ploidía se puede referir a un posible estado de ploidía del cromosoma o segmento del cromosoma del individuo diana. El conjunto de hipótesis puede incluir algunos o todos los posibles estados de ploidía que cabe esperar que presente el cromosoma del individuo diana. Algunos de los posibles estados de ploidía pueden incluir nulisomía, monosomía, disomía, disomía uniparental, euploidía, trisomía, trisomía emparejada, trisomía no emparejada, trisomía materna, trisomía paterna, tetrasomía, tetrasomía equilibrada (2:2), tetrasomía no equilibrada (3:1), pentasomía, hexasomía, otra aneuploidía y combinaciones de estas. Cualquiera de estos estados de aneuploidía puede ser una aneuploidía mixta o parcial, como translocaciones no equilibradas, translocaciones equilibradas, translocaciones robertsonianas, recombinaciones, deleciones, inserciones, cruces y combinaciones de estas.

En algunas realizaciones, el conocimiento del estado de ploidía determinado se puede utilizar para tomar una decisión clínica. Este conocimiento, típicamente almacenado como una disposición física de la materia en un dispositivo de memoria, se puede transformar después en un informe. Posteriormente se puede actuar en función de este informe. Por ejemplo, la decisión clínica puede ser poner fin al embarazo; alternatively, la decisión clínica puede ser continuar el embarazo. En algunas realizaciones, la decisión clínica puede implicar una intervención diseñada para reducir la gravedad de la presentación fenotípica de un trastorno genético, o la decisión de emprender las acciones pertinentes para prepararse para un niño con necesidades especiales.

En una realización de la presente divulgación, cualquiera de los métodos descritos en el presente documento se puede modificar para permitir que múltiples dianas procedan del mismo individuo diana, por ejemplo, múltiples extracciones de sangre de la misma mujer embarazada. Esto puede mejorar la precisión del modelo, dado que múltiples mediciones

genéticas pueden proporcionar más datos con los que se puede determinar el genotipo diana. En una realización, un conjunto de datos genéticos diana se proporcionan como datos primarios de un informe y los demás se proporcionan como datos para una doble comprobación de los datos genéticos primarios. En una realización, una pluralidad de conjuntos de datos genéticos, cada uno de ellos medidos del material genético tomado del individuo diana, se consideran en paralelo, y así ambos conjuntos de datos genéticos diana contribuyen a determinar qué secciones de los datos genéticos parentales, medidos con una alta precisión, componen el genoma fetal.

En una realización, el método se puede utilizar para realizar las pruebas de paternidad. Por ejemplo, dada la información genotípica basada en el SNP de la madre y de un hombre que puede ser o no el padre genético, y la información genotípica medida de la muestra mezclada, se puede determinar si la información genotípica del hombre representa de hecho la del padre genético real del feto en gestación. Una forma sencilla de hacerlo consiste simplemente en analizar los contextos donde la madre es AA y el posible padre es AB o BB. En estos casos, cabe esperar que la contribución del padre sea la mitad (AA|AB) o la totalidad (AA|BB) del tiempo, respectivamente. Teniendo en cuenta la ADO prevista, resulta sencillo determinar si los SNP fetales observados están correlacionados o no con los del posible padre.

Una realización de la presente divulgación podría ser como sigue: una mujer embarazada quiere saber si su feto está afectado por el síndrome de Down y/o si sufrirá fibrosis quística, y no desea tener un hijo que esté afectado por ninguna de estas condiciones. Un doctor le extrae sangre y tiñe la hemoglobina con un marcador de forma que aparezca claramente roja y tiñe el material nuclear con otro marcador para que aparezca claramente azul. Sabiendo que los glóbulos rojos maternos son típicamente anucleares, mientras que una elevada proporción de células fetales contienen un núcleo, el médico puede aislar visualmente una serie de glóbulos rojos nucleados identificando las células que presentan tanto el color rojo como el azul. El médico recoge estas células del portaobjetos con un micromanipulador y las envía a un laboratorio que amplifica y determina los genotipos de diez células individuales. Utilizando las mediciones genéticas, el método PARENTAL SUPPORT™ es capaz de determinar que seis de las células son células sanguíneas maternas y cuatro de las diez células son células fetales. Si una mujer embarazada ya ha tenido un hijo, PARENTAL SUPPORT™ se puede utilizar también para determinar que las células fetales son distintas de las células del niño nacido marcando las determinaciones de alelos fiables en las células fetales y mostrando que son distintas de las del niño nacido. Cabe señalar que el concepto de este método es similar al de la realización de la prueba de paternidad de la presente divulgación. Los datos genéticos medidos de las células fetales pueden tener muy mala calidad, al comprender muchas pérdidas de alelos, debido a la dificultad de determinar el genotipo de células únicas. El médico es capaz de utilizar el ADN fetal medido junto con las mediciones de ADN fiables de los progenitores para deducir aspectos del genoma del feto con una alta precisión utilizando PARENTAL SUPPORT™, transformando así los datos genéticos contenidos en el material genético del feto en el estado genético previsto del feto, almacenado en un ordenador. El médico es capaz de determinar tanto el estado de ploidía del feto como la presencia o ausencia de una pluralidad de genes asociados con la enfermedad de interés. Resulta que el feto es euploide y no es portador de la fibrosis quística y la madre decide continuar el embarazo.

En una realización de la presente divulgación, a una mujer embarazada le gustaría determinar si su feto está afectado por alguna anomalía del cromosoma completo. Acude al médico y le extrae una muestra de sangre, y se toman muestras a ella y su marido de su propio ADN con hisopos bucales. Un investigador del laboratorio determina los genotipos del ADN parental utilizando el protocolo MDA para amplificar el ADN parental y arrays ILLUMINA INFINIUM para medir los datos genéticos de los progenitores en un gran número de SNP. A continuación, el investigador centrifuga la sangre, recoge el plasma y aísla una muestra de ADN flotante libre utilizando la cromatografía de exclusión por tamaño. Alternativamente, el investigador utiliza uno o más anticuerpos fluorescentes, como uno que es específico para la hemoglobina fetal, para aislar un glóbulo rojo fetal nucleado. A continuación, el investigador toma el material genético fetal aislado o enriquecido y lo amplifica utilizando una biblioteca de oligonucleótidos de 70-mer convenientemente diseñada de forma que dos extremos de cada oligonucleótido se correspondan con las secuencias que flanquean cualquier lado de un alelo diana. Tras la adición de una polimerasa, ligasa y los reactivos adecuados, los oligonucleótidos se sometieron a circularización con llenado de huecos, capturando el alelo deseado. Se añadió una exonucleasa, inactivada por calor, y los productos se utilizaron directamente como plantilla para la amplificación por PCR. Los productos de la PCR se secuenciaron en un ILLUMINA GENOME ANALYZER. Las lecturas de secuencia se utilizaron como datos de entrada para el método PARENTAL SUPPORT™, que posteriormente predijo el estado de ploidía del feto.

En otra realización, una pareja, en la que la madre que está embarazada tiene una edad materna avanzada, quiere saber si el feto en gestación padece síndrome de Down, síndrome de Turner, síndrome de Prader Willi o alguna otra anomalía del cromosoma completo. El obstetra toma una muestra de sangre a la madre y al padre. La sangre se envía a un laboratorio, donde un técnico centrifuga la muestra materna para aislar el plasma y la capa leucocitaria. El ADN de la capa leucocitaria y la muestra de sangre paterna se transforman mediante amplificación y los datos genéticos codificados en el material genético amplificado se transforman también a partir de los datos genéticos molecularmente almacenados en datos genéticos electrónicamente almacenados procesando el material genético en un secuenciador de alto rendimiento para medir los genotipos parentales. La muestra de plasma es enriquecida preferentemente en un conjunto de loci utilizando un método de PCR focalizada hemi-anidada de 5000-plex. La mezcla de fragmentos de ADN se prepara creando una biblioteca de ADN adecuada para la secuenciación. A continuación, se secuencian los datos de ADN utilizando un método de secuenciación de alto rendimiento, por ejemplo, el ILLUMINA GAIIx GENOMA ANALYZER. La secuenciación transforma la información que está codificada molecularmente en el ADN en información que está

codificada electrónicamente en un software informático. Se puede utilizar una técnica basada en la informática que incluye las realizaciones divulgadas en el presente documento, como PARENTAL SUPPORT™, para determinar el estado de ploidía del feto. Esto puede implicar el cálculo, en un ordenador, de probabilidades del recuento de alelos en la pluralidad de loci polimórficos de las mediciones de ADN realizadas en la muestra preparada; la creación, en un ordenador, de una pluralidad de hipótesis de ploidía pertenecientes cada una de ellas a un posible estado de ploidía diferente del cromosoma; la creación en un ordenador, de un modelo de distribución conjunto para los recuentos de alelos previstos en la pluralidad de loci polimórficos del cromosoma para cada una de las hipótesis de ploidía; la determinación, en un ordenador, de una probabilidad relativa de cada una de las hipótesis de ploidía utilizando el modelo de distribución conjunto y los recuentos de alelos medidos en la muestra preparada; y la determinación del estado de ploidía del feto seleccionando el estado de ploidía correspondiente a la hipótesis con la probabilidad más elevada. Se determina que el feto tiene síndrome de Down. Se imprime un informe o se envía por medios electrónicos al obstetra, que transmite el diagnóstico a la mujer. La mujer, su marido y el médico se sientan a hablar sobre sus opciones. La pareja decide poner fin al embarazo basándose en el conocimiento de que el feto está afectado por una condición trisómica.

En una realización, una empresa puede decidir ofrecer una tecnología de diagnóstico diseñada para detectar la aneuploidía en un feto en gestación a partir de la extracción de una muestra de sangre materna. Su producto puede implicar que una madre visite a su obstetra, que puede extraerle una muestra de sangre. El obstetra también puede recoger una muestra genética del padre del feto. Un médico puede aislar el plasma de la sangre materna y purificar el ADN del plasma. Un médico también puede aislar la capa leucocitaria de la sangre materna y preparar el ADN de la capa leucocitaria. Un médico también puede preparar ADN de la muestra genética paterna. El médico puede utilizar técnicas de biología molecular descritas en la presente divulgación para unir etiquetas de amplificación universales al ADN obtenido de la muestra de plasma. El especialista clínico puede amplificar el ADN etiquetado universalmente. El médico puede enriquecer de forma preferente el ADN a través de diversas técnicas incluyendo la captura mediante hibridación y PCR focalizada. La PCR focalizada puede implicar un planteamiento de animado, hemi-anidado o semi-anidado, o cualquier otro planteamiento para conseguir un enriquecimiento eficiente del ADN derivado del plasma. La PCR focalizada puede ser masivamente multiplexada, por ejemplo, con 10.000 cebadores en un volumen de reacción, donde los cebadores van dirigidos a los SNP de los cromosomas 13, 18, 21, X y aquellos loci que son comunes tanto para X como para Y, y opcionalmente también para otros cromosomas. El enriquecimiento selectivo y/o la amplificación pueden implicar el etiquetado de cada molécula individual con diferentes etiquetas, códigos de barras moleculares, etiquetas para la amplificación y/o etiquetas para la secuenciación. A continuación, el médico puede secuenciar la muestra de plasma y también posiblemente el ADN materno y/o paterno preparado. Los pasos de biología molecular se pueden ejecutar parcial o totalmente en un cuadro de diagnóstico. Los datos de la secuencia se pueden introducir en un único ordenador u otro tipo de plataforma informática, como las que se pueden encontrar en "la nube". La plataforma informática puede calcular recuentos de alelos en los loci polimórficos focalizados de las mediciones realizadas por el secuenciador. La plataforma informática puede crear una pluralidad de hipótesis de ploidía correspondientes a la nulismía, monosomía, disomía, trisomía emparejada, y trisomía no emparejada para cada uno de los cromosomas 13, 18, 21, X e Y. La plataforma informática puede crear un modelo de distribución conjunto para los recuentos de alelos previstos en los loci focalizados del cromosoma para cada una de las hipótesis de ploidía para cada uno de los cinco cromosomas que se van a interrogar. La plataforma informática puede determinar una probabilidad de que cada una de las hipótesis de ploidía sea cierta utilizando el modelo de distribución conjunto y los recuentos de alelos medidas en el ADN enriquecido preferentemente derivado de la muestra de plasma. La plataforma informática puede determinar el estado de ploidía del feto, para cada uno de los cromosomas 13, 18, 21, X e Y, seleccionando el estado de ploidía correspondiente a la hipótesis relevante con la probabilidad más alta. Se puede generar un informe que comprende los estados de ploidía determinados, que puede ser enviado al obstetra por medios electrónicos, desplegado en un dispositivo de salida o se puede imprimir una copia en papel del informe que se puede entregar al obstetra. El obstetra puede informar a la paciente y, opcionalmente, al padre del feto, y ambos pueden decidir de qué opciones clínicas disponen, y cuál es la más aconsejable.

En otra realización, una mujer embarazada, en adelante denominada "la madre" puede decidir que desea saber si su feto o fetos llevan o no alguna anomalía genética u otras condiciones. Es posible que ella quiera asegurarse de que no hay anomalías importantes antes de decidir si quiere continuar con el embarazo. La madre puede acudir a su obstetra para que le tome una muestra de sangre. También puede tomar una muestra genética, por ejemplo, con un hisopo bucal, de la mejilla. Es posible que también quiera tomar una muestra genética del padre del feto, por ejemplo, con un hisopo bucal, una muestra de esperma o una muestra de sangre. El obstetra puede enviar las muestras a un médico. El médico puede enriquecer la fracción de ADN fetal flotante libre de la muestra de sangre materna. El médico puede enriquecer la fracción de células sanguíneas fetales enucleadas que contiene la muestra de sangre materna. El médico puede utilizar diversos aspectos de los métodos descritos en el presente documento para determinar los datos genéticos del feto. Esos datos genéticos pueden incluir el estado de ploidía del feto y/o la identidad de una o una serie de alelos relacionados con enfermedades en el feto. Se puede generar un informe resumiendo los resultados del diagnóstico prenatal. El informe puede ser transmitido o enviado por correo al doctor, que podrá informar a la madre del estado genético del feto. La madre podrá decidir interrumpir el embarazo basándose en el hecho de que el feto presente una o más anomalías cromosómicas o genéticas, o condiciones no deseables. También puede decidir continuar el embarazo basándose en el hecho de que el feto no presenta ninguna anomalía cromosómica o genética grave, ni condiciones genéticas de interés.

Otro ejemplo puede implicar una mujer embarazada que se ha sometido a una inseminación artificial con espermatozoides de un donante. Quiere minimizar el riesgo de que el feto tenga una enfermedad genética. Se va a extraer sangre a un flebotomista y las técnicas descritas en la presente divulgación se utilizan para aislar tres glóbulos rojos fetales nucleados, y también se toma una muestra de tejido de la madre y del padre genético. El material genético del feto y de la madre y el padre se amplifican según sea necesario y se determina el genotipo utilizando ILLUMINA INFINIUM BEADARRAY, y los métodos descritos en el presente documento limpian y determinan la fase del genotipo parental y fetal con una elevada precisión, además de determinar el estado de ploidía del feto. Se descubre que el feto es euploide y las susceptibilidades fenotípicas se predicen a partir del genotipo fetal reconstruido, y se genera un informe que es enviado al médico de la madre para que puedan decidir cuál es la mejor decisión clínica.

En una realización, el material genético bruto de la madre y del padre se transforma a través de la amplificación en una cantidad de ADN que tiene una secuencia similar pero en una cantidad mayor. A continuación, a través de un método de determinación del genotipo, los datos genotípicos codificados por los ácidos nucleicos se transforman en mediciones genéticas que pueden ser almacenadas física y/o electrónicamente en un dispositivo de memoria, como los que se han descrito anteriormente. Los algoritmos relevantes que componen el algoritmo de PARENTAL SUPPORT™, cuyas partes relevantes se exponen detalladamente en el presente documento, se trasladan a un programa informático utilizando un lenguaje de programación. A continuación, ejecutando el programa informático en el hardware informático, en lugar de ser bits y bytes físicamente codificados, dispuestos en un patrón que representa datos de medición en bruto, se transforman en un patrón que representa una determinación de alta certeza del estado de ploidía del feto. Los detalles de esta transformación se basarán en los propios datos y en el lenguaje informático y el sistema de hardware utilizados para ejecutar el método descrito en el presente documento. A continuación, los datos que están físicamente configurados para representar una determinación del estado de ploidía de alta calidad del feto se transforman en un informe que se puede enviar a un profesional sanitario. Esta transformación se puede realizar utilizando una impresora o una pantalla de ordenador. El informe puede ser una copia impresa, en papel u otro medio adecuado, o en soporte electrónico. En el caso de un informe en soporte electrónico, se puede transmitir, se puede almacenar físicamente en un dispositivo de memoria en una ubicación del ordenador a la que puede acceder el profesional sanitario; o también se puede mostrar en una pantalla para que se pueda leer. En caso de que se muestre en una pantalla, los datos se pueden transformar en un formato legible causando la transformación física de los píxeles en el dispositivo de visualización. La transformación se puede realizar activando físicamente los electrones en una pantalla fosforescente, alterando una carga eléctrica que cambia físicamente la transparencia de un conjunto específico de píxeles sobre una pantalla que puede encontrarse frente a un sustrato que emite o absorbe fotones. Esta transformación se puede realizar cambiando la orientación a nanoescala de las moléculas en un cristal líquido, por ejemplo, de una fase nemática a colestérica o esméctica, a un conjunto específico de píxeles. Esta transformación se puede conseguir mediante una corriente eléctrica que provoca que los fotones se emitan desde un conjunto específico de píxeles producidos a partir de una pluralidad de diodos emisores de luz dispuestos en un patrón significativo. Esta transformación se puede conseguir a través de cualquier otra forma utilizada para mostrar información, como la pantalla de un ordenador o algún otro dispositivo de salida o manera de transmitir información. A continuación, el profesional sanitario puede actuar en función del informe, de forma que los datos del informe se transforman en una acción. La acción puede ser continuar o interrumpir el embarazo, en cuyo caso un feto en gestación con una anomalía genética se transformará en un feto no vivo. Las transformaciones enumeradas en el presente documento se pueden agregar, de forma que, por ejemplo, se puede transformar el material genético de una madre embarazada y el padre, a través de una serie de pasos descritos en esta divulgación, en una decisión médica consistente en abordar el feto con anomalías genéticas o continuar el embarazo. Alternativamente, se puede transformar un conjunto de mediciones genotípicas en un informe que ayuda a un médico a tratar a esta paciente embarazada.

En una realización de la presente divulgación, el método descrito en el presente documento se puede utilizar para determinar el estado de ploidía de un feto incluso cuando la madre hospedadora, es decir, la mujer que está embarazada, no es la madre biológica del feto que lleva. En una realización de la presente divulgación, el método descrito en el presente documento se puede utilizar para determinar el estado de ploidía de un feto que utiliza únicamente la muestra de sangre materna, y sin necesidad de una muestra genética paterna.

Parte de las matemáticas de las realizaciones de la presente realización generan las hipótesis relativas a un número limitado de estados de aneuploidía. En algunos casos, por ejemplo, se espera que solo cero, uno o dos cromosomas procedan de cada progenitor. En algunas realizaciones de la presente divulgación, las derivaciones matemáticas se pueden ampliar para tener en cuenta otras formas de aneuploidía, como la cuatrisonía, donde tres cromosomas proceden de un progenitor, la pentasomía, la hexasomía, etc., sin variar los conceptos fundamentales de la presente divulgación. Al mismo tiempo, es posible centrarse en un número más reducido de estados de ploidía, por ejemplo, solo trisomía y disomía. Cabe señalar que las determinaciones del estado de ploidía que indican un número no entero de cromosomas pueden indicar mosaicismos en una muestra de material genético.

En algunas realizaciones, la anomalía genética es un tipo de aneuploidía, como el síndrome de Down (o trisomía 21), el síndrome de Edwards (trisomía 18), el síndrome de Patau (trisomía 13), el síndrome de Turner (45X), el síndrome de Klinefelter (un hombre con 2 cromosomas X), el síndrome de Prader-Willi y el síndrome de DiGeorge (UPD15). Por lo general, los trastornos congénitos, como los enumerados en la frase anterior, suelen ser indeseados, y el conocimiento de que un feto está afectado por una o más anomalías fenotípicas puede proporcionar la base para la decisión de poner fin a un embarazo, para tomar las precauciones necesarias para prepararse para el nacimiento de un niño con

necesidades especiales, o para adoptar algún planteamiento terapéutico orientado a aliviar la gravedad de una anomalía cromosómica.

En algunas realizaciones, los métodos descritos en el presente documento se pueden utilizar en una edad gestacional muy temprana, por ejemplo, tan solo a las cuatro semanas, tan solo a las cinco semanas, tan solo a las seis semanas, tan solo a las siete semanas, tan solo a las ocho semanas, tan solo a las nueve semanas, tan solo a las diez semanas, tan solo a las once semanas y tan solo a las doce semanas.

En algunas realizaciones, se utiliza un método divulgado en el presente documento en el contexto del diagnóstico genético previo a la implantación (PGD) para la selección del embrión durante la fertilización in vitro, donde el individuo diana es un embrión, y los datos genotípicos parentales se pueden utilizar para realizar determinaciones del estado de ploidía sobre el embrión a partir de los datos de secuenciación de una biopsia de una o dos células de un embrión de tres días o una biopsia del trofotodermo de un embrión de cinco o seis días. En el contexto del PGD, solamente se mide el ADN del niño y solamente se somete a ensayo una pequeña cantidad de células, por lo general entre una y cinco, aunque pueden ser hasta diez, veinte o cincuenta. A continuación, se determina el número total de copias de inicio de los alelos A y B (en el SNP) de forma trivial a través del genotipo del niño y del número de células. En el NPD, el número de copias de inicio es muy elevado y por tanto se espera que el ratio de alelos tras la PCR refleje de forma precisa el ratio de partida. Sin embargo, el reducido número de copias de inicio en el PGD hace que la contaminación y la escasa eficiencia de la PCR tengan un efecto no trivial sobre el ratio de alelos tras la PCR. Este efecto puede ser más importante que la profundidad de la lectura en la predicción de la varianza en el ratio de alelos medido tras la secuenciación. La distribución del ratio de alelos medido dado un genotipo conocido del niño se puede crear mediante simulación de Monte Carlo del proceso de PCR basándose en la eficiencia de la sonda de la PCR y en la probabilidad de contaminación. Dada una distribución del ratio de alelos para cada posible genotipo del niño, la probabilidad de diversas hipótesis se puede calcular tal y como se describe para el NIPD.

Cálculos de la probabilidad máxima

La mayoría de los métodos conocidos en la técnica para detectar la presencia o ausencia de fenómenos biológicos o condiciones médicas implican el uso de una prueba de rechazo de una única hipótesis, donde se mide un parámetro que está correlacionado con la condición, y si el parámetro se encuentra a un lado de un determinado umbral, la condición está presente, mientras que si se encuentra al otro lado del umbral está ausente. Una prueba de rechazo de una única hipótesis solo analiza la distribución nula al decidir entre las hipótesis nulas y alternativas. Sin tener en cuenta la distribución alternativa, no se puede estimar la probabilidad de cada hipótesis dados los datos observados y, por tanto, no se puede calcular la certeza de la determinación. Por tanto, con una prueba de rechazo de una única hipótesis, se obtiene una respuesta positiva o negativa sin una indicación de la certeza asociada al caso específico.

En algunas realizaciones, el método divulgado en el presente documento puede detectar la presencia o ausencia de fenómenos biológicos o condiciones médicas utilizando un método de probabilidad máxima. Esto supone una mejora sustancial con respecto a un método que utiliza una técnica de rechazo de una única hipótesis, dado que el umbral para determinar la ausencia o presencia de la condición se puede ajustar según corresponda a cada caso. Esto es particularmente relevante para las técnicas de diagnóstico que pretenden determinar la presencia o ausencia de aneuploidía en un feto en gestación con los datos genéticos disponibles de la mezcla del ADN fetal y materno presente en el ADN flotante libre que se encuentra en el plasma materno. Esto se debe a que cuando cambia la fracción de ADN fetal de la fracción obtenida del plasma, varía el umbral óptimo para determinar la aneuploidía frente a la euploidía. Cuando cae la fracción fetal, la distribución de los datos asociados con una aneuploidía es cada vez más similar a la distribución de los datos asociados a la euploidía.

El método de la estimación de la probabilidad máxima utiliza las distribuciones asociadas con cada hipótesis para estimar la probabilidad de los datos condicionados de cada hipótesis. Estas probabilidades condicionales se pueden convertir en la determinación de una hipótesis y de la certeza. Del mismo modo, un método de estimación máxima a posteriori utiliza las mismas probabilidades condicionales que la estimación de la probabilidad máxima, pero también incorpora datos previos de la población a la hora de elegir la mejor hipótesis y determinar la certeza.

Por tanto, el uso de una técnica de estimación de la probabilidad máxima (MLE), o la técnica estrechamente relacionada de la máxima a posteriori (MAP) ofrece dos ventajas: la primera que aumenta la probabilidad de una determinación correcta y la segunda que también permite un cálculo de la certeza para cada determinación. En una realización, la selección del estado de ploidía correspondiente a la hipótesis con la probabilidad más elevada se realiza utilizando estimaciones de probabilidad máxima o estimaciones de la máxima a posteriori. En una realización, se divulga un método para determinar el estado de ploidía de un feto en gestación que implica la utilización de cualquier método actualmente conocido en la técnica que utiliza una técnica de rechazo de una única hipótesis y la reformulación de dicho método para que utilice una técnica de MLE o MAP. Algunos ejemplos de los métodos que se pueden mejorar de forma significativa aplicando estas técnicas se pueden encontrar en la Patente USA 8.008.018, la Patente USA 7.888.017 o la Patente USA 7.332.277.

En una realización, se describe un método para determinar la presencia o ausencia de aneuploidía fetal en una muestra de plasma materno que comprende ADN genómico fetal y materno, donde el método consiste en lo siguiente: obtener una muestra de plasma materno; medir los fragmentos de ADN que se encuentran en la muestra de plasma con un secuenciador de alto rendimiento; correlacionar las secuencias con el cromosoma y determinar el número de lecturas de secuencia que corresponden a cada cromosoma; calcular la fracción de ADN fetal en la muestra de plasma; calcular

una distribución prevista de la cantidad de un cromosoma diana que cabría esperar que estuviese presente si el segundo cromosoma diana fuese euploide y uno o una pluralidad de distribuciones previstas que cabría esperar si ese cromosoma fuese aneuploide, utilizando la fracción fetal y el número de lecturas de secuencia que corresponden a una o una pluralidad de cromosomas de referencia que se espera que sean euploides; y utilizando una MLE o MAP para determinar cuáles de las distribuciones es más probable que sean correctas, indicando así la presencia o ausencia de aneuploidía fetal. En una realización, la medición de ADN del plasma puede implicar la realización de una secuenciación por fuerza bruta masivamente paralela. En una realización, la medición del ADN de la muestra de plasma puede implicar la secuenciación de ADH que se ha enriquecido preferentemente, por ejemplo, mediante amplificación focalizada, en una pluralidad de loci polimórficos o no polimórficos. La pluralidad de loci puede estar diseñada para focalizar uno o un pequeño número de cromosomas que se sospecha que son aneuploides y uno o un pequeño número de cromosomas de referencia. El propósito del enriquecimiento preferente consiste en aumentar el número de lecturas de secuencia que son informativas para la determinación de la ploidía.

Métodos informáticos para la determinación de la ploidía

En el presente documento se describe un método para determinar el estado de ploidía de un feto en función de los datos de secuencia. En algunas realizaciones, los datos de esta secuencia se pueden medir en un secuenciador de alto rendimiento. En algunas realizaciones, los datos de la secuencia se pueden medir en ADN procedentes del ADN flotante libre aislado de la sangre materna, donde el ADN flotante libre comprende ADN de origen materno y ADN de origen fetal/placentario. Esta sección describirá una realización de la presente divulgación en la que el estado de ploidía del feto se determina asumiendo que la fracción de ADN fetal de la mezcla que ha sido analizada se desconoce y se calculará a partir de los datos. También se describirá una realización en la que la fracción de ADN fetal ("fracción fetal") o el porcentaje de ADN fetal de la mezcla se puede medir mediante otro método y se asume que es conocida para determinar el estado de ploidía del feto. En algunas realizaciones, la fracción fetal se puede calcular utilizando solo mediciones del genotipo realizadas con la muestra de la sangre materna, que es una mezcla de ADN fetal y materno. En algunas realizaciones, la fracción se puede calcular utilizando también el genotipo medido o conocido de otro modo de la madre y/o el genotipo medido o conocido de otro modo del padre. En otra realización, el estado de ploidía del feto se puede determinar basándose exclusivamente en la fracción calculada de ADN fetal para el cromosoma en cuestión en comparación con la fracción calculada de ADN fetal para el cromosoma de referencia que se considera disómico.

En la realización preferible, se supone que, para un cromosoma concreto, observamos y analizamos N SNP, para los que tenemos:

Un conjunto de NR mediciones de la secuencia de ADN flotante libre $S=(s_1, \dots, s_N)$. Dado que este método utiliza las mediciones de SNP, todos los datos de la secuencia que corresponden a loci no polimórficos se pueden ignorar. En una versión simplificada, donde tenemos recuentos (A,B) en cada SNP, donde A y B corresponden a los dos alelos presentes en un determinado locus, S se puede escribir como $S=((a_1, b_1), \dots, (a_N, b_N))$, donde a_i es el recuento de A en el

$$\sum_{i=1}^N (a_i + b_i) = NR$$

SNP i , b_i es el recuento de B en el SNP i , y

Los datos parentales se componen de

- Genotipos de un microarray de SNP u otra plataforma de determinación del genotipo basada en la intensidad: madre $M=(m_1, \dots, m_N)$ padre $F=(f_1, \dots, f_N)$, donde $m_i, f_i \in \{AA, AB, BB\}$.
- Y/O mediciones de los datos de la secuencia: NRM mediciones de la madre $SM=(sm_1, \dots, sm_{nrm})$, NRF mediciones del padre $SF=(sf_1, \dots, sf_{nrf})$. De forma similar a la simplificación anterior, si tenemos recuentos (A,B) de cada SNP $SM=((am_1, bm_1), \dots, (am_N, bm_N))$, $SF=((af_1, bf_1), \dots, (af_N, bf_N))$

Colectivamente, los datos de la madre, del padre y del niño se denotan como $D = (M, F, SM, SF, S)$. Cabe señalar que los datos parentales son deseables y aumentan la precisión del algoritmo, aunque NO son necesarios, en especial los datos del padre. Esto significa que incluso en ausencia de datos de la madre y/o el padre, es posible obtener resultados muy precisos del número de copias.

Se puede obtener la mejor estimación del número de copias (H^*) maximizando la probabilidad del logaritmo de datos $LIK(D|H)$ con respecto a todas las hipótesis (H) consideradas. En particular, se puede determinar la probabilidad relativa de cada una de las hipótesis de ploidía utilizando el modelo de distribución conjunto y los recuentos de alelos medidos en la muestra preparada, y utilizando esas probabilidades relativas determinar la hipótesis que es más probable que sea correcta del modo siguiente:

$$H^* = \underset{H}{\operatorname{argmax}} LIK(D|H)$$

De forma similar, la probabilidad de las hipótesis a posteriori en función de los datos se pueden escribir como sigue:

$$H^* = \underset{H}{\operatorname{argmax}} LIK(D|H) * \text{priorprob}(H)$$

Donde $\text{priorprob}(H)$ es la probabilidad previa asignada a cada hipótesis H , basada en el diseño del modelo y en el conocimiento previo.

También se pueden utilizar datos previos para hallar la estimación de la máxima a posteriori:

$$H_{MA} = \underset{H}{\operatorname{argmax}} \operatorname{LIK}(D|H)$$

En una realización, las hipótesis del número de copias que se pueden considerar son:

- Monosomía:

- materna H10 (una copia de la madre)

- paterna H01 (una copia del padre)

5 • Disomía: H11 (una copia de la madre y una copia del padre)

- Trisomía simple (sin considerar los cruces):

- Materna: H21_matched (dos copias idénticas de la madre, una copia del padre), H21_unmatched (AMBAS copias de la madre, una copia del padre)

10 -Paterna: H12_matched (una copia de la madre, dos copias idénticas del padre), H12_unmatched (una copia de la madre, ambas copias del padre)

- Trisomía compuesta, permitiendo cruces (utilizando un modelo de distribución conjunto):

- materna H21 (dos copias de la madre, una del padre),

- paterna H12 (una copia de la madre, dos copias del padre)

15 En otras realizaciones, se pueden considerar otros estados de ploidía, como la nulisomía (H00), disomía uniparental (H20 y H02), y tetrasomía (H04, H13, H22, H31 y H40).

Si no hay cruces, cada trisomía, con independencia de que el origen haya sido la mitosis, meiosis I o meiosis II, sería una de las trisomías emparejadas o no emparejadas. Debido a los cruces, la verdadera trisomía suele ser una combinación de ambas. En primer lugar, se describe un método para obtener probabilidades de hipótesis para las hipótesis simples. A continuación, se describe un método para obtener probabilidades de hipótesis para las hipótesis compuestas, que combinan la probabilidad individual del SNP con los cruces.

LIK(D\H) para una hipótesis simple

En una realización, se puede determinar LIK(D|H) para las hipótesis simples del modo siguiente: Para las hipótesis simples H, LIK(H), la probabilidad logarítmica de la hipótesis H en un cromosoma completo, se puede calcular como la suma de las probabilidades logarítmicas de los SNP individuales, asumiendo la fracción cf conocida o derivada del niño. En una realización se puede obtener cf a partir de los datos.

$$\operatorname{LIK}(D|H) = \sum_i \operatorname{LIK}(D|H, cf, i)$$

Esta hipótesis no suma ningún enlace entre los SNP y, por tanto, no utiliza un modelo de distribución conjunto.

En algunas realizaciones, la probabilidad logarítmica se puede determinar para cada SNP. En un SNP i concreto, asumiendo la hipótesis de ploidía fetal H y un porcentaje de ADN fetal cf, la probabilidad logarítmica de los datos observados D se define como:

$$\operatorname{LIK}(D|H, i) = \log P(D|H, cf, i) = \log \left(\sum_{m,f,c} P(D|m, f, c, H, cf, i) P(c|m, f, H) P(m|i) P(f|i) \right)$$

30 donde m son posibles genotipos verdaderos de la madre, f son posibles genotipos verdaderos del padre, donde m, f e {AA, AB, BB}, y donde c son posibles genotipos del niño dada la hipótesis H. En concreto, para la monosomía c e {A, B}, para la disomía c e {AA, AB, BB}, para la trisomía c e {AAA, AAB, ABB, BBB},

Frecuencia previa del genotipo: p(m|i) es la probabilidad previa general del genotipo de la madre m en el SNP i, basándose en la frecuencia de la población conocida en SNP I, denotada como pAi. En concreto

$$35 \quad p(AA|pA_i) = (pA_i)^2, p(AB|pA_i) = 2(pA_i) * (1 - pA_i), p(BB|pA_i) = (1 - pA_i)^2$$

La probabilidad del genotipo del padre, p(f|i), se puede determinar de forma análoga.

Probabilidad verdadera del niño: p(c|m, f, H) es la probabilidad de obtener el genotipo verdadero del niño = c, dados los progenitores m, f, y asumiendo la hipótesis H, que se puede calcular fácilmente. Por ejemplo, para H11, H21 emparejado y H21 no emparejado, p(c|m, f, H) se proporciona a continuación.

40

p(c m,f,H)													
		H11			H21 emparejado				H21 no emparejado				
m	f	AA	AB	BB	AAA	AAB	ABB	BBB	AAA	AAB	ABB	BBB	
AA	AA	1	0	0	1	0	0	0	1	0	0	0	
AB	AA	0.5	0.5	0	0.5	0	0.5	0	0	1	0	0	
BB	AA	0	1	0	0	0	1	0	0	0	1	0	
AA	AB	0.5	0.5	0	0.5	0.5	0	0	0.5	0.5	0	0	
AB	AB	0.25	0.5	0.25	0.25	0.25	0.25	0.25	0	0.5	0.5	0	
BB	AB	0	0.5	0.5	0	0	0.5	0.5	0	0	0.5	0.5	
AA	BB	0	1	0	0	1	0	0	0	1	0	0	
AB	BB	0	0.5	0.5	0	0.5	0	0.5	0	0	1	0	
BB	BB	0	0	1	0	0	0	1	0	0	0	1	

Probabilidad de los datos: $P(D|m, f, c, H, i, cf)$ es la probabilidad de los datos dados D en el SNP i, dado el genotipo verdadero de la madre m, el genotipo verdadero del padre f, el genotipo verdadero del niño c, la hipótesis H y la fracción del niño cf. Se pueden dividir en los datos de probabilidad de la madre, del padre y del niño como sigue:

$$P(D|m, f, c, H, cf, i) = P(SM|m, i)P(M|m, i)P(SF|f, i)P(F|f, i)P(S|m, c, H, cf, i)$$

- 5 Probabilidad de los datos del array del SNP de la madre: La probabilidad de los datos m del genotipo del array del SNP de la madre en SNP ⁱ en comparación con el genotipo verdadero m, asumiendo que los genotipos del array de SNP son correctos, es simplemente

$$P(M|m, i) = \begin{cases} 1 & m_i = m \\ 0 & m_i \neq m \end{cases}$$

Probabilidad de los datos de la secuencia de la madre: la probabilidad de los datos de la secuencia de la madre en SNP i, en el caso de los recuentos $S_i=(a_{m_i}, b_{m_i})$, sin ningún ruido o sesgo adicional implicado, es la probabilidad binomial definida como $P(SM|m, i) = P_{X|m}(a_{m_i})$ donde $X|m \sim \text{Binom}(p_m(A), a_{m_i} + b_{m_i})$ con $p_m(A)$ se define como

m	AA	AB	BB	A	B	Sin determinar
P(A)	1	0,5	0	1	0	0,5

Probabilidad de los datos del padre: una ecuación similar se aplica a la probabilidad de los datos del padre. Cabe señalar que se puede determinar el genotipo del hijo sin los datos parentales, especialmente los datos del padre. Por ejemplo, si no se dispone de los datos del genotipo del padre F, se puede utilizar simplemente $P(F|f, i) = 1$. Si no se dispone de los datos de la secuencia del padre SF, se puede utilizar simplemente $P(SF|f, i) = 1$.

En algunas realizaciones, el método implica la elaboración de un modelo de distribución conjunto para los recuentos de alelos previstos en una pluralidad de loci polimórficos del cromosoma para cada una de las hipótesis de ploidía; en el presente documento se describe un método para ello. Probabilidad de los datos del ADN fetal libre: $P(S|m, c, H, cf, i)$ es la probabilidad de los datos de la secuencia de ADN fetal libre en SNP i, dado el genotipo verdadero de la madre m, el genotipo verdadero del hijo c, la hipótesis del número de copias del hijo H, y asumiendo una fracción del hijo cf. De hecho, se trata de la probabilidad de los datos de la secuencia S en SNP i, dada la probabilidad verdadera de contenido A en SNP i $\mu(m, c, cf, H)$

$$P(S|m, c, H, cf, i) = P(S|\mu(m, c, cf, H), i)$$

Para los recuentos, donde $S_i=(a_i, b_i)$, sin ningún ruido ni sesgo adicional de los datos,

$$P(S|\mu(m, c, cf, H), i) = P_X(a_i)$$

donde $X \sim \text{Binom}(p(A), a_i + b_i)$ con $p(A) = \mu(m, c, cf, H)$. En un caso más complejo donde la alineación exacta y los recuentos (A,B) por SNP se desconocen, $P(S|\mu(m, c, cf, H), i)$ es una combinación de binomios integrados.

Probabilidad verdadera de contenido A: $\mu(m, c, cf, H)$, la probabilidad verdadera de contenido A en SNP i de esta mezcla madre/hijo, asumiendo el genotipo verdadero de la madre = m, el genotipo verdadero del niño = c, y la fracción del niño total = cf, se define como

$$\mu(m, c, cf, H) = \frac{\#A(m) * (1 - cf) + \#A(c) * cf}{n_m * (1 - cf) + n_c * cf}$$

donde $\#A(g)$ = número de A en el genotipo g, $n_m = 2$ es la suma de la madre y n_c es la ploidía del hijo bajo la hipótesis H (1 para monosomía, 2 para disomía, 3 para trisomía).

Utilización de un modelo de distribución conjunto: $LIK(D|H)$ para una hipótesis compuesta

En algunas realizaciones, el método implica la elaboración de un modelo de distribución conjunto para los recuentos de alelos previstos en una pluralidad de loci polimórficos del cromosoma para cada una de las hipótesis de ploidía; en el presente documento se describe un método para ello. En muchos casos, habitualmente la trisomía no está puramente emparejada o no emparejada, debido a los cruces, por lo que en esta sección se obtienen los resultados de las hipótesis compuestas H21 (trisomía materna) y H12 (trisomía paterna), que combinan la trisomía emparejada y no emparejada, teniendo en cuenta posibles cruces.

En el caso de la trisomía, si no hubiese cruces, la trisomía sería simplemente una trisomía emparejada o no emparejada. La trisomía emparejada es cuando el niño hereda dos copias del segmento del cromosoma idéntico de uno de los progenitores. La trisomía no emparejada es cuando el niño hereda una copia de cada segmento del cromosoma homólogo del progenitor. Debido a los cruces, algunos segmentos de un cromosoma pueden tener trisomía emparejada y otras partes presentar una trisomía no emparejada.

En esta sección se describe cómo elaborar un modelo de distribución conjunto para los índices de heterocigosidad de un conjunto de alelos; es decir, para los recuentos de alelos previstos en una serie de loci para una o más hipótesis.

Suponiendo que SNP i , $LIK(D|H_m, i)$ corresponde a la hipótesis H_m emparejada, y $LIK(D|H_u, i)$ corresponde a la hipótesis H_u no emparejada, y $pc(i)$ = probabilidad de cruce entre SNPs $i-1$ and i . A continuación, se puede calcular la probabilidad completa como:

$$LIK(D|H) = \sum_E LIK(D|E, 1:N)$$

donde $LIK(D|E, 1:N)$ es la probabilidad de acabar en la hipótesis E , para SNP $1:N$. E = hipótesis del último SNP, E e (H_m, H_u) . Repetidamente, se puede calcular:

$$LIK(D|E, 1:i) = LIK(D|E, i) + \log(\exp(LIK(D|E, 1:i-1)) * (1 - pc(i)) + \exp(LIK(D|\sim E, 1:i-1)) * pc(i))$$

donde $\sim E$ es la hipótesis distinta de E (no E), donde las hipótesis consideradas son H_m y H_u . En particular, se puede calcular la probabilidad de $1:i$ SNPs, basándose en la probabilidad de 1 a $(i-1)$ SNP con la misma hipótesis y ningún cruce, o la hipótesis opuesta y un cruce, multiplicada por la probabilidad del SNP i

Para SNP $1 = i1$ (aquí falta un símbolo o fórmula que aparece en el pdf original)
Para SNP $2, i=2$, y así sucesivamente para $i=3:N$.

$$LIK(D|E, 1:2) = LIK(D|E, 2) + \log(\exp(LIK(D|E, 1)) * (1 - pc(2)) + \exp(LIK(D|\sim E, 1)) * pc(2)),$$

En algunas realizaciones, se puede determinar la fracción del niño. La fracción del niño se puede referir a la proporción de secuencias en una mezcla de ADN procedente del niño. En el contexto del diagnóstico prenatal no invasivo, la fracción del niño se puede referir a la proporción de secuencias del plasma materno que proceden del feto o la proporción de la placenta con genotipo fetal. Se puede referir a la fracción del niño en una muestra de ADN que se ha preparado con plasma materno y que puede ser enriquecida en ADN fetal. Un propósito de la determinación de la fracción del niño en la muestra de ADN es el uso de un algoritmo que puede realizar determinaciones del estado de ploidía del feto. Por tanto, la fracción del niño se podría referir a cualquier muestra de ADN analizada por secuenciación con el fin de realizar un diagnóstico prenatal no invasivo.

Algunos de los algoritmos presentados en esta divulgación que forman parte de un diagnóstico de aneuploidía prenatal no invasivo asumen una fracción del niño conocida, que no siempre es el caso. En una realización, se puede hallar la fracción del niño más probable maximizando la probabilidad para la disomía en cromosomas seleccionados, con o sin la presencia de los datos parentales.

En concreto, suponiendo que $LIK(D|H11, cf, chr)$ = probabilidad logarítmica descrita anteriormente, para las hipótesis de disomía, y para la fracción del niño cf en el cromosoma chr . Para cromosomas seleccionados en $Cset$ (normalmente $1:16$), que se supone que son euploides, la probabilidad completa es:

$$LIK(cf) = \sum_{chr \in Cset} LIK(D|H11, cf, chr)$$

La fracción del niño más probable se obtiene como $cf^* = \arg\max_{cf} LIK(cf)$.

Se puede utilizar cualquier conjunto de cromosomas. También se puede obtener la fracción del niño sin asumir la euploidía de los cromosomas de referencia. Utilizando este método, se puede determinar la fracción del niño para cualquiera de las situaciones siguientes: (1) se dispone de datos del array de los progenitores y datos de la secuenciación por fuerza bruta del plasma materno; (2) se dispone de datos del array de los progenitores y datos de la secuenciación focalizada del plasma materno; (3) se dispone de datos de la secuenciación focalizada de ambos progenitores y del plasma materno; (4) se dispone de datos de la secuenciación focalizada de la madre y de la fracción de plasma materno; (5) se dispone de datos de la secuenciación focalizada de la fracción del plasma materno; (6) otras combinaciones de mediciones de la fracción parental y del niño.

En algunas realizaciones, el método informático puede incorporar pérdidas de datos; esto puede proporcionar determinaciones de la ploidía con una mayor precisión. En otro apartado de esta divulgación se ha asumido que la probabilidad de obtener una A es una función directa del genotipo verdadero de la madre, el genotipo verdadero del niño, la fracción del niño en la mezcla y el número de copias del niño. También es posible que se produzca una pérdida de alelos de la madre o del niño, por ejemplo, en lugar de medir el verdadero AB del niño en la mezcla, se puede dar el caso de que solo se midan las secuencias correspondientes al alelo A. Se puede denotar la tasa de pérdida parental para los datos genómicos de Illumina d_{pg} , la tasa de pérdida parental para los datos de la secuencia d_{ps} y la tasa de pérdida del niño para los datos de la secuencia d_{cs} . En algunas realizaciones, la tasa de pérdida de la madre se puede asumir que es cero y las tasas de pérdida del niño son relativamente bajas; en este caso, los resultados no se ven gravemente afectados por las pérdidas. En algunas realizaciones, la posibilidad de pérdidas de alelos puede ser suficientemente grande como para que produzca un efecto significativo sobre la determinación del estado de ploidía previsto. En este caso, las pérdidas de alelos se han incorporado al algoritmo como sigue:

Pérdidas de los datos del array del SNP parental: Para los datos genómicos de la madre M, se supone que el genotipo tras la pérdida es m_d , por tanto

$$P(M|m, i) = \sum_{m_d} P(M|m_d, i)P(m_d|m)$$

donde $P(M|m_d, i) = \begin{cases} 1 & m_i = m_d \\ 0 & m_i \neq m_d \end{cases}$ como antes, y $P(m_d|m)$ es la probabilidad del genotipo m_d tras la posible pérdida dado el verdadero genotipo m, definido más abajo, para la tasa de pérdida d

md						
m	AA	AB	BB	A	B	sin determinar
AA	$(1-d)\lambda^2$	0	0	$2d(1-d)$	0	$d\lambda^2$
AB	0	$(1-d)\lambda^2$	0	$d(1-d)$	$d(1-d)$	$d\lambda^2$
BB	0	0	$(1-d)\lambda^2$	0	$2s(1-d)$	$d\lambda^2$

Una ecuación similar se aplica a los datos del array del SNP del padre.

Pérdidas de datos de la secuencia parental: Para los datos de la secuencia de la madre SM

$$P(SM|m, i) = \sum_{m_d} P_{X|m_d}(am_i)P(m_d|m)$$

donde $P(m_d|m)$ se define como en la sección anterior y $P_{X|m_d}(am_i)$ la probabilidad de una distribución binomial se define como anteriormente en la sección de la probabilidad de los datos parentales. Una ecuación similar se aplica a los datos de la secuencia paterna.

Pérdida de datos de la secuencia de ADN flotante libre:

$$P(S|m, c, H, cf, i) = \sum_{m_d, c_d} P(S|\mu(m_d, c_d, cf, H), i)P(m_d|m)P(c_d|c)$$

donde $P(S|\mu(m_d, c_d, cf, H), i)$ es como se define en la sección sobre probabilidad de los datos flotantes libres.

En una realización, $p(m_d|m)$ es la probabilidad del genotipo observado de la madre, dado el genotipo verdadero de la madre m, asumiendo una tasa de pérdida d_{ps} , y $p(cd|c)$ es la probabilidad del genotipo observado del niño cd , dado el genotipo verdadero del niño c, asumiendo una tasa de pérdida d_{cs} . Si nA_T = número de alelos A en el genotipo verdadero c, nA_D = número de alelos A en el genotipo observado ^cd , donde $nA_T \geq nA_D$, y de forma similar nB_T = número de alelos B en el genotipo verdadero c, nB_D = número de alelos B en el genotipo observado ^cd , donde $nB_T \geq nB_D$ y d = tasa de pérdida, entonces

$$p(c_d|c) = \binom{nA_T}{nA_D} * d^{nA_T-nA_D} * (1-d)^{nA_D} * \binom{nB_T}{nB_D} * d^{nB_T-nB_D} * (1-d)^{nB_D}$$

En una realización, el método informático puede incorporar un sesgo aleatorio y constante. En una situación ideal no existe ningún sesgo de muestreo constante por SNP ni ruido aleatorio (además de la variación de la distribución binomial) en el número de recuentos de secuencias. En concreto, con respecto a SNP i, para el genotipo de la madre m, el genotipo verdadero del niño c y la fracción del niño cf, y X = el número de A en el conjunto de lecturas de (A+B) en SNP i, X actúa como un X~Binómico (p, A+B), donde $p = \mu, c, cf, H$ = probabilidad verdadera de contenido de A.

En una realización, el método informático puede incorporar un sesgo aleatorio. Como suele ocurrir, se supone que existe un sesgo en las mediciones, de forma que la probabilidad de obtener una A en este SNP es igual a q, que es

ligeramente diferente de p que se ha definido anteriormente. El punto hasta el que p es diferente de q depende de la precisión del proceso de medición y de varios otros factores y se puede cuantificar a través de las desviaciones estándar de q con respecto a p . En una realización, se puede elaborar un modelo de q que tiene una distribución beta, con los parámetros α, β dependientes de la media de esa distribución centrada en p , y cierta desviación estándar especificada s . En concreto, esto da $X|q \sim \text{Bin}(q, D_i)$, donde $q \sim \text{Beta}(\alpha, \beta)$. Si dejamos $E(q) = p, V(q) = s^2$ y los

parámetros se pueden derivar como $\alpha = pN, \beta = (1 - p)N$ donde $N = \frac{p(1-p)}{s^2} - 1$

Esta es la definición de una distribución beta-binomial, donde se realiza un muestreo de una distribución binomial con un parámetro variable q , donde q sigue una distribución beta como una media p . Por tanto, en una configuración sin sesgo, con respecto a SNP i , la probabilidad de los datos de la secuencia paterna (SM) asumiendo el genotipo verdadero de la madre (m), dado el recuento de A de la secuencia de la madre en SNP i ($a_{m,i}$) y el recuento de B de la secuencia de la madre en SNP i ($b_{m,i}$) se puede calcular como:

$$P(\text{SM}|m,i) = P_{X|m}(a_{m,i}) \text{ donde } X|m \sim \text{Binom}(p_m(A), a_{m,i} + b_{m,i})$$

Ahora, incluyendo un sesgo aleatorio con una desviación estándar s , esto se convierte en: $X|m \sim \text{BetaBinom}(p_m(A), a_{m,i} + b_{m,i}, s)$

En caso de ausencia de sesgo, la probabilidad de los datos de la secuencia de ADN de plasma materno (S), asumiendo el genotipo verdadero de la madre (m), el genotipo verdadero del niño (c) y la fracción del niño (cf), asumiendo la hipótesis del niño H , dado un recuento de A de la secuencia de ADN flotante libre en SNP i (a_i) y un recuento de B en la secuencia de ADN flotante libre en SNP i (b_i), se puede calcular del modo siguiente:

$$P(S|m, c, cf, H, i) = P_X(a_i)$$

donde $X \sim \text{Binom}(p(A), a_i + b_i)$ with $p(A) = \mu(m, c, cf, H)$.

En una realización, incluyendo un sesgo aleatorio con una desviación estándar s , esto se convierte en $X \sim \text{BetaBinom}(p(A), a_i + b_i, s)$, donde la cantidad de variación adicional viene determinada por el parámetro de desviación s , o equivalentemente N . Cuanto menor es el valor de s (o mayor el valor de N) más se aproximará esta distribución a la distribución binomial ordinaria. Es posible estimar la cantidad de sesgo, es decir, estimar $\hat{N}$ anterior, a partir de contextos no ambiguos $AA|AA, BB|BB, AA|BB, BB|AA$ y utilizar el valor $\hat{N}$ estimado en la probabilidad anterior. Dependiendo del comportamiento de los datos, N puede ser una constante con independencia de la profundidad de lectura $a_i + b_i$, o una función de $a_i + b_i$, haciendo el sesgo menor para las profundidades de lectura mayores.

En una realización, el método informático puede incorporar un sesgo constante por SNP. Debido a los artefactos del proceso de secuenciación, algunos SNP pueden tener unos recuentos más bajos o más altos con independencia de la cantidad verdadera de contenido de A. Supongamos que SNP i añade de forma constante un sesgo de un porcentaje w_i al número de recuentos A. En algunas realizaciones, este sesgo se puede estimar a partir del conjunto de datos de formación derivados en las mismas condiciones, y añadirse a la estimación de los datos de la secuencia de los progenitores como sigue:

$$P(\text{SM}|m,i) = P_{X|m}(a_{m,i}) \text{ donde } X|m \sim \text{BetaBinom}(p_m(A) + w_i, a_{m,i} + b_{m,i}, s)$$

y con la estimación de la probabilidad de los datos de la secuencia de ADN flotante libre como:

$$P(S|m, c, cf, H, i) = P_N(a_i) \text{ donde } X \sim \text{BetaBinom}(p(A) + w_i, a_i + b_i, s),$$

En algunas realizaciones, el método puede estar escrito para tener en cuenta específicamente el ruido adicional, la calidad de la muestra diferencial, la calidad del SNP diferencial y el sesgo de muestreo aleatorio. Un ejemplo de esto se proporciona en el presente documento. Este método ha demostrado ser particularmente útil en el contexto de los datos generados utilizando el protocolo de mini-PCR masivamente multiplexada, y se utilizó en los Experimentos 7 a 13. El método implica varios pasos que introducen cada uno diferentes tipos de ruido y/o sesgo con respecto al modelo final:

(1) Se asume una primera muestra que contiene una mezcla de ADN materno y fetal que comprende una cantidad original de ADN de tamaño N_0 moléculas, normalmente en el rango 1.000-40.000, donde $p = \% \text{refs verdaderas}$

(2) En la amplificación utilizando adaptadores de unión universales, se asume que se someten a muestreo N_1 moléculas, normalmente $N_1 \sim N_0/2$ moléculas y el sesgo de muestreo aleatorio se introduce debido al muestreo. La muestra amplificada puede contener un número de moléculas N_2 donde $N_2 \gg N_1$. Suponiendo que X_1 representa la cantidad de loci de referencia (por SNP) de N_1 moléculas sometidas a muestreo, con una variación en $p_1 = X_1/N_1$ que introduce un sesgo de muestreo aleatorio durante el resto del protocolo. Este sesgo de muestreo se incluye en el modelo utilizando una distribución Beta-binomial (BB) en lugar de utilizar un modelo de distribución binomial simple. El parámetro N de la distribución Beta-binomial se puede estimar posteriormente para cada muestra a partir de los datos de formación, tras haber realizado el correspondiente ajuste por el sesgo de pérdida y amplificación, en los SNP con $0 < p < 1$. La pérdida es la tendencia de un SNP a ser leído incorrectamente.

(3) El paso de amplificación amplificará cualquier sesgo alélico y, por tanto, el sesgo de amplificación introducido debido a una posible amplificación heterogénea. Asumiendo que un alelo de un locus sea amplificado f veces, otro alelo de ese locus sea amplificado g veces, donde $f = ge^b$, donde $b=0$ indica ausencia de sesgo. El parámetro del sesgo, b , está centrado en 0 e indica en qué grado mayor o mayor se ha amplificado el alelo A en comparación con el alelo B en un

SNP concreto. El parámetro b puede diferir de un SNP a otro SNP. El parámetro del sesgo b se puede estimar por cada SNP, por ejemplo, a partir de los datos de formación.

(4) El paso de secuenciación implica la secuenciación de una muestra de moléculas amplificadas. En este paso puede haber una pérdida, donde la pérdida es la situación en la que un SNP es leído de forma incorrecta. La pérdida puede producirse por cualquier número de problemas y puede provocar que un SNP no sea leído como el alelo A correcto, sino como otro alelo B que se encuentra en ese locus o como un alelo C o D que no se encuentra típicamente en ese locus. Suponiendo que la secuenciación mide los datos de la secuencia de una serie de moléculas de ADN de una muestra amplificada de tamaño N_3 , donde $N_3 < N_2$. En algunas realizaciones, N_3 se puede encontrar en el rango de 20.000 a 100.000; 100.000 a 500.000; 500.000 a 4.000.000; 4.000.000 a 20.000.000; o 20.000.000 a 100.000.000. Cada molécula sometida a un muestreo tiene una probabilidad p_g de ser leída correctamente, en cuyo caso aparecerá correctamente como alelo A. La muestra será leída incorrectamente como un alelo no relacionado con la molécula original con probabilidad $1-p_g$, y aparecerá como un alelo A con probabilidad p_r , un alelo B con probabilidad p_m o un alelo C o alelo D con probabilidad p_o , donde $p_r+p_m+p_o=1$. Los parámetros p_g , p_r , p_m , p_o se estiman para cada SNP a partir de los datos de formación.

Diferentes protocolos pueden implicar pasos similares con variaciones en los pasos de biología molecular que dan como resultado diferentes cantidades de muestreo aleatorias, diferentes niveles de amplificación y diferente sesgo de pérdidas. El siguiente modelo también se puede aplicar perfectamente a cada uno de estos casos. El modelo para la cantidad de ADN sometido a muestreo, por cada SNP, viene dado por:

$$X \sim \text{BetaBinomial}(L(F(p,b), p_r, p_g), N \cdot H(p,b))$$

donde p = la cantidad verdadera de ADN de referencia, b = sesgo por SNP, y tal y como se ha descrito anteriormente, p_g es la probabilidad de una lectura correcta, p_r es la probabilidad de que una lectura sea leída de forma incorrecta pero accidentalmente parezca el alelo correcto, en caso de una mala lectura, tal y como se ha descrito anteriormente, y:

$$F(p,b) = pe^b / (pe^b + (1-p)), H(p,b) = (e^b p + (1-p))^2 / e^b, L(p, p_r, p_g) = p \cdot p_g + p_r \cdot (1-p_g).$$

En algunas realizaciones, el método utiliza una distribución Beta-binomial en lugar de una distribución binomial simple; de este modo se presta atención al sesgo de muestreo aleatorio. El parámetro N de la distribución Beta-binomial es estimado para cada muestra o cuando resulta necesario. Utilizando una corrección del sesgo $F(p,b)$, $H(p,b)$, en lugar de tan solo p , se presta atención al sesgo de amplificación. El parámetro b del sesgo es estimado para cada SNP a partir de los datos de formación con antelación.

En algunas realizaciones, el método utiliza la corrección de pérdidas $L(p, p_r, p_g)$, en lugar de tan solo p ; de este modo se presta atención al sesgo de pérdidas, es decir, la variación del SNP y de la calidad de la muestra. En algunas realizaciones, los parámetros p_g , p_r , p_o se estiman para cada SNP a partir de los datos de formación con antelación. En algunas realizaciones, los parámetros p_g , p_r , p_o se pueden actualizar con la muestra actual sobre la marcha, para tomar en cuenta la variación de la calidad de la muestra.

El modelo descrito en el presente documento es bastante general y puede tener en cuenta tanto la calidad diferencial de la muestra como la calidad diferencial del SNP. Las diferentes muestras y SNP se tratan de forma diferente, tal y como se pone de manifiesto por el hecho de que algunas realizaciones utilizan distribuciones Beta-binomiales cuya media y varianza son una función de la cantidad original de ADN, así como la calidad de la muestra y del SNP.

Elaboración de modelos de plataformas

Se asume un único SNP en el que el ratio del alelo previsto presente en el plasma es r (basándose en los genotipos maternos y fetales). El ratio del alelo previsto se define como la fracción prevista de los alelos A en el ADN combinado materno y fetal. Para el genotipo materno g_m y el genotipo del niño g_c , el ratio del alelo previsto viene dado por la ecuación 1, asumiendo que los genotipos son representados también como ratios de alelos.

$$r = fg_c + (1-f)g_m(1)$$

La observación en el SNP se compone del número de lecturas correlacionadas con cada alelo presente, n_a y n_b , que se suman a la profundidad de lectura d . Se asume que los umbrales ya se han aplicado a las probabilidades de correlación y las puntuaciones phred de forma que las correlaciones y las observaciones de alelos se pueden considerar correctas. Una puntuación phred es una medida numérica que se refiere a la probabilidad de que una medición concreta en una base concreta sea errónea. En una realización, cuando la base ha sido medida mediante secuenciación, la puntuación phred se puede calcular a partir del ratio de la intensidad del tinte correspondiente a la base determinada para la intensidad del tinte de las otras bases. El modelo más sencillo para la probabilidad de observación es una distribución binomial que asume que cada una de las lecturas d se extrae independientemente de un amplio conjunto que tiene un ratio de alelos r . La ecuación 2 describe este modelo.

$$P(n_a, n_b | r) = p_{\text{bino}}(n_a; n_a + n_b, r) = \binom{n_a + n_b}{n_a} r^{n_a} (1-r)^{n_b} (2)$$

El modelo binomial se puede ampliar de varias maneras. Cuando los genotipos materno y fetal son todos A o todos B, el ratio alélico previsto en el plasma será 0 o 1 y la probabilidad binomial no estará bien definida. En la práctica, en ocasiones se observan alelos imprevistos. En una realización, es posible utilizar un ratio alélico corregido $P = 1/(n_a + n_b)$

para permitir un pequeño número de alelos imprevistos. En una realización, se pueden utilizar los datos de formación para elaborar un modelo de la tasa de alelos imprevistos que aparece en cada SNP y utilizar este modelo para corregir el ratio alélico previsto. Cuando el ratio alélico previsto no es 0 o 1, el ratio alélico observado puede no coincidir con una profundidad de lectura suficientemente elevada para el ratio alélico previsto, debido al sesgo de amplificación u otros fenómenos. A continuación, se puede elaborar un modelo del ratio alélico como una distribución beta centrada en el ratio alélico previsto, lo que conduce a una distribución beta-binomial para $P(n_a, n_b|r)$ que tiene una varianza más elevada que la binomial.

El modelo de la plataforma para la respuesta en un único SNP se definirá como $F(a, b, g_c, g_m, f)$ (3), o la probabilidad de observar $n_a = a$ y $n_b = b$ dados los genotipos materno y fetal, que también depende de la fracción fetal a través de la ecuación 1. La forma funcional de F puede ser una distribución binomial, una distribución beta-binomial o funciones similares como las anteriormente descritas.

$$F(a, b, g_c, g_m, f) = P(n_a = a, n_b = b|g_c, g_m, f) = P(n_a = a, n_b = b|r(g_c, g_m, f)) \quad (3)$$

En una realización, se puede determinar la fracción del niño como sigue. Una estimación de la probabilidad máxima de la fracción fetal f para una prueba prenatal se puede obtener sin el uso de información paterna. Esto puede resultar relevante cuando no se dispone de los datos genéticos paternos, por ejemplo, cuando el padre registrado no es realmente el padre genético del feto. La fracción fetal se estima a partir del conjunto de SNP en los que el genotipo materno es 0 o 1, lo que resulta en un conjunto de solamente dos genotipos fetales posibles. Se define S_0 como el conjunto de SNP con el genotipo materno 0 y S_1 como el conjunto de SNP con el genotipo materno 1. Los genotipos fetales posibles en S_0 son 0 y 0,5, lo que resulta en un conjunto de ratios alélicos posibles $R_0(f) = \{0, f/2\}$. De forma similar, $R_1(f) = \{1-f/2, 1\}$.

Este método se puede ampliar de forma trivial para incluir los SNP donde el genotipo materno es 0,5, pero estos SNP serán menos informativos debido al conjunto mayor de ratios alélicos posibles.

Se define N_{a0} y N_{b0} como los vectores formados por n_{as} y n_{bs} para los SNP s en S_0 , y N_{a1} y N_{b1} similarmente para S_1 . La estimación de la probabilidad máxima $\hat{f}$ de f se define por la ecuación 4.

$$\hat{f} = \arg \max_f P(N_{a0}, N_{b0}|f) P(N_{a1}, N_{b1}|f) \quad (4)$$

Asumiendo que los recuentos de alelos en cada SNP son independientes condicionados por el ratio alélico del plasma en los SNP, las probabilidades se pueden expresar como productos sobre los SNP en cada conjunto (5).

$$P(N_{a0}, N_{b0}|f) = \prod_{s \in S_0} P(n_{as}, n_{bs}|f) \quad (5)$$

$$P(N_{a1}, N_{b1}|f) = \prod_{s \in S_1} P(n_{as}, n_{bs}|f)$$

La dependencia de f se produce a través de los posibles ratios alélicos $R_0(f)$ y $R_1(f)$. La probabilidad del SNP $P(n_{as}, n_{bs}|f)$ se puede aproximar asumiendo el genotipo de la probabilidad máxima condicionado de f . A una fracción fetal y una profundidad de lectura razonablemente elevadas, la selección del genotipo de la probabilidad máxima será de alta certeza. Por ejemplo, a una fracción fetal del 10 por ciento y una profundidad de lectura de 1000, se considera un SNP donde la madre tiene el genotipo cero. Los ratios alélicos previstos son 0 y 5 por ciento, que serán fácilmente distinguibles a una profundidad de lectura suficientemente elevada. La sustitución del genotipo del niño estimado en la ecuación 5 resulta en la ecuación completa (6) para la estimación de la fracción fetal.

$$\hat{f} = \arg \max_f \left[\prod_{s \in S_0} \left(\max_{r_s \in R_0(f)} P(n_{as}, n_{bs}|r_s) \right) \prod_{s \in S_1} \left(\max_{r_s \in R_1(f)} P(n_{as}, n_{bs}|r_s) \right) \right] \quad (6)$$

La fracción fetal debe estar en el rango $[0,1]$ y por tanto la optimización se puede implementar fácilmente a través de una búsqueda unidimensional limitada.

En presencia de una baja profundidad de lectura o un elevado nivel de ruido, puede ser preferible no asumir el genotipo de la probabilidad máxima, que puede proporcionar certezas artificialmente elevadas. Otro método consistiría en sumar los posibles genotipos de cada SNP, resultando en la siguiente expresión (7) para $P(n_a, n_b|f)$ para un SNP en S_0 . La probabilidad previa $P(r)$ se podría asumir uniforme en $R_0(f)$ o se podría basar en las frecuencias de la población. La extensión al grupo S_1 es insignificante.

$$P(n_a, n_b|f) = \sum_{r \in R_0(f)} P(n_a, n_b|r) P(r) \quad (7)$$

En algunas realizaciones, se pueden determinar las probabilidades del modo siguiente. Se puede calcular una certeza a partir de las probabilidades de los datos de las dos hipótesis H_t y H_r . La probabilidad de cada hipótesis se obtiene basándose en el modelo de respuesta, la fracción fetal estimada, los genotipos maternos, las frecuencias alélicas de la población y los recuentos alélicos del plasma.

Se define la siguiente notación:

G_m, G_c genotipos verdaderos materno y del niño

G_{af}, G_{tf} genotipos verdaderos del supuesto padre y del verdadero padre

$G(g_c, g_m, g_{tf}) = P(G_c = g_c | G_m = g_m, G_{tf} = g_{tf})$ probabilidades de herencia

$P(g) = P(G_{tf} = g)$ frecuencia de la población del genotipo g en un SNP concreto

5 Asumiendo que la observación en cada SNP es independiente condicionada del ratio alélico del plasma, la probabilidad de una hipótesis de paternidad es el producto de las probabilidades de los SNP. Las siguientes ecuaciones calculan la probabilidad para un único SNP. La ecuación 8 es una expresión general para la probabilidad de cualquier hipótesis h , que posteriormente se dividirá en casos específicos de H_t y H_f .

$$\begin{aligned} P(n_a, n_b | h, G_m, G_{tf}, f) &= \sum_{g_c \in (0,0.5,1)} P(n_a, n_b | G_c = g_c, G_m, G_{tf}, h, f) P(G_c = g_c | G_m, G_{tf}, h, f) \\ &= \sum_{g_c \in (0,0.5,1)} P(n_a, n_b | G_c = g_c, G_m, f) P(G_c = g_c | G_m, G_{tf}, h) \\ 10 &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf}, h) (8) \end{aligned}$$

En el caso de H_t , el supuesto padre es el verdadero padre y los genotipos fetales son heredados de los genotipos maternos y de los genotipos del supuesto padre en función de la ecuación 9.

$$\begin{aligned} P(n_a, n_b | H_t, G_m, G_{tf}, f) &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf}, H_t) (9) \\ 15 &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) G(g_c, G_m, G_{tf}) \end{aligned}$$

En el caso de H_f , el supuesto padre no es el verdadero padre. La mejor estimación de los genotipos del verdadero padre viene dada por las frecuencias de la población en cada SNP. Por tanto, las probabilidades de los genotipos del niño están determinadas por los genotipos de la madre conocidos y las frecuencias de la población, como en la ecuación 10.

$$\begin{aligned} P(n_a, n_b | H_f, G_m, G_{tf}, f) &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf}, H_f) \\ 20 &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m) \\ &= \sum_{g_c \in (0,0.5,1)} \sum_{g_{tf} \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf} = \square_{tf}) P(G_{tf} = g_{tf}) \\ &= \sum_{g_c \in (0,0.5,1)} \sum_{g_{tf} \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) G(g_c, G_m, g_{tf}) P(g_{tf}) \end{aligned}$$

La certeza C_p de la paternidad correcta se calcula a partir del producto de los SNP de las dos probabilidades utilizando la regla de Bayes (11).

$$25 \quad C_p = \frac{\prod_s P(n_{as}, n_{bs} | H_t, G_{ms}, G_{tf}, f)}{\prod_s P(n_{as}, n_{bs} | H_t, G_{ms}, G_{tf}, f) + \prod_s P(n_{as}, n_{bs} | H_f, G_{ms}, G_{tf}, f)} (11)$$

Modelo de la probabilidad máxima utilizando el porcentaje de la fracción fetal

La determinación del estado de ploidía de un feto midiendo el ADN flotante libre contenido en suero materno o midiendo el material genotípico de cualquier muestra mezclada, es un ejercicio nada trivial. Existen diversos métodos, por ejemplo, para realizar un análisis del recuento de lecturas donde se presume que si el feto es trisómico en un determinado cromosoma, entonces la cantidad total de ADN de ese cromosoma que se encuentra en la sangre materna será elevada con respecto al cromosoma de referencia. Una forma de detectar la trisomía en estos fetos consiste en normalizar la cantidad de ADN prevista para cada cromosoma, por ejemplo, en función del número de SNP del conjunto de análisis que corresponde a un cromosoma dado o en función del número de porciones del cromosoma correlacionables de forma única. Una vez que se han normalizado las mediciones, se determina que cualquiera de los cromosomas para los que la cantidad de ADN medida supere un determinado umbral son trisómicos. Este planteamiento se describe en Fan, et al. PNAS, 2008; 105(42); pp. 16266-16271, y también en Chiu et al. BMJ 2011;342:c7401. En el documento de Chiu et al., la normalización se realizó calculando una puntuación Z del modo siguiente:

40 Puntuación Z para el porcentaje del cromosoma 21 en el caso de ensayo = ((porcentaje del cromosoma 21 en el caso de ensayo) - (porcentaje medio del cromosoma 21 en controles de referencia)) / (desviación estándar del porcentaje del cromosoma 21 en controles de referencia).

Estos métodos determinan el estado de ploidía del feto utilizando un método de rechazo por hipótesis única. Sin embargo, sufren algunas carencias significativas. Puesto que estos métodos para la determinación del estado de ploidía en el feto son invariables en función del porcentaje de ADN fetal en la muestra, utilizan un valor de corte; el resultado de

esto es que las precisiones de las determinaciones no son óptimas y aquellos casos en los que el porcentaje de ADN fetal en la mezcla es relativamente bajo sufrirán las peores precisiones.

En una realización, se utiliza un método de la presente divulgación para determinar el estado de ploidía del feto que implica tener en cuenta la fracción de ADN fetal de la muestra. En otra realización de la presente divulgación, el método implica el uso de estimaciones de la probabilidad máxima. En una realización, un método de la presente divulgación implica calcular el porcentaje de ADN en una muestra que es de origen fetal o placentario. En una realización, el umbral para la determinación del estado de aneuploidía se ajusta adaptativamente sobre el porcentaje de ADN fetal calculado. En algunas realizaciones, el método para estimar el porcentaje de ADN que es de origen fetal en una mezcla de ADN comprende la obtención de una muestra mezclada que contiene material genético de la madre y material genético del feto, la obtención de una muestra genética del padre del feto, la medición del ADN en la muestra mezclada, la medición del ADN en la muestra del padre, y el cálculo del porcentaje de ADN que es de origen fetal en la muestra mezclada utilizando las mediciones de ADN de la muestra mezclada y de la muestra del padre.

En una realización de la presente divulgación, la fracción de ADN fetal, o el porcentaje de ADN fetal en la mezcla se pueden medir. En algunas realizaciones, la fracción se puede calcular utilizando solo mediciones del genotipo realizadas con la propia muestra de plasma materno, que es una mezcla de ADN fetal y materno. En algunas realizaciones, la fracción se puede calcular utilizando también el genotipo medido o conocido de otro modo de la madre y/o el genotipo medido o conocido de otro modo del padre. En algunas realizaciones, el porcentaje de ADN fetal se puede calcular utilizando mediciones realizadas en la mezcla de ADN materno y fetal, junto con el conocimiento de los contextos parentales. En una realización, la fracción de ADN fetal se puede calcular utilizando frecuencias de la población para ajustar el modelo de la probabilidad a mediciones alélicas específicas.

En una realización de la presente divulgación, se puede calcular una certeza de la precisión de la determinación del estado de ploidía del feto. En una realización, la certeza de la hipótesis de mayor probabilidad ($H_{\text{principal}}$) se puede calcular como $(1 - H_{\text{principal}}) / \sum (\text{todas las } H)$. Se puede determinar la certeza de una hipótesis si se conocen las distribuciones de todas las hipótesis. Se puede determinar la distribución de todas las hipótesis, si se conoce la información del genotipo parental. Se puede calcular una certeza de la determinación del estado de ploidía si se conoce la distribución de los datos prevista para el feto euploide y la distribución de los datos prevista para el feto aneuploide. Se pueden calcular estas distribuciones previstas si se conocen los datos del genotipo parental. En una realización, se puede utilizar el conocimiento de la distribución de una estadística de ensayo con respecto a una hipótesis normal y con respecto a una hipótesis anómala para determinar tanto la fiabilidad de la determinación, así como también para reajustar el umbral y obtener una determinación más fiable. Esto resulta particularmente útil cuando la cantidad y/o el porcentaje de ADN fetal en la mezcla son bajos. Ayudará a evitar la situación en la que un feto que es realmente aneuploide se determina que es euploide porque una estadística de ensayo, como la estadística Z, no supera un umbral que se establece basándose en un umbral que está optimizado para el caso en el que existe un porcentaje de ADN fetal superior.

En una realización, se puede utilizar un método divulgado en el presente documento para determinar una aneuploidía fetal estableciendo el número de copias de cromosomas diana maternos y fetales en una mezcla de material genético materno y fetal. Este método puede implicar la obtención de tejido materno que comprende tanto material genético materno como fetal; en algunas realizaciones, este tejido materno puede ser plasma materno o un tejido aislado de sangre materna. Este método también puede implicar la obtención de una mezcla de material genético materno y fetal de dicho tejido materno mediante el procesamiento del tejido materno mencionado anteriormente. Este método puede implicar la distribución del material genético obtenido en una pluralidad de mezclas de reacción para proporcionar aleatoriamente muestras de reacción individuales que comprenden una secuencia diana de un cromosoma diana y muestras de reacción individuales que no comprenden una secuencia diana de un cromosoma diana, por ejemplo, realizando una secuenciación de alto rendimiento de la muestra. Este método puede implicar el análisis de secuencias diana de material genético presente o ausente en dichas muestras de reacción individuales para proporcionar un primer número de resultados binarios que representan la presencia o ausencia de un cromosoma fetal presumiblemente euploide en las muestras de la reacción y un segundo número de resultados binarios que representan la presencia o ausencia de un cromosoma fetal posiblemente aneuploide en las muestras de la reacción. Cualquiera de los números de resultados binarios se puede calcular, por ejemplo, mediante una técnica informática que recuenta las lecturas de secuencia que corresponden a un determinado cromosoma, a una determinada región de un cromosoma, a un determinado locus o conjunto de loci. El método puede implicar la normalización del número de acontecimientos binarios basándose en la longitud del cromosoma, la longitud de la región del cromosoma o el número de loci del conjunto. Este método puede implicar el cálculo de una distribución prevista del número de resultados binarios para un cromosoma fetal presumiblemente euploide en las muestras de la reacción utilizando el primer número. Este método puede implicar el cálculo de una distribución prevista del número de resultados binarios para un cromosoma fetal presumiblemente aneuploide en las muestras de la reacción utilizando el primer número y una fracción estimada de ADN fetal que se encuentra en la muestra, por ejemplo, multiplicando la distribución del recuento leído previsto del número de resultados binarios para un cromosoma fetal presumiblemente euploide por $(1 + n/2)$ donde n es la fracción fetal prevista. En algunas realizaciones, las lecturas de secuencia se pueden tratar en correspondencias probabilísticas más que resultados binarios; este método proporcionaría precisiones más elevadas, aunque requiere una mayor potencia de cálculo. La fracción fetal se puede estimar a través de una pluralidad de métodos, algunos de los cuales se describen en otro apartado de la presente divulgación. Este método puede implicar el uso de un planteamiento de la probabilidad máxima para determinar si el segundo número corresponde al cromosoma fetal posiblemente aneuploide que es

euploide o aneuploide. Este método puede implicar que la determinación del estado de ploidía del feto sea el estado de ploidía que corresponde a la hipótesis con la probabilidad máxima de ser correcta dados los datos medidos.

Cabe señalar que se puede utilizar un modelo de probabilidad máxima para aumentar la precisión de cualquier método que determine el estado de ploidía de un feto. De forma similar, se puede calcular una certeza para cualquier método que determine el estado de ploidía del feto. El uso de un modelo de probabilidad máxima supondría una mejora de la precisión de cualquier método en el que se determine el estado de ploidía utilizando una técnica de rechazo de una única hipótesis. Se puede utilizar un modelo de probabilidad máxima para cualquier método en el que se pueda calcular una distribución de la probabilidad tanto para los casos normales como para los anómalos. El uso de un modelo de probabilidad máxima implica la capacidad de calcular una certeza para la determinación del estado de ploidía.

Exposición más detallada del método

En una realización, un método divulgado en el presente documento utiliza una medición cuantitativa del número de observaciones independientes de cada alelo en un locus polimórfico, donde esto no implica el cálculo del ratio de los alelos. Esto contrasta con otros métodos, como algunos métodos basados en microarrays, que proporcionan información sobre el ratio de dos alelos en un locus, pero no cuantifican el número de observaciones independientes de ninguno de los alelos. Algunos métodos conocidos en la técnica pueden proporcionar información cuantitativa sobre el número de observaciones independientes, pero los cálculos que conducen a la determinación del estado de ploidía utilizan únicamente los ratios de alelos y no utilizan la información cuantitativa. Para ilustrar la importancia de retener información sobre el número de observaciones independientes, consideremos el locus de la muestra con dos alelos, A y B. En un primer experimento se observan 20 alelos A y 20 alelos B, y en un segundo experimento se observan 200 alelos A y 200 alelos B. En ambos experimentos el ratio $(A/(A+B))$ es igual a 0,5; sin embargo, el segundo experimento proporciona más información que el primero acerca de la certidumbre de la frecuencia del alelo A o B. El método instantáneo, en lugar de utilizar los ratios de los alelos, utiliza datos cuantitativos para elaborar un modelo más preciso de las frecuencias alélicas más probables en cada locus polimórfico.

En una realización, los métodos instantáneos crean un modelo genético para sumar las mediciones de múltiples loci polimórficos, con el fin de distinguir mejor la trisomía de la disomía y también para determinar el tipo de trisomía. Por otra parte, el método instantáneo incorpora información de la unión genética para mejorar la precisión del método. Esto contrasta con algunos métodos conocidos en la técnica, donde se establece la media de los ratios de los alelos de todos los loci polimórficos de un cromosoma. El método divulgado en el presente documento establece un modelo explícitamente de las distribuciones de la frecuencia alélica prevista en la disomía, así como la trisomía resultante de la ausencia de disyunción durante la meiosis I, la ausencia de disyunción durante la meiosis II, y la ausencia de disyunción durante la mitosis temprana en el desarrollo fetal. Para ilustrar por qué esto es importante, si no se produjesen cruces, la ausencia de disyunción durante la meiosis I resultaría en una trisomía en la que dos homólogos diferentes se heredaron de un progenitor; la ausencia de disyunción durante la meiosis II o durante la mitosis temprana en el desarrollo fetal resultaría en dos copias del mismo homólogo de un progenitor. Cada escenario daría como resultado unas frecuencias alélicas previstas diferentes en cada locus polimórfico y también en todos los loci unidos genéticamente (es decir, los loci en el mismo cromosoma) considerados conjuntamente. Los cruces, que provocan el intercambio de material genético entre homólogos, hacen que el patrón de herencia sea más complejo; sin embargo, el método instantáneo tiene esto en cuenta utilizando información sobre la unión genética, es decir, información sobre la tasa de recombinación y la distancia física entre loci. Para distinguir mejor entre la ausencia de disyunción de la meiosis I y la ausencia de disyunción de la meiosis II o la mitosis, el método instantáneo incorpora en el modelo una probabilidad creciente de cruce dado que la distancia desde el centrómero aumenta. La ausencia de disyunción de la meiosis II y la mitosis se puede distinguir por el hecho de que la ausencia de disyunción mitótica típicamente resulta en copias idénticas o prácticamente idénticas de un homólogo, mientras que los dos homólogos presentes tras un evento de ausencia de disyunción de la meiosis II a menudo difieren debido a uno o más cruces durante la gametogénesis.

En una realización, un método de la presente divulgación no puede determinar los haplotipos de los padres si se asume una disomía. En una realización, en caso de trisomía, el método instantáneo puede realizar una determinación sobre los haplotipos de uno de los progenitores o ambos utilizando el hecho de que el plasma toma dos copias de un progenitor, y la información de fase del progenitor se puede determinar analizando qué dos copias se han heredado del progenitor en cuestión. En concreto, un niño puede heredar dos de las mismas copias del progenitor (trisomía emparejada) o ambas copias del progenitor (trisomía no emparejada). En cada SNP se puede calcular la probabilidad de la trisomía emparejada y de la trisomía no emparejada. Un método para determinar el estado de ploidía que no utiliza el modelo de enlace que tiene en cuenta los cruces calcularía la probabilidad total de la trisomía como una media simple ponderada de las trisomías emparejadas y no emparejadas de todos los cromosomas. Sin embargo, debido a los mecanismos biológicos que provocan el error de disyunción y cruce, la trisomía puede cambiar de emparejada a no emparejada (y viceversa) en un cromosoma únicamente si se produce un cruce. El método instantáneo probabilísticamente tiene en cuenta la probabilidad de cruce, por lo que produce determinaciones del estado de ploidía más precisas que los métodos que no tienen esto en cuenta.

En una realización, se utiliza un cromosoma de referencia para determinar la fracción del niño y la cantidad del nivel de ruido o distribución de la probabilidad. En una realización, la fracción del niño, el nivel de ruido y/o la distribución de la probabilidad se determina utilizando únicamente la información genética disponible de los cromosomas cuyo estado de ploidía se está determinando. El método instantáneo funciona sin el cromosoma de referencia y también sin establecer la fracción del niño o el nivel de ruido específicos. Esto representa una mejora significativa y un punto de diferenciación

con respecto a los métodos conocidos en la técnica, donde los datos genéticos de un cromosoma de referencia son necesarios para calibrar la fracción del niño y el comportamiento del cromosoma.

En una realización donde un cromosoma de referencia no es necesario para determinar la fracción fetal, la determinación de la hipótesis se realiza del modo siguiente:

$$H^* = \operatorname{argmax}_H \operatorname{LIK}(D|H) * \operatorname{priorprob}(H)$$

Con el algoritmo con el cromosoma de referencia, normalmente se asume que el cromosoma de referencia es una disomía y entonces se puede (a) determinar la fracción del niño más probable y el nivel de ruido aleatorio N basándose en este supuesto y en los datos del cromosoma de referencia:

$$[cfr^*, N^*] = \operatorname{argmax}_{cfr, N} \operatorname{LIK}(D(\text{ref. chrom})|H11, cfr, N)$$

Y, a continuación, reducir

$$\operatorname{LIK}(D|H) = \operatorname{LIK}(D|H, cfr^*, N^*)$$

o (b) estimar la distribución de la fracción del niño y el nivel de ruido basándose en este supuesto y en los datos del cromosoma de referencia. En concreto, no se determinaría un solo valor para cfr y N, sino que se asignaría la probabilidad $p(cfr, N)$ para el rango más amplio de posibles valores cfr, N:

$$p(cfr, N) \sim \operatorname{LIK}(D(\text{ref. chrom})|H11, cfr, N) * \operatorname{priorprob}(cfr, N)$$

donde $\operatorname{priorprob}(cfr, N)$ es la probabilidad previa de la fracción del niño y el nivel de ruido concretos, determinados por el conocimiento y los experimentos anteriores. Si se desea, simplemente uniformes con respecto al rango de cfr, N. A continuación, se puede escribir:

$$\operatorname{LIK}(D|H) = \sum_{cfr, N} \operatorname{LIK}(D|H, cfr, N) * p(cfr, N)$$

Los dos métodos anteriores dan buenos resultados.

Cabe señalar que en algunos casos el uso de un cromosoma de referencia no resulta recomendable, posible o factible. En este caso, se puede obtener la mejor determinación del estado de ploidía para cada cromosoma por separado. En concreto:

$$\operatorname{LIK}(D|H) = \sum_{cfr, N} \operatorname{LIK}(D|H, cfr, N) * p(cfr, N|H)$$

$p(cfr, N|H)$ Se puede determinar como se ha indicado anteriormente, para cada cromosoma por separado, asumiendo la hipótesis H, no solo para el cromosoma de referencia que asume la disomía. Utilizando este método se pueden mantener fijados los parámetros del ruido y la fracción del niño, se puede fijar uno de los dos parámetros o se pueden mantener ambos parámetros en forma probabilística para cada cromosoma y cada hipótesis.

Las mediciones de ADN son propensas a los ruidos y/o errores, en especial las mediciones donde la cantidad de ADN es pequeña o donde el ADN está mezclado con ADN contaminante. Este ruido genera datos genotípicos menos precisos y determinaciones del estado de ploidía menos precisas. En algunas realizaciones, se pueden utilizar los modelos de las plataformas o algún otro método de modelos de ruido para contrarrestar los efectos nocivos del ruido sobre la determinación del estado de ploidía. El método instantáneo utiliza un modelo conjunto de ambos canales, que tiene en cuenta el ruido aleatorio debido a la cantidad de ADN aportado, la calidad del ADN y/o la calidad del protocolo.

Esto contrasta con algunos métodos conocidos en la técnica donde las determinaciones del estado de ploidía se realizan utilizando el ratio de intensidades alélicas en un locus. Este método excluye los modelos de ruido de SNP precisos. En concreto, por lo general los errores de las mediciones no dependen específicamente del ratio de intensidad del canal medido, lo que reduce el modelo al uso de la información unidimensional. La elaboración de modelos precisos del ruido, la calidad del canal y la interacción del canal requiere un modelo conjunto bidimensional, que no se puede modelar utilizando ratios alélicos.

En concreto, la proyección de la información de dos canales para el ratio r donde $f(x, y)$ es $r = x/y$, no resulta recomendable para los modelos precisos de ruido del canal y sesgo. El ruido de un SNP concreto no es una función del ratio, es decir, $\text{ruido}(x, y) \neq f(x, y)$ sino que se trata de hecho de una función conjunta de ambos canales. Por ejemplo, en el modelo binomial, el ruido del ratio medido tiene una varianza de $r(1-r)/(x+y)$ que no es una función puramente de r . En este modelo, cuando se incluye cualquier ruido o sesgo del canal, se supone que en SNP i , el valor X del canal observado es $x = a_i X + b_i$, donde X es el valor verdadero del canal, b_i es el sesgo del canal y el ruido aleatorio adicionales. De forma similar, se supone que $y = c_i Y + d_i$. El ratio observado $r = x/y$ no puede predecir de forma precisa el ratio verdadero X/Y ni elaborar un modelo del ruido restante, dado que $(a_i X + b_i)/(c_i Y + d_i)$ no es una función de X/Y .

El método divulgado en el presente documento describe una forma efectiva de elaborar un modelo del ruido y el sesgo utilizando distribuciones binomiales conjuntas de todos los canales de medición individualmente. Las ecuaciones relevantes se pueden encontrar en otro apartado del presente documento en las secciones que tratan sobre el sesgo constante por SNP, $P(\text{buena})$ y $P(\text{ref}|mala)$, $P(\text{mut}|mala)$ que se ajustan efectivamente al comportamiento de los SNP.

En una realización, un método de la presente divulgación utiliza una distribución Beta-binomial, que evita la práctica restrictiva de confiar exclusivamente en los ratios de los alelos, pero que elabora modelos del comportamiento basándose en los recuentos de ambos canales.

En una realización, un método divulgado en el presente documento puede determinar el estado de ploidía de un feto en gestación a partir de los datos genéticos que se encuentran en el plasma materno utilizando todas las mediciones disponibles. En una realización, un método divulgado en el presente documento puede determinar el estado de ploidía de un feto en gestación a partir de los datos genéticos que se encuentran en el plasma materno utilizando las mediciones de solo un subconjunto de contextos parentales. Algunos métodos conocidos en la técnica solamente utilizan datos genéticos medidos en los que el contexto parental procede del contexto AA|BB, es decir, cuando ambos progenitores son homocigotos en un determinado locus, pero para un alelo diferente. Un problema de este método es que una pequeña proporción de loci polimórficos procede del contexto AA|BB, típicamente menos del 10%. En una realización de un método divulgado en el presente documento, el método no utiliza mediciones genéticas del plasma materno realizadas en loci en los que el contexto parental es AA|BB. En una realización, el método instantáneo utiliza mediciones de plasma solo para aquellos loci polimórficos con el contexto parental AA|AB, AB|AA, y AB|AB.

Algunos métodos conocidos en la técnica implican la determinación de la media de los ratios de los alelos de los SNP en el contexto AA|BB, donde los genotipos de ambos progenitores se encuentran presentes, y reivindican la determinación del estado de ploidía a partir del ratio medio de los alelos en estos SNP. Este método presenta una imprecisión significativa debido al comportamiento diferencial de los SNP. Cabe señalar que este método asume que se conocen los genotipos de ambos progenitores. Por el contrario, en algunas realizaciones, el método instantáneo utiliza un modelo de distribución de canales conjunto que no asume la presencia de ninguno de los progenitores ni asume el comportamiento uniforme de los SNP. En algunas realizaciones, el método instantáneo tiene en cuenta la diferente ponderación/comportamiento de los SNP. En algunas realizaciones, el método instantáneo no requiere el conocimiento de uno o ambos genotipos parentales. Un ejemplo de cómo puede conseguir esto el método instantáneo es el siguiente:

En algunas realizaciones, la probabilidad logarítmica de una hipótesis se puede determinar para cada SNP. En un SNP i concreto, asumiendo la hipótesis de ploidía fetal H y un porcentaje de ADN fetal cf, la probabilidad logarítmica de los datos observados D se define como:

$$\text{LIK}(D|H, i) = \log P(D|H, cf, i) = \log \left(\sum_{m,f,c} P(D|m, f, c, H, cf, i) P(c|m, f, H) P(m|i) P(f|i) \right)$$

donde m son posibles genotipos verdaderos de la madre, f son posibles genotipos verdaderos del padre, donde m, f c {AA,AB,BB}, y donde c son posibles genotipos del niño dada la hipótesis H. En concreto, para la monosomía c (A,B), para la disomía c c (AA,AB,BB), para la trisomía c e (AAA,AAB,ABB,BBB). Cabe señalar que la inclusión de los datos genotípicos parentales típicamente proporciona determinaciones de la ploidía más precisas; sin embargo, los datos genotípicos parentales no son necesarios para que el método instantáneo funcione bien.

Algunos métodos conocidos en la técnica implican la determinación de la media de los ratios de alelos de los SNP en los que la madre es homocigota, pero un alelo diferente se mide en el plasma (contextos AA|AB o AA|BB) y reivindican determinar los estados del ploidía a partir del ratio de alelos medio de estos SNP. Este método está pensado para los casos en los que el genotipo parental no está disponible. Cabe señalar que es cuestionable la precisión con la que se puede reivindicar que el plasma es heterocigoto en un SNP concreto sin la presencia de un padre homocigoto y opuesto BB: para los casos con una escasa fracción del niño, lo que parece la presencia de un alelo B podría ser simplemente la presencia de ruido; por otra parte, lo que parece la ausencia de B podría ser una simple pérdida de alelos de las mediciones fetales. Incluso en el caso de que se pueda determinar realmente la heterocigosidad del plasma, este método no será capaz de distinguir trisomías paternas. En concreto, para los SNP en los que la madre es AA, y donde algún B se mide en el plasma, si el padre es GG, el genotipo del niño resultante es AGG, lo que produce un ratio medio del 33% de A (para una fracción del niño=100%). Sin embargo, en el caso en el que el padre es AG, el genotipo del niño resultante podría ser AGG para la trisomía emparejada, contribuyendo al ratio del 33% de A, o AAG para la trisomía no emparejada, arrastrando el ratio medio más hacia un 66% de A. Dado que muchas trisomías se encuentran en los cromosomas con cruces, el cromosoma total puede encontrarse en cualquier estado entre la ausencia de trisomía no emparejada y todas las trisomías no emparejadas, y este ratio puede variar entre el 33 y el 66%. Para una disomía plana, el ratio se debería encontrar en torno al 50%>. Sin el uso de un modelo de enlace o un modelo de error preciso de la media, este método no detectaría muchos casos de trisomía parental. Por el contrario, el método divulgado en el presente documento asigna probabilidades del genotipo parental para cada candidato genotípico parental, basándose en la información genotípica disponible y la frecuencia de la población, y no requiere explícitamente genotipos parentales. Por otra parte, el método divulgado en el presente documento es capaz de detectar la trisomía incluso en ausencia o presencia de datos genotípicos parentales y puede compensar identificando los puntos de posibles cruces de trisomía emparejada a no emparejada utilizando un modelo de enlace.

Algunos métodos conocidos en la técnica reivindican un método para determinar la media de los ratios alélicos de los SNP cuando no se conoce el genotipo materno o paterno, y para determinar el estado de ploidía a partir del ratio medio en estos SNP. Sin embargo, no se divulga un método para lograr este fin. El método divulgado en el presente documento es capaz de realizar determinaciones precisas del estado de ploidía en esta situación, y la reducción a la práctica se divulga en otro apartado del documento, utilizando un método de la probabilidad máxima conjunto y opcionalmente utiliza modelos de sesgo y ruido de los SNP, así como un modelo de enlace.

Algunos métodos conocidos en la técnica implican la estimación de la media de los ratios alélicos y reivindican la determinación de los estados de ploidía a partir del ratio alélico medio en uno o unos cuantos SNP. Sin embargo, estos métodos no utilizan el concepto del enlace. Los métodos divulgados en el presente documento no presentan estas desventajas.

5 Utilización de la longitud de la secuencia como dato previo para determinar el origen del ADN

Se ha determinado que la distribución de la longitud de las secuencias difiere entre el ADN materno y fetal, siendo por lo general más cortas en el fetal. En una realización de la presente divulgación, se puede utilizar el conocimiento previo en forma de datos empíricos y construir una distribución previa para la longitud prevista del ADN tanto materno ($P(X| \text{materno})$) como fetal ($P(X| \text{fetal})$). Dada una nueva secuencia de ADN no identificada de longitud x , se puede asignar una probabilidad de que una determinada secuencia de ADN sea ADN materno o fetal, basándose en la probabilidad
10 previa de que x sea materna o fetal. En concreto, si $P(x|\text{materna}) > P(x|\text{fetal})$, entonces la secuencia de ADN se puede clasificar como materna, con $P(x|\text{materna}) = P(x|\text{materna}) / [P(x|\text{materna}) + P(x|\text{fetal})]$, y si $p(x|\text{materna}) < p(x|\text{fetal})$, entonces la secuencia de DNA se puede clasificar como fetal, $P(x|\text{fetal}) = P(x|\text{fetal}) / [P(x|\text{materna}) + P(x|\text{fetal})]$. En una realización de la presente divulgación, una distribución de las longitudes de secuencia materna y fetal se puede
15 determinar que es específica para esa muestra considerando las secuencias que se pueden clasificar como maternas o fetales con una elevada probabilidad y, a continuación, esa distribución específica de la muestra se puede utilizar como la distribución del tamaño prevista para esa muestra.

Profundidad de lectura variable para minimizar el coste de la secuenciación

En múltiples ensayos clínicos relativos al diagnóstico, por ejemplo, en Chiu et al. BMJ 2001;342:c7401, se establece un
20 protocolo con un número de parámetros y, a continuación, el mismo protocolo se ejecuta con los mismos parámetros para cada uno de los pacientes del ensayo. En el caso de la determinación del estado de ploidía de un feto en gestación en una madre utilizando la secuenciación como método para medir el material genético, un parámetro pertinente es el número de lecturas. El número de lecturas se puede referir al número de lecturas reales, el número de lecturas previstas, líneas parciales, líneas completas o células del flujo completas en un secuenciador. En estos estudios, el
25 número de lecturas se fija típicamente a un nivel que garantizará que todas o prácticamente todas las muestras alcancen el nivel de precisión deseado. En estos momentos la secuenciación es una tecnología cara, con un coste de unos 200 dólares por cinco millones de lecturas correlacionables, y a pesar de que el precio está bajando, cualquier método que permita un diagnóstico basado en la secuenciación para operar con un nivel de precisión similar pero menos lecturas supondrá necesariamente un ahorro considerable de dinero.

La precisión de la determinación del estado de ploidía depende típicamente de una serie de factores, incluyendo el número de lecturas y la fracción del ADN fetal en la mezcla. La precisión es típicamente superior cuando la fracción de ADN fetal de la mezcla es superior. Al mismo tiempo, la precisión es típicamente superior si el número de lecturas es mayor. Es posible tener una situación con dos casos en la que el estado de ploidía se determine con precisiones comparables, donde el primer caso tiene una fracción más baja de ADN fetal en la mezcla que el segundo, y se han
35 secuenciado más lecturas en el primer caso que en el segundo. Se puede utilizar la fracción estimada de ADN fetal de la mezcla como guía para determinar el número de lecturas necesarias para conseguir un determinado nivel de precisión.

En una realización de la presente divulgación, se puede utilizar un conjunto de muestras donde las diferentes muestras del conjunto se secuencian a diferentes profundidades de lectura, donde el número de lecturas ejecutado en cada una
40 de las muestras se selecciona para conseguir un determinado nivel de precisión teniendo en cuenta la fracción calculada de ADN fetal en cada mezcla. En una realización de la presente divulgación, esto puede implicar realizar una medición de la muestra mezclada para determinar la fracción de ADN fetal en la mezcla; esta estimación de la fracción fetal se puede realizar con secuenciación, se puede realizar con TAQMAN, se puede realizar con qPCR, se puede realizar con arrays de SNP, se puede realizar con cualquier método que pueda distinguir diferentes alelos en un determinado loci. La necesidad de la estimación de la fracción fetal se puede eliminar incluyendo hipótesis que cubran
45 todas las fracciones fetales o un conjunto de fracciones fetales del conjunto de hipótesis consideradas cuando se comparan con los datos medidos reales. Una vez que se ha determinado la fracción de ADN fetal de la muestra, se puede determinar el número de secuencias a leer para cada muestra.

En una realización de la presente divulgación, 100 mujeres embarazadas visitan a sus respectivos obstetras que les extraen sangre en tubos de análisis con un anti-lisante y/o algún elemento para inactivar la ADNasa. Se llevan a casa un
50 kit para el padre de su feto en gestación y el padre proporciona una muestra de saliva. Ambos conjuntos de materiales genéticos de las 100 parejas se envían al laboratorio, donde se agita la sangre materna y se aísla la capa leucocitaria, así como el plasma. El plasma comprende una mezcla de ADN materno y ADN de origen placentario. Se determina el genotipo de la capa leucocitaria materna y la sangre paterna utilizando un array de SNP, y el ADN de las muestras de plasma materno se focaliza con sondas de hibridación SURESELECT. El ADN extraído con las sondas se utiliza para generar 100 bibliotecas etiquetadas, una para cada una de las muestras maternas, donde cada muestra es etiquetada con una etiqueta diferente. Se extrae una fracción de cada biblioteca, cada una de esas fracciones se mezcla y añade a dos líneas de un secuenciador de ADN ILLUMINA HISEQ de forma multiplexada, donde cada línea da como resultado
55 unos 50 millones de lecturas correlacionables, lo que proporciona unos 100 millones de lecturas correlacionables en las 100 mezclas multiplexadas o aproximadamente un millón de lecturas por muestra. Las lecturas de secuencia se utilizaron para determinar la fracción de ADN fetal de cada mezcla. 50 de las muestras tenían más de un 15% de ADN

fetal en la mezcla y un millón de lecturas fue suficiente para determinar el estado de ploidía de los fetos con una certeza del 99,9%.

De las mezclas restantes, 25 tenían entre un 10 y un 15% de ADN fetal; una fracción de cada una de las bibliotecas relevantes preparada a partir de estas mezclas se multiplexó y se procesó en la línea del HISEZ generando otros dos millones de lecturas para cada muestra. Los dos conjuntos de datos de secuencias para cada una de las mezclas con un 10-15% de ADN fetal se juntaron y los tres millones de lecturas por muestra resultantes fueron suficientes para determinar el estado de ploidía de esos fetos con una certeza del 99,9%.

De las mezclas restantes, 13 tenían entre un 6 y un 10% de ADN fetal; una fracción de cada una de las bibliotecas relevantes preparada a partir de estas mezclas se multiplexó y se procesó en la línea del HISEZ generando otros cuatro millones de lecturas para cada muestra. Los dos conjuntos de datos de secuencias para cada una de las mezclas con un 6-10% de ADN fetal se juntaron y los 5 millones de lecturas totales por mezcla resultantes fueron suficientes para determinar el estado de ploidía de esos fetos con una certeza del 99,9%.

De las mezclas restantes, 8 tenían entre un 4 y un 6% de ADN fetal; una fracción de cada una de las bibliotecas relevantes preparada a partir de estas mezclas se multiplexó y se procesó en la línea del HISEZ generando otros seis millones de lecturas para cada muestra. Los dos conjuntos de datos de secuencias para cada una de las mezclas con un 4-6% de ADN fetal se juntaron y los siete millones de lecturas totales por mezcla resultantes fueron suficientes para determinar el estado de ploidía de esos fetos con una certeza del 99,9%.

De las cuatro mezclas restantes, todas ellas tenían entre un 2 y un 4% de ADN fetal; una fracción de cada una de las bibliotecas relevantes preparada a partir de estas mezclas se multiplexó y se procesó en la línea del HISEZ generando otros 12 millones de lecturas para cada muestra. Los dos conjuntos de datos de secuencias para cada una de las mezclas con un 2-4% de ADN fetal se juntaron y los 13 millones de lecturas totales por mezcla resultantes fueron suficientes para determinar el estado de ploidía de esos fetos con una certeza del 99,9%.

Este método precisó seis líneas de secuenciación en una máquina HISEQ para conseguir una precisión del 99,9% con las 100 muestras. Si se hubiese requerido el mismo número de procesamientos para cada muestra, para garantizar una determinación del estado de ploidía con una precisión del 99,9% se habrían necesitado 25 líneas de secuenciación, y si se hubiese tolerado una tasa de no determinación o una tasa de error del 4% se habrían podido obtener 14 líneas de secuenciación.

Uso de datos en bruto del genotipo

Hay una serie de métodos que pueden realizar un NPD utilizando información genética fetal medida en ADN fetal que se encuentra en la sangre materna. Algunos de estos métodos implican la realización de mediciones del ADN fetal utilizando arrays de SNP, algunos métodos implican la secuenciación no focalizada y algunos métodos implican la secuenciación focalizada. La secuenciación focalizada puede focalizarse en los SNP, puede focalizarse en STR, puede focalizarse en otros loci polimórficos, puede focalizarse en loci no polimórficos o en alguna combinación de estos. Algunos de estos métodos pueden implicar el uso de un instrumento de determinación alélica propio o comercializado que determine la identidad de los alelos a partir de los datos de intensidad procedentes de los sensores de la máquina que realiza la medición. Por ejemplo, el sistema ILLUMINA INFINIUM o el sistema de microarrays AFFYMETRIX GENECHIP utilizan perlas o microchips con secuencias de ADN adjuntas que se pueden hibridar con segmentos complementarios de ADN; tras la hibridación, se produce un cambio en las propiedades fluorescentes de la molécula sensora que puede ser detectada. También existen métodos de secuenciación, por ejemplo, ILLUMINA SOLEXA GENOME SEQUENCER o ABI SOLID GENOME SEQUENCER, donde la secuencia genética de fragmentos de ADN es secuenciada; tras la ampliación de la cadena de ADN complementaria a la cadena que se va a secuenciar, la identidad del nucleótido ampliado se detecta típicamente a través de una etiqueta fluorescente o radioetiqueta anexada al nucleótido complementario. En todos estos métodos los datos genotípicos o de secuenciación se determinan típicamente sobre la base de señales fluorescentes o de otro tipo, o de la ausencia de estas. Estos sistemas se combinan típicamente con paquetes de software de bajo nivel que realizan determinaciones alélicas específicas (datos genéticos secundarios) a partir de los datos analógicos del dispositivo de detección de fluorescencia o de otro tipo (datos genéticos primarios). Por ejemplo, en el caso de un alelo determinado en un array de SNP, el software determinará, por ejemplo, que un determinado SNP se encuentra presente o no presente si la intensidad de la fluorescencia se encuentra por encima o por debajo de un cierto umbral. De forma similar, el producto de un secuenciador es un cromatograma que indica el nivel de fluorescencia detectado para cada uno de los tintes, y el software determinará que un determinado par de bases es A o T o C o G. Los secuenciadores de alto rendimiento típicamente realizan una serie de mediciones de este tipo, denominadas lecturas, que representan la estructura más probable de la secuencia de ADN que se ha secuenciado. El producto análogo directo del cromatograma se define aquí como datos genéticos primarios, y la determinación del par de bases/SNP realizada por el software se considera en el presente documento los datos genéticos secundarios. En una realización, por datos primarios se entiende los datos de intensidad bruta que son el producto no procesado de una plataforma de determinación del genotipo, donde la plataforma de determinación del genotipo se puede referir a un array de SNP o a una plataforma de secuenciación. Los datos genéticos secundarios se refieren a los datos genéticos procesados, donde se ha realizado una determinación alélica, o se han asignado pares de bases a los datos de la secuencia, y/o las lecturas de secuencia se han correlacionado con el genoma.

Muchas aplicaciones de nivel superior aprovechan estas determinaciones alélicas, determinaciones de los SNP y lecturas de secuencia, es decir, los datos genéticos secundarios que produce el software de determinación del genotipo. Por ejemplo, DNA NEXUS, ELAND o MAQ realizarán las lecturas de secuenciación y las correlacionarán con el genoma. Por ejemplo, en el contexto del diagnóstico prenatal no invasivo, un sistema informático complejo, como PARENTAL SUPPORT™, puede aprovechar un gran número de determinaciones de SNP para determinar el genotipo de un individuo. Por otra parte, en el contexto del diagnóstico genético previo al implante, se puede tomar un conjunto de lecturas de secuencia correlacionadas con el genoma y realizando un recuento normalizado de las lecturas que están correlacionadas con cada cromosoma, o con una sección de un cromosoma, se puede determinar el estado de ploidía de un individuo. En el contexto del diagnóstico prenatal no invasivo, se puede tomar un conjunto de lecturas de secuencia que se han medido en el ADN presente en el plasma materno y correlacionarlas con el genoma. A continuación, se puede realizar un recuento normalizado de las lecturas correlacionadas con cada cromosoma, o sección de un cromosoma, y utilizar esos datos para determinar el estado de ploidía de un individuo. Por ejemplo, se puede concluir que los cromosomas que tiene un número desproporcionadamente grande de lecturas son trisómicos en el feto que se está gestando en la madre a la que se le ha extraído la sangre.

Sin embargo, en realidad el producto inicial de los instrumentos de medición es una señal analógica. Cuando se determina un determinado par de bases a través del software asociado con el software de secuenciación, por ejemplo, el software puede determinar el par de bases T, en realidad la determinación es la que el software considera más probable. Sin embargo, en algunos casos, la determinación puede ser de baja certeza, por ejemplo, la señal analógica puede indicar que el par de bases concreto tiene solo un 90% de probabilidad de ser T y un 10% de probabilidad de ser A. En otro ejemplo, el software de determinación del genotipo que está asociado con un lector del array de SNP puede determinar que un alelo específico es G. Sin embargo, en realidad la señal analógica subyacente puede indicar que tiene solo un 70% de probabilidad de ser G y un 30% de probabilidad de ser T. En estos casos, cuando las aplicaciones de nivel superior utilizan las determinaciones del genotipo y las determinaciones de la secuencia realizadas por el software de nivel inferior, pierden cierta información. Es decir, los datos genéticos primarios, medidos directamente por la plataforma de determinación del genotipo, pueden ser menos fiables que los datos genéticos secundarios determinados por los paquetes de software adjuntos, aunque contienen más información. Para correlacionar las secuencias de datos genéticos secundarios con el genoma, muchas lecturas se descartan porque algunas bases no se leen con suficiente claridad y/o porque la correlación no está clara. Cuando se utilizan lecturas de secuencia de datos genéticos primarios, todas o muchas de esas lecturas que se han descartado cuando se convirtieron por primera vez en lecturas de secuencia de datos genéticos secundarios se pueden utilizar tratando las lecturas de manera probabilística.

En una realización de la presente divulgación, el software de nivel superior no se basa en las determinaciones alélicas, las determinaciones del SNP o las lecturas de secuencia determinadas por el software de nivel inferior.

En vez de esto, el software de nivel superior basa sus cálculos en las señales analógicas medidas directamente por la plataforma de determinación del genotipo. En una realización de la presente divulgación, un método basado en la informática como PARENTAL SUPPORT™ se modifica para que su capacidad para reconstruir datos genéticos del embrión/feto/niño sea organizada para utilizar directamente los datos genéticos primarios medidos por la plataforma de determinación del genotipo. En una realización de la presente divulgación, un método basado en la informática como PARENTAL SUPPORT™ puede realizar determinaciones alélicas y/o determinar el número de copias del cromosoma utilizando datos genéticos primarios y sin usar datos genéticos secundarios. En una realización de la presente divulgación, todas las determinaciones genéticas, determinaciones de SNP, lecturas de secuencia, correlaciones de secuencia se tratan de manera probabilística utilizando los datos de intensidad brutos medidos directamente por la plataforma de determinación del genotipo, en lugar de convertir los datos genéticos primarios en determinaciones genéticas secundarias. En una realización, las mediciones de ADN de la muestra preparada utilizada para calcular las probabilidades del recuento de alelos y determinar la probabilidad relativa de cada hipótesis comprenden los datos genéticos primarios.

En algunas realizaciones, el método puede aumentar la precisión de los datos genéticos de un individuo diana que incorporan los datos genéticos de al menos un individuo relacionado, donde el método comprende la obtención de datos genéticos primarios específicos del genoma de un individuo diana y datos genéticos específicos del genoma o genomas del individuo o individuos relacionados, la creación de un conjunto de una o más hipótesis sobre qué segmentos de qué cromosomas del individuo o individuos relacionados corresponden posiblemente a esos segmentos en el genoma del individuo diana, la determinación de la probabilidad de cada una de las hipótesis teniendo en cuenta los datos genéticos primarios del individuo diana y los datos genéticos del individuo o individuos relacionados, y la utilización de las probabilidades asociadas con cada hipótesis para determinar el estado más probable del material genético real del individuo diana. En algunas realizaciones, el método puede determinar el número de copias de un segmento de un cromosoma del genoma de un individuo diana, donde el método comprende la creación de un conjunto de hipótesis del número de copias sobre cuántas copias del segmento del cromosoma se encuentran presentes en el genoma de un individuo diana, la incorporación de datos genéticos primarios del individuo diana e información genética de uno o más individuos relacionados al conjunto de datos, la estimación de las características de la respuesta de la plataforma asociada con el conjunto de datos, donde la respuesta de la plataforma puede variar de un experimento a otro, el cálculo de las probabilidades condicionales de cada hipótesis del número de copias, teniendo en cuenta el conjunto de datos y las características de la respuesta de la plataforma, y la determinación del número de copias del segmento del cromosoma basándose en la hipótesis del número de copias más probable. En una realización, un método de la presente divulgación puede determinar un estado de ploidía de al menos un cromosoma en un individuo diana, el

método comprende la obtención de datos genéticos primarios del individuo diana y de uno o más individuos relacionados, la creación de un conjunto de al menos una hipótesis del estado de ploidía para cada uno de los cromosomas del individuo diana, la utilización de una o más técnicas especializadas para determinar una probabilidad estadística para cada hipótesis del estado de ploidía del conjunto, para cada técnica especializada utilizada, teniendo en cuenta los datos genéticos obtenidos, la combinación para cada hipótesis del estado de ploidía de las probabilidades estadísticas determinadas mediante la técnica o técnicas especializadas, y la determinación del estado de ploidía para cada uno de los cromosomas del individuo diana basándose en las probabilidades estadísticas combinadas de cada una de las hipótesis del estado de ploidía. En una realización, un método de la presente divulgación puede determinar un estado alélico de un conjunto de alelos en un individuo diana y a partir de uno o los dos progenitores del individuo diana, y opcionalmente de uno o más individuos relacionados, donde el método comprende la obtención de datos genéticos primarios del individuo diana, y de uno o los dos progenitores, y de cualquier individuo relacionado, la creación de un conjunto de al menos una hipótesis alélica para el individuo diana, y para uno o los dos progenitores, y opcionalmente para uno o más individuos relacionados, donde las hipótesis describen los posibles estados alélicos del conjunto de alelos, la determinación de una probabilidad estadística para cada hipótesis alélica del conjunto de hipótesis, teniendo en cuenta los datos genéticos obtenidos, y la determinación del estado alélico para cada uno de los alelos del conjunto de alelos para el individuo diana, y para uno o los dos progenitores, y opcionalmente para uno o más individuos relacionados, basándose en las probabilidades estadísticas de cada una de las hipótesis alélicas.

En algunas realizaciones, los datos genéticos de la muestra mezclada pueden comprender datos de secuencia donde los datos de secuencia pueden no corresponder de forma exclusiva al genoma humano. En algunas realizaciones, los datos genéticos de la muestra mezclada pueden comprender datos de secuencia donde los datos de secuencia corresponden a una pluralidad de ubicaciones en el genoma, donde cada posible correlación está asociada con una probabilidad de que la correlación en cuestión es correcta. En algunas realizaciones, no se asume que las lecturas de secuencia están asociadas con una posición concreta en el genoma. En algunas realizaciones, las lecturas de secuencia están asociadas con una pluralidad de posiciones en el genoma y una probabilidad asociada perteneciente a esa posición.

Método de recuento para determinar el número de copias de cromosoma

En un aspecto, la divulgación presenta métodos de realizar pruebas para determinar la distribución anómala de un cromosoma fetal comparando el número de etiquetas de secuencia que se alinean con diferentes cromosomas (véase, por ejemplo, la Pat. USA número 8.296.076, presentada el 20 de abril de 2012). Como se conoce en la técnica, el término "etiqueta de secuencia" se refiere a una secuencia de ácido nucleico relativamente corta (por ejemplo, 15-100) que se puede utilizar para identificar una determinada secuencia más larga, por ejemplo, que esté correlacionada con un cromosoma o gen o región genómica. En algunas realizaciones, el método consiste en (i) poner en contacto una muestra que incluye una mezcla de ADN materno y fetal con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana diferentes para producir una mezcla de reacción; en la que los loci diana son de una pluralidad de cromosomas diferentes; y en la que la pluralidad de cromosomas diferentes incluye al menos un primer cromosoma sospechoso de tener una distribución anómala en la muestra y al menos un segundo cromosoma que se supone que está distribuido normalmente en la muestra; (ii) someter la mezcla de reacción a las condiciones de reacción de extensión del cebador para producir productos amplificados; (iii) secuenciar los productos amplificados para obtener una pluralidad de etiquetas de secuencia que se alinean con los loci diana; en donde las etiquetas de secuencia tienen la longitud suficiente para ser asignadas a un locus diana específico; (iv) asignar en un ordenador la pluralidad de etiquetas de secuencia a sus loci diana correspondientes; (v) determinar en un ordenador un número de etiquetas de secuencia que se alinean con los loci diana del primer cromosoma y un número de etiquetas de secuencia que se alinean con los loci diana del segundo cromosoma; y (vi) comparar los números del paso (v) para determinar la presencia o ausencia de una distribución anómala del primer cromosoma.

En un aspecto, la divulgación proporciona métodos para detectar la presencia o ausencia de una aneuploidía fetal comparando la frecuencia relativa de amplicones diana entre cromosomas (véase, por ejemplo, la publ. PCT número WO 2012/103031, presentada el 23 de enero de 2012). En algunas realizaciones, el método consiste en (i) poner en contacto una muestra que incluye una mezcla de ADN materno y fetal con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci diana no polimórficos diferentes para producir una mezcla de reacción; en la que los loci diana son de una pluralidad de cromosomas diferentes; (ii) someter la mezcla de reacción a las condiciones de reacción de extensión del cebador para producir productos amplificados que incluyen amplicones diana; (iii) cuantificar en un ordenador una frecuencia relativa de los amplicones diana del primer y segundo cromosomas de interés; (iv) comparar en un ordenador la frecuencia relativa de los amplicones diana del primer y segundo cromosomas de interés; e (v) identificar la presencia o ausencia de una aneuploidía en base a las frecuencias relativas comparadas del primer y segundo cromosoma de interés. En algunas realizaciones, el primer cromosoma es un cromosoma sospechoso de ser euploide. En algunas realizaciones, el segundo cromosoma es un cromosoma sospechoso de ser aneuploidía.

Combinación de métodos para el diagnóstico prenatal

Hay muchos métodos que se pueden utilizar para el diagnóstico prenatal y la detección prenatal de la aneuploidía u otros defectos genéticos. En otra parte de este documento, y en la Solicitud de patente USA con el número de serie 11/603.406, presentada el 28 de noviembre de 2006; la Solicitud de patente USA con el número de serie 12/076.348,

presentada el 17 de marzo de 2008, y la Solicitud PCT con el número de serie PCT/S09/52730 se describe un método que utiliza los datos genéticos de individuos relacionados para aumentar la precisión con la que se conocen o estiman los datos genéticos de un individuo diana. Otros métodos utilizados para el diagnóstico prenatal implican la medición de los niveles de determinadas hormonas en la sangre materna, donde estas hormonas están correlacionadas con diversas anomalías genéticas. Un ejemplo de esto se denomina la prueba triple, una prueba en la que se miden los niveles de varias (normalmente dos, tres, cuatro o cinco) hormonas diferentes en la sangre materna. En un caso en el que se utilizan múltiples métodos para determinar la probabilidad de un determinado resultado, donde ninguno de los métodos es definitivo por sí solo, se puede combinar la información proporcionada por estos métodos para realizar una predicción que es más precisa que cualquiera de los métodos individuales. En la prueba triple, la combinación de la información proporcionada por las tres hormonas diferentes puede proporcionar una predicción de anomalías genéticas que es más precisa que la efectuada por los niveles de hormonas individuales.

En el presente documento se describe un método para hacer predicciones más precisas sobre el estado genético de un feto, específicamente la posibilidad de anomalías genéticas en un feto que comprende la combinación de predicciones de anomalías genéticas en un feto donde esas predicciones se han realizado utilizando una variedad de métodos. Un método "más preciso" se puede referir a un método para el diagnóstico de una anomalía que tiene una tasa de falsos negativos inferior para una determinada tasa de falsos positivos. En una realización preferida de la presente divulgación, una o más de las predicciones se realizan en base a los datos genéticos conocidos sobre el feto, donde el conocimiento genético ha sido determinado utilizando el método PARENTAL SUPPORT™, es decir, utilizando datos genéticos del individuo relacionado con el feto para determinar los datos genéticos del feto con mayor precisión. En algunas realizaciones, los datos genéticos pueden incluir estados de ploidía del feto. En algunas realizaciones, los datos genéticos se pueden referir a un conjunto de determinaciones alélicas sobre el genoma del feto. En algunas realizaciones, algunas de las predicciones se pueden haber realizado utilizando la prueba triple. En algunas de las realizaciones, algunas de las predicciones se pueden haber realizado utilizando mediciones de otros niveles de hormonas en la sangre materna. En algunas realizaciones, las predicciones realizadas mediante métodos considerados diagnósticos se pueden combinar con predicciones realizadas mediante métodos considerados de detección. En algunas realizaciones, el método implica la medición de los niveles en sangre materna de alfa-fetoproteína (AFP). En algunas realizaciones, el método implica la medición de los niveles de sangre materna de estriol no conjugado (UE3). En algunas realizaciones, el método implica la medición de los niveles en sangre materna de gonadotropina coriónica humana beta (beta-hCG). En algunas realizaciones, el método implica la medición de los niveles en sangre materna de antígeno trofoblástico invasivo (ITA). En algunas realizaciones, el método implica la medición de los niveles en sangre materna de inhibina. En algunas realizaciones, el método implica la medición de los niveles en sangre materna de proteína A plasmática asociada al embarazo (PAPP-A). En algunas realizaciones, el método implica la medición de los niveles en sangre materna de otras hormonas o marcadores del suero materno. En algunas de las realizaciones, algunas de las predicciones se pueden haber realizado utilizando otros métodos. En algunas realizaciones, algunas de las predicciones se pueden haber realizado utilizando una prueba plenamente integrada, como una que combina los ultrasonidos y el análisis de sangre a las 12 semanas aproximadas de embarazo y un segundo análisis de sangre sobre las 16 semanas de embarazo. En algunas realizaciones, el método implica la medición de la translucencia nual (NT) fetal. En algunas realizaciones, el método implica el uso de los niveles medidos de las mencionadas hormonas para realizar predicciones. En algunas realizaciones, el método implica una combinación de los métodos mencionados anteriormente.

Existen muchas formas de combinar las predicciones, por ejemplo, se podrían convertir las mediciones de las hormonas en un múltiplo de la mediana (MoM) y a continuación en ratios de probabilidad (LR). De forma similar, otras mediciones se podrían transformar en LR utilizando el modelo de la mezcla de las distribuciones de NT. Los LR para NET y los marcadores bioquímicos se podrían multiplicar por la edad y el riesgo relacionado con la gestación para obtener el riesgo para las diversas condiciones, como la trisomía 21. Las tasas de detección (DR) y las tasas de falsos positivos (FPR) se podrían calcular tomando las proporciones con los riesgos anteriores que superan un determinado umbral de riesgo.

En una realización, un método para determinar el estado de ploidía implica la combinación de las probabilidades relativas de cada una de las hipótesis de ploidía determinadas utilizando el modelo de distribución conjunto y las probabilidades del recuento de alelos con las probabilidades relativas de cada una de las hipótesis de ploidía que se calculan utilizando técnicas estadísticas tomadas de otros métodos que determinan una puntuación de riesgo para un feto que es trisómico, incluyendo, entre otros: un análisis del recuento de lecturas, la comparación de las tasas de heterocigosidad, una estadística que solo está disponible cuando se utiliza la información genética parental, la probabilidad de señales normalizadas del genotipo para determinados contextos parentales, una estadística que se calcula utilizando la fracción fetal estimada de la primera muestra o de la muestra preparada y combinaciones de estas.

Otro método podría implicar una situación con cuatro niveles de hormonas medidos, donde la distribución de la probabilidad con respecto a esas hormonas es conocida: $p(x_1, X_2, X_3, X_4|e)$ para el caso euploide y $p(x_1, x_2, x_3, x_4|a)$ para el caso aneuploide. A continuación, se podría medir la distribución de la probabilidad para las mediciones del ADN, $g(y|e)$ y $g(y|a)$ para el caso euploide y el caso aneuploide, respectivamente. Asumiendo que son independientes dado el supuesto de euploide/aneuploide, se podría combinar como $p(x_1, X_2, X_3, X_4|a)g(y|a)$ y $p(x_1, X_2, X_3, X_4|e)g(y|e)$ y, a continuación, multiplicar cada uno por el anterior $p(a)$ y $p(e)$ dada la edad materna. Se podría elegir el más elevado.

En una realización, es posible evocar el teorema de límite central para asumir que la distribución en $g(y|a)$ o e es gaussiana, y medir la desviación estándar y media analizando múltiples muestras. En otra realización, se puede asumir

que no son independientes dado el resultado y recoger suficientes muestras para estimar la distribución conjunta $p(x_1, x_2, x_3, x_4|a \text{ o } e)$.

En una realización, el estado de ploidía para el individuo diana está determinado que sea el estado de ploidía que está asociado con la hipótesis cuya probabilidad es mayor. En algunos casos, una hipótesis tendrá una probabilidad normalizada combinada superior al 90%. Cada hipótesis está asociada con uno o varios estados de ploidía, y el estado de ploidía asociado con la hipótesis cuya probabilidad normalizada combinada sea superior al 90%, o algún otro valor umbral, como el 50%, 80%, 95%, 98%, 99% o 99,9% podrá ser seleccionado como el umbral requerido para establecer que una hipótesis es el estado de ploidía determinado.

ADN de hijos de embarazos anteriores en la sangre materna

Una dificultad del diagnóstico prenatal no invasivo consiste en diferenciar las células fetales del embarazo actual de las células fetales de embarazos anteriores. Algunos piensan que el material genético de los embarazos anteriores desaparecerá tras algún tiempo, pero no se han obtenido pruebas concluyentes. En una realización de la presente divulgación, se puede determinar el ADN fetal presente en la sangre materna de origen paterno (es decir, ADN del feto heredado del padre) utilizando el método PARENTAL SUPPORT™ (PS) y el conocimiento del genoma paterno. Este método puede utilizar información genética parental por fases. Es posible obtener por fases el genotipo parental de la información genotípica no determinada por fases utilizando datos genéticos del abuelo (como datos genéticos medidos a partir del esperma del abuelo), o datos genéticos de otros niños nacidos, o una muestra de un aborto. También se podría obtener información genética por fases a partir de la información genética no determinada por fases mediante HapMap o determinando el haplotipo de las células paternas. La determinación con éxito del haplotipo se ha demostrado deteniendo las células en la fase de mitosis cuando los cromosomas son agrupaciones apretadas y utilizando microfluidicos para poner los cromosomas separados en pocillos separados. En una realización se pueden utilizar los datos del haplotipo parental por fases para detectar la presencia de más de un homólogo del padre, lo que implica que el material genético de más de un niño se encuentra presente en la sangre. Al centrarse en los cromosomas que se espera que sean euploides en un feto, se podría descartar la posibilidad de que el feto se vea afectado por una trisomía. También se puede determinar si el ADN fetal no procede del padre actual, en cuyo caso se podrían utilizar otros métodos como la prueba triple para predecir anomalías genéticas.

Puede haber otras fuentes de material genético fetal disponibles a través de métodos distintos de la extracción de sangre. En el caso del material genético fetal disponible en la sangre materna, existen dos categorías principales: (1) células fetales completas, por ejemplo, glóbulos rojos fetales nucleados o eritroblastos, y (2) ADN fetal flotante libre. En el caso de las células fetales completas, existen algunas evidencias de que las células fetales pueden persistir en la sangre materna durante un periodo de tiempo prolongado de forma que se puede aislar una célula de una mujer embarazada que contenga ADN de un niño o un feto de un embarazo anterior. También existen evidencias de que el ADN fetal flotante se elimina del sistema en unas semanas. Un problema consiste en determinar la identidad del individuo cuyo material genético se encuentra en la célula, concretamente para garantizar que el material genético medido no corresponda a un feto de un embarazo anterior. En una realización de la presente divulgación, el conocimiento del material genético materno se puede utilizar para garantizar que el material genético en cuestión no es material genético materno. Existen diversos métodos para ello, incluyendo el método basado en la informática PARENTAL SUPPORT™, descrito en este documento o en cualquiera de las patentes mencionadas en el mismo.

En una realización de la presente divulgación, la sangre extraída de la madre embarazada se puede separar en una fracción que comprende ADN fetal flotante libre, y una fracción que comprende glóbulos rojos nucleados. El ADN flotante libre se puede opcionalmente enriquecer, y la información genotípica del ADN se puede medir. A partir de la información genotípica medida del ADN flotante libre, el conocimiento del genotipo materno se puede utilizar para determinar aspectos del genotipo fetal. Estos aspectos se pueden referir al estado de ploidía y/o a las identidades de un conjunto de alelos. A continuación, se puede determinar el genotipo de los glóbulos rojos nucleados utilizando métodos descritos en otro apartado de este documento y en otras patentes mencionadas, especialmente las mencionadas en la primera parte de este documento. El conocimiento del genoma materno permitiría determinar si un glóbulo rojo dado es genéticamente materno o no. Y los aspectos del genotipo fetal determinados como se ha descrito anteriormente permitirían determinar si la célula sanguínea individual procede genéticamente del feto que se está gestando en estos momentos. Básicamente, este aspecto de la presente divulgación permite utilizar el conocimiento genético de la madre y posiblemente la información genética de otros individuos relacionados, como el padre, junto con la información genética medida del ADN flotante libre que se encuentra en la sangre materna para determinar si una célula nucleada aislada que se encuentra en la sangre materna es a) genéticamente materna, b) genéticamente del feto que se está gestando en estos momentos, o c) genéticamente de un feto de un embarazo anterior.

Determinación de aneuploidía del cromosoma sexual prenatal

En métodos conocidos en la técnica, las personas que desean determinar el sexo de un feto en gestación a partir de la sangre materna han utilizado el hecho de que el ADN flotante libre fetal (ffDNA) se encuentra presente en el plasma de la madre. Si se pueden detectar los loci específicos de Y en el plasma materno, esto implica que el feto en gestación es un varón. Sin embargo, la falta de detección de loci específicos de Y en el plasma no siempre garantiza que el feto en gestación es femenino cuando se utilizan los métodos conocidos en la técnica, ya que en algunos casos la cantidad de ffDNA es demasiado baja para garantizar que los loci específicos de Y se detectarían en el caso de un feto varón.

En el presente documento se proporciona un método novedoso que no requiere la medición de ácidos nucleicos específicos de Y, es decir ADN procedente de loci de origen exclusivamente paterno. El método Parental Support, anteriormente divulgado, utiliza datos de frecuencia de cruces, datos genotípicos parentales y técnicas informáticas para determinar el estado de ploidía de un feto en gestación. El sexo del feto es simplemente el estado de ploidía del feto en los cromosomas sexuales. Un feto que es XX es mujer y XY es varón. El método descrito en el presente documento también se puede utilizar para determinar el estado de ploidía del feto. Cabe señalar que la determinación del sexo es efectivamente sinónimo de la determinación del estado de ploidía de los cromosomas sexuales; en el caso de la determinación del sexo, normalmente se presupone que el niño es euploide y, por tanto, hay menos hipótesis posibles.

El método divulgado en el presente documento implica analizar los loci que son comunes a los cromosomas X e Y para crear una línea de base en términos de cantidad prevista de ADN fetal presente para un feto. A continuación, las regiones que son específicas solo para el cromosoma X se pueden analizar para determinar si el feto es mujer o varón. En el caso de un varón, esperamos observar menos ADN fetal de loci que son específicos para el cromosoma X que de loci que son específicos tanto para el cromosoma X como para el cromosoma Y. Por el contrario, en los fetos de mujeres, esperamos que el ADN para cada uno de estos grupos sea el mismo. El ADN en cuestión se puede medir a través de cualquier técnica que pueda cuantificar la cantidad de ADN presente en una muestra, por ejemplo, qPCR, arrays de SNP, arrays para la determinación del genotipo o secuenciación. Para el ADN que es exclusivamente de un individuo esperaríamos observar lo siguiente:

	ADN específico de X	ADN específico de X e Y	ADN específico de Y
Varón (XY)	A	2A	A
Mujer (XX)	2A	2A	0

En el caso del ADN de un feto mezclado con ADN de la madre y donde la fracción de ADN fetal de la mezcla es F y donde la fracción de ADN materno de la mezcla es M, de forma que $F+M = 100\%$, esperaríamos observar lo siguiente:

	ADN específico de X	ADN específico de X e Y	ADN específico de Y
Feto varón (XY)	$M + \frac{1}{2} F$	$M + F$	$\frac{1}{2} F$
Feto mujer (XX)	$M + F$	$M + F$	0

En el caso en el que F y M son conocidos, los ratios previstos se pueden calcular y los datos observados se pueden comparar con los datos previstos. En el caso en el que F y M no son conocidos, se puede seleccionar un umbral basado en datos históricos. En ambos casos, la cantidad de ADN medida en loci específicos tanto para X como para Y se pueden utilizar como línea de base, y la prueba para determinar el sexo del feto se puede basar en la cantidad de ADN observada en loci específicos exclusivamente para el cromosoma X. Si esa cantidad es inferior que la línea de base en una cantidad aproximadamente igual a $\frac{1}{2} F$ o en una cantidad que hace que caiga por debajo de un umbral predefinido, se determina que el feto es varón, y si esa cantidad es aproximadamente igual a la línea de base o si no es inferior en una cantidad que hace que caiga por debajo de un umbral predefinido, se determina que el feto es mujer.

En otra realización, se puede analizar sólo aquellos loci que son comunes tanto a los cromosomas X como a los Y, a menudo denominado el cromosoma Z. Típicamente un subconjunto de los loci del cromosoma Z son siempre A en el cromosoma X y B en el cromosoma Y. Si se descubre que los SNP del cromosoma Z tienen el genotipo B, entonces se dirá que el feto es un varón; si se descubre que los SNP del cromosoma Z solo tienen el genotipo A, se dirá que el feto es una mujer. En otra realización, se pueden analizar los loci que se encuentran únicamente en el cromosoma X. Contextos como AA|B resultan particularmente informativos, dado que la presencia de una B indica que el feto tiene un cromosoma X del padre. Contextos como AB|B también son informativos, dado que esperamos ver B presente únicamente la mitad de las veces en el caso de un feto de mujer en comparación con el feto de un varón. En otra realización, se pueden analizar los SNP del cromosoma Z donde tanto los alelos A como B se encuentran presentes en el cromosoma X y el cromosoma Y, y donde se sabe qué SNP proceden del cromosoma Y paterno y cuáles proceden del cromosoma X paterno.

En una realización, es posible amplificar las posiciones de un solo nucleótido que se conoce que varían entre la región homóloga que no se vuelve a combinar (HNR) compartida por el cromosoma Y y el cromosoma X. La secuencia de la región HNR es en gran medida idéntica en los cromosomas X e Y. Dentro de esta región idéntica están posiciones de un solo nucleótido que, aunque no varían entre los cromosomas X y entre los cromosomas Y de la población, son diferentes entre los cromosomas X e Y. Cada ensayo por PCR podría amplificar una secuencia de loci que se encuentran presentes tanto en los cromosomas X como Y. Dentro de cada secuencia amplificada estaría una base individual que se puede detectar utilizando la secuenciación o cualquier otro método.

En una realización, el sexo del feto se podría determinar a partir del ADN flotante libre fetal que se encuentra en el plasma materno, el método comprende algunos o todos los pasos siguientes: 1) Diseñar cebadores para una PCR (ordinaria o mini-PCR, más multiplexado si se desea) para amplificar las posiciones de un único nucleótido variable X/Y dentro de la región HNR, 2) obtener plasma materno, 3) amplificar por PCR diana del plasma materno utilizando ensayos de PCR para X/Y de HNR, 4) secuenciar los amplicones, 5) examinar los datos de secuencia para determinar la presencia del alelo Y en una o más de las secuencias amplificadas. La presencia de uno o más indicaría un feto varón. La ausencia de todos los alelos Y de los amplicones indica un feto mujer.

En una realización, se puede utilizar la secuenciación focalizada para medir el ADN en el plasma materno y/o los genotipos parentales. En una realización se podrían ignorar todas las secuencias que proceden claramente de ADN de origen paterno. Por ejemplo, en el contexto AA|AB, se podría recontar el número de secuencias A e ignorar todas las secuencias B. Para determinar la tasa de heterocigosidad del anterior algoritmo, se podría comparar el número de secuencias A observadas con el número previsto de secuencias totales para la sonda en cuestión. Hay muchas formas de calcular un número previsto de secuencias para cada sonda y para cada muestra. En una realización, es posible utilizar datos históricos para determinar qué fracción de todas las lecturas de secuencia pertenece a cada sonda específica y, a continuación, utilizar esta fracción empírica, combinada con el número total de lecturas de secuencia, para estimar el número de secuencias en cada sonda. Otro planteamiento sería focalizar algunos alelos homocigotos conocidos y, a continuación, utilizar los datos históricos relativos al número de lecturas en cada sonda con el número de lecturas en los alelos homocigotos conocidos. A continuación, para cada muestra se podría medir el número de lecturas en los alelos homocigotos y después utilizar esta medición, junto con las relaciones obtenidas empíricamente, para estimar el número de lecturas de secuencia en cada sonda.

En algunas realizaciones, es posible determinar el sexo del feto combinando las predicciones realizadas mediante una pluralidad de métodos. En algunas realizaciones, la pluralidad de métodos se selecciona de los métodos descritos en esta divulgación. En algunas realizaciones, al menos uno de la pluralidad de métodos se selecciona de los métodos descritos en esta divulgación.

En algunas realizaciones, el método descrito en el presente documento también se puede utilizar para determinar el estado de ploidía del feto en gestación. En una realización, el método de determinación del estado de ploidía utiliza loci que son específicos para el cromosoma X, o comunes para el cromosoma X y el cromosoma Y, pero no emplea los loci específicos de Y. En una realización, el método de determinación del estado de ploidía utiliza uno o más de los siguientes loci: loci que son específicos del cromosoma X, loci que son comunes para el cromosoma X y el cromosoma Y, y loci que son específicos para el cromosoma Y. En una realización, donde los ratios de cromosomas sexuales son similares, por ejemplo, 45,X (síndrome de Turner), 46,XX (mujer normal) y 47,XXX (trisomía X), la diferenciación se puede realizar comparando las distribuciones alélicas con las distribuciones alélicas previstas en función de las diversas hipótesis. En otra realización, esto se puede conseguir comparando el número relativo de lecturas de secuencia para los cromosomas sexuales con uno o una pluralidad de cromosomas de referencia que se presuponen euploides. Cabe señalar también que estos métodos se pueden ampliar para incluir los casos aneuploides.

Detección de enfermedades en un único gen

En una realización, un método para determinar el estado de ploidía del feto se puede ampliar para permitir la realización de pruebas simultáneas para trastornos en un único gen. El diagnóstico de enfermedades en un único gen utiliza el mismo planteamiento focalizado que se emplea para las pruebas de la aneuploidía y requiere dianas específicas adicionales. En una realización, el diagnóstico de NPD en un único gen se realiza a través del análisis del enlace. En muchos casos, las pruebas directas de la muestra de ADN libre de células no resultan fiables, dado que la presencia de ADN materno hace que resulte prácticamente imposible determinar si el feto ha heredado la mutación de la madre. La detección de un alelo único de origen paterno resulta menos problemática, pero solamente resulta plenamente informativa si la enfermedad es dominante y portada por el padre, lo que limita la utilidad del planteamiento. En una realización el método implica la PCR o planteamientos de amplificación relacionados.

En algunas realizaciones, el método implica el ajuste por fases de los alelos anómalos con los SNPs circundantes enlazados muy estrechamente en los progenitores utilizando información de familiares en primer grado. A continuación se puede ejecutar el método Parental Support en los datos de secuenciación focalizados obtenidos de estos SNP para determinar qué homólogos, normales o anómalos, han sido heredados por el feto de ambos progenitores. Siempre que los SNP están suficientemente unidos, la herencia del genotipo del feto se puede determinar de manera muy fiable. En algunas realizaciones, el método comprende a) la adición de un conjunto de loci de SNP para flanquear de forma densa un conjunto especificado de enfermedades comunes a nuestro grupo multiplexado para la prueba de la aneuploidía; b) determinación por fases de manera fiable de los alelos de estos SNP añadidos con los alelos normales y anómalos basados en los datos genéticos de diversos parientes; y (c) reconstrucción del haplotipo fetal, o del conjunto de alelos de SNP por fases en los homólogos paternos y maternos heredados en la región que rodea al locus de la enfermedad para determinar el genotipo fetal. En algunas realizaciones, las sondas adicionales que están estrechamente unidas a un locus vinculado a una enfermedad se añaden al conjunto de loci polimórficos que se van a utilizar para la prueba de la aneuploidía.

La reconstrucción del diplotipo fetal resulta difícil porque la muestra es una mezcla de ADN materno y fetal. En algunas realizaciones, el método incorpora información relativa para determinar las fases de los SNP y los alelos de la enfermedad, a continuación tiene en cuenta la distancia física de los SNP y los datos de recombinación de las

probabilidades de recombinación específicas de una ubicación y los datos observados de las mediciones genéticas del plasma materno para obtener el genotipo más probable del feto.

En una realización, se incluyen un número de sondas adicionales por locus vinculado a una enfermedad en el conjunto de loci polimórficos focalizados; el número de sondas adicionales por locus vinculada a la enfermedad puede ser de entre 4 y 10, entre 11 y 20, entre 21 y 40, entre 41 y 60, entre 61 y 80, o combinaciones de estos.

El ajuste por fases de los datos diploides de los progenitores puede ser un reto, y hay varias maneras de que esto se pueda conseguir. Algunas se describen en esta divulgación, otras se describen más detalladamente en otras divulgaciones (véase, por ejemplo, la publ. PCT número W02009105531, presentada el 9 de febrero de 2009, y la publ. PCT número W02010017214, presentada el 4 de agosto de 2009). En una realización, un progenitor se puede determinar por fases por inferencia midiendo el tejido del progenitor que es haploide, por ejemplo, midiendo uno o más espermatozoides u óvulos. En una realización el progenitor se puede determinar por fases por inferencia utilizando los datos genotípicos medidos de un familiar de primer grado como los padres o hermanos del progenitor. En una realización, el progenitor se puede determinar por fases por dilución donde el ADN se diluye, en uno o una pluralidad de pocillos, hasta el punto donde se espera que no haya más de aproximadamente una copia de cada haplotipo en cada pocillo y, a continuación, se mide el ADN en uno o más pocillos. En una realización, el genotipo del progenitor se puede determinar por fases utilizando programas informáticos que utilizan frecuencias de haplotipo basadas en la población para inferir la fase más probable. En una realización, el progenitor se puede determinar por fases si los datos haplotípicos por fases son conocidos para el otro progenitor, junto con los datos genéticos no determinados por fases de una o más descendencia genética de los progenitores. En algunas realizaciones, la descendencia genética de los progenitores puede ser uno o más embriones, fetos, y/o niños nacidos. Algunos de estos métodos y otros métodos para la determinación por fases de uno o ambos progenitores se divulgan más detalladamente, por ejemplo, en la Publ. USA número 2011/0033862, presentada el 19 de agosto de 2010; la Publ. USA número 2011/0178719, presentada el 3 de febrero de 2011; la Publ. USA número 2007/0184467, presentada el 22 de noviembre de 2006; la Publ. USA número 2008/0243398, presentada el 17 de marzo de 2008.

Reconstrucción del genoma fetal

En un aspecto, la divulgación presenta métodos para determinar un haplotipo de un feto. En varias realizaciones, este método permite determinar qué loci polimórficos (como SNPs) ha heredado el feto y reconstruir qué homólogos (incluyendo los eventos de recombinación) están presentes en el feto (e interpolar así la secuencia entre los loci polimórficos). Si se desea, se puede reconstruir básicamente todo el genoma del feto. Si queda alguna ambigüedad restante en el genoma del feto (como en intervalos con un cruce), esta ambigüedad se puede minimizar si se desea analizando loci polimórficos adicionales. En distintas realizaciones, los loci polimórficos se eligen para cubrir uno o más de los cromosomas a una densidad para reducir cualquier ambigüedad a un nivel deseado. Este método tiene importantes aplicaciones para la detección de polimorfismos u otras mutaciones de interés en un feto ya que permite su detección en base a un enlace (como la presencia de loci polimórficos enlazados en el genoma fetal) en lugar de dirigir la detección del polimorfismo u otra mutación de interés en el genoma fetal. Por ejemplo, si un progenitor es un portador para una mutación asociada a fibrosis quística (CF), se puede analizar una muestra de ácido nucleico que incluye ADN materno de la madre del feto y ADN fetal del feto para determinar si el ADN fetal incluye el haplotipo que contiene la mutación CF. En concreto, los loci polimórficos se pueden analizar para determinar si el ADN fetal incluye el haplotipo que contiene la mutación CF sin tener que detectar la propia mutación CF en el ADN fetal.

En algunas realizaciones, el método implica la determinación de un haplotipo parental (por ejemplo, un haplotipo de la madre o padre del feto). En algunas realizaciones, esta determinación se realiza sin utilizar datos de un familiar de la madre o padre. En algunas realizaciones, un haplotipo parental se determina utilizando un planteamiento de dilución seguido de una determinación del genotipo de SNP o secuenciación como se describe en el presente documento y en otra parte (véase, por ejemplo, la Publ. USA número 2011/0033862, presentada el 19 de agosto de 2010). Puesto que el ADN está diluido, es poco probable que más de un haplotipo esté en la misma fracción (o tubo). Por tanto, puede haber efectivamente una única molécula de ADN en el tubo, lo que permite determinar el haplotipo en una única molécula de ADN. En algunas realizaciones, el método incluye dividir una muestra de ADN en una pluralidad de fracciones de tal modo que al menos una de las fracciones incluye un cromosoma o un segmento de cromosoma de una par de los cromosomas, y una determinación del genotipo (por ejemplo, la determinación de la presencia de dos o más loci polimórficos) de la muestra de ADN en al menos una de las fracciones, determinando así un haplotipo parental. En algunas realizaciones, la determinación del genotipo implica una secuenciación (como, por ejemplo, la secuenciación por fuerza bruta). En algunas realizaciones, la determinación del genotipo consiste en utilizar un array del SNP para detectar loci polimórficos, por ejemplo, al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes. En algunas realizaciones, la determinación del genotipo implica el uso de PCR multiplexada. En algunas realizaciones, el método consiste en poner en contacto la muestra en una fracción con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes (como SNPs) para producir una mezcla de reacción; y someter la mezcla de reacción a las condiciones de reacción de extensión del cebador para producir productos amplificados que se miden con un secuenciador de alto rendimiento para producir datos de secuenciación.

En algunas realizaciones, un haplotipo de la madre se determina mediante cualquiera de los métodos que se describen en el presente documento utilizando datos de un familiar de la madre. En algunas realizaciones, un haplotipo del padre

se determina mediante cualquiera de los métodos que se describen en el presente documento utilizando datos de un familiar del padre. En algunas realizaciones, un haplotipo se determina tanto para el padre como para la madre. En algunas realizaciones, se utiliza un array de SNP para determinar la presencia de al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes en una muestra de ADN de la madre (o padre) y un familiar de la madre (o padre). En algunas realizaciones, el método consiste en poner en contacto una muestra de ADN de la madre (o padre) y/o un familiar de la madre (o padre) con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes (como SNPs) para producir una mezcla de reacción; y someter la mezcla de reacción a las condiciones de reacción de extensión del cebador para producir productos amplificados que se miden con un secuenciador de alto rendimiento para producir datos de secuenciación. El haplotipo parental se puede determinar en base al array de SNP o a los datos de secuenciación. En algunas realizaciones, los datos parentales se pueden determinar por fases mediante métodos descritos o a los que se hace referencia en otra parte de este documento.

Estos datos de haplotipo parental se pueden utilizar para determinar si el feto ha heredado el haplotipo parental. En algunas realizaciones, se analiza una muestra de ácido nucleico que incluye ADN materno de la madre del feto y ADN fetal utilizando un array de SNP para detectar al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes. En algunas realizaciones, se analiza una muestra de ácido nucleico que incluye ADN materno de la madre del feto y ADN fetal poniendo en contacto la muestra con una biblioteca de cebadores que se hibridan simultáneamente a al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes (como SNPs) para producir una mezcla de reacción. En algunas realizaciones, la mezcla de reacción se somete a condiciones de reacción de extensión del cebador para producir productos amplificados. En algunas realizaciones, los productos amplificados se miden con un secuenciador de alto rendimiento para producir datos de secuenciación. En varias realizaciones, el array de SNP o los datos de secuenciación se utilizan para determinar un haplotipo parental utilizando datos sobre la probabilidad de que los cromosomas se crucen en diferentes ubicaciones en un cromosoma (como, por ejemplo, utilizando datos de recombinación como los que se pueden encontrar en la base de datos HapMap para crear una puntuación de riesgo de recombinación para cualquier intervalo) para modelar la dependencia entre alelos polimórficos en el cromosoma. En algunas realizaciones, los recuentos de alelos en los loci polimórficos se calculan en un ordenador en base a los datos de secuenciación. En algunas realizaciones, en un ordenador se crean una pluralidad de hipótesis de ploidía cada una de ellas perteneciente a un estado de ploidía posible diferente del cromosoma; un modelo (como, por ejemplo, un modelo de distribución conjunto) para los recuentos de alelos esperados en los loci polimórficos en el cromosoma se genera en un ordenador para cada hipótesis de ploidía; una probabilidad relativa de cada hipótesis de ploidía se determina en un ordenador utilizando el modelo de distribución conjunto y los recuentos de alelos; y el estado de ploidía del feto se determina seleccionando el estado de ploidía correspondiente a la hipótesis con la mayor probabilidad. En algunas realizaciones, la creación de un modelo de distribución conjunto para los recuentos de alelos y el paso de determinación de la probabilidad relativa de cada hipótesis se realiza utilizando un método que no requiere el uso de un cromosoma de referencia.

En algunas realizaciones, un haplotipo fetal se determina para uno o más cromosomas tomados del grupo compuesto por los cromosomas 13, 18, 21, X, e Y. En algunas realizaciones, un haplotipo fetal se determina para todos los cromosomas fetales. En varias realizaciones, el método determina básicamente todo el genoma del feto. En algunas realizaciones, el haplotipo se determina para al menos un 30, 40, 50, 60, 70, 80, 90, o un 95% del genoma del feto. En algunas realizaciones, la determinación del haplotipo del feto incluye información sobre qué alelo está presente para al menos 1.000; 2.000; 5.000; 7.500; 10.000; 20.000; 25.000; 30.000; 40.000; 50.000; 75.000; o 100.000 loci polimórficos diferentes.

Composiciones de ADN

Cuando se realiza un análisis informático en datos de secuenciación medidos en una mezcla de sangre fetal y materna para determinar información genómica que pertenece al feto, por ejemplo, el estado de ploidía del feto, puede ser beneficioso medir las distribuciones alélicas en un conjunto de alelos. Lamentablemente, en muchos casos, por ejemplo, cuando se intenta determinar el estado de ploidía de un feto a partir de la mezcla de ADN que se encuentra en el plasma de una muestra de sangre materna, la cantidad de ADN disponible no es suficiente como para medir directamente las distribuciones alélicas de una forma fiable en la muestra. En estos casos, la amplificación de la mezcla de ADN proporcionará un número suficiente de moléculas de ADN como para medir de forma fiable las distribuciones alélicas deseadas. Sin embargo, los métodos actuales de amplificación que se utilizan típicamente en la amplificación de ADN para la secuenciación a menudo presentan sesgos importantes, lo que significa que no amplifican los dos alelos de un locus polimórfico de forma homogénea. Una amplificación sesgada puede producir distribuciones alélicas bastante diferentes de las presentes en la mezcla original. Para la mayoría de las aplicaciones, no se necesitan unas mediciones altamente precisas de las cantidades relativas de alelos presentes en los loci polimórficos. Por el contrario, en una realización de la presente divulgación, la amplificación o los métodos de enriquecimiento que enriquecen específicamente alelos polimórficos y preservan ratios alélicos es beneficiosa.

En el presente documento se describen diversos métodos que se pueden utilizar para enriquecer preferentemente una muestra de ADN en una pluralidad de loci de forma que se minimice el sesgo alélico. Algunos ejemplos utilizan sondas circularizantes para focalizar una pluralidad de loci donde los extremos 3' y los extremos 5' de la sonda precircularizada están diseñados para hibridarse a bases que están a una o unas cuantas posiciones de distancia de los puntos

polimórficos del alelo focalizado. Otro ejemplo consiste en utilizar sondas para la PCR donde la sonda para la PCR del extremo 3' está diseñada para hibridarse a bases que se encuentran a una o unas cuantas posiciones de distancia de los puntos polimórficos del alelo focalizado. Otro ejemplo consiste en utilizar un planteamiento de división y reagrupación para crear mezclas de ADN en las que los loci enriquecidos preferentemente son enriquecidos con un escaso sesgo alélico sin las desventajas del multiplexado directo. Otro ejemplo es el uso de un planteamiento de captura híbrida, donde las sondas de captura están diseñadas de forma que la región de la sonda de captura que está diseñada para hibridarse con el ADN que flanquea el punto polimórfico de la diana está separada del punto polimórfico por uno o un número reducido de bases.

Cuando las distribuciones alélicas medidas en un conjunto de loci polimórficos se utilizan para determinar el estado de ploidía de un individuo, resulta recomendable preservar las cantidades relativas de alelos de una muestra de ADN mientras se prepara para las mediciones genéticas. Esta preparación puede implicar la amplificación por WGA, la amplificación focalizada, técnicas de enriquecimiento selectivo, técnicas de captura híbrida, sondas de circularización u otros métodos diseñados para amplificar la cantidad de ADN y/o mejorar selectivamente la presencia de moléculas de ADN correspondientes a determinados alelos.

En algunas realizaciones de la presente divulgación, se presenta un conjunto de sondas de ADN diseñadas para loci diana, donde los loci tienen unas frecuencias alélicas menores máximas. En algunas realizaciones de la presente divulgación, se presenta un conjunto de sondas que están diseñadas para una diana donde los loci tienen la probabilidad máxima de que el feto tenga un SNP altamente informativo en esos loci. En algunas realizaciones de la presente divulgación, se presenta un conjunto de sondas diseñadas para loci diana, donde las sondas están optimizadas para un subgrupo de población determinado. En algunas realizaciones de la presente divulgación, se presenta un conjunto de sondas diseñadas para loci diana, donde las sondas están optimizadas para una mezcla determinada de subgrupos de población. En algunas realizaciones de la presente divulgación, se presenta un conjunto de sondas diseñadas para loci diana, donde las sondas están optimizadas para una determinada pareja de progenitores que pertenecen a subgrupos de población diferentes que tienen perfiles de frecuencia alélica menor diferentes. En algunas realizaciones de la presente divulgación, se presenta una cadena circularizada de ADN que comprende al menos un par de bases que se hibrida a una porción de ADN que es de origen fetal. En algunas realizaciones de la presente divulgación, se presenta una cadena circularizada de ADN que comprende al menos un par de bases que se hibrida a una porción de ADN que es de origen placentario. En algunas realizaciones de la presente divulgación, se presenta una cadena circularizada de ADN que se circularizó mientras que al menos algunos de los nucleótidos se hibridaban con ADN de origen fetal. En algunas realizaciones de la presente divulgación, se presenta una cadena circularizada de ADN que se circularizó mientras que al menos algunos de los nucleótidos se hibridaban con ADN de origen placentario. En algunas realizaciones de la presente divulgación, se presenta un conjunto de sondas donde algunas de las sondas están dirigidas a repeticiones en tándem simples y algunas de las sondas están dirigidas a polimorfismos de un solo nucleótido. En algunas realizaciones, los loci se seleccionan para los fines del diagnóstico prenatal no invasivo. En algunas realizaciones, las sondas se utilizan para los fines del diagnóstico prenatal no invasivo. En algunas realizaciones, los loci se focalizan utilizando un método que podría incluir sondas circularizantes, MIP, captura mediante sondas de hibridación, sondas de un array de SNP o combinaciones de estas. En algunas realizaciones, las sondas se utilizan como sondas circularizantes, MIP, captura mediante sondas de hibridación, sondas de un array de SNP o combinaciones de estas. En algunas realizaciones, los loci se secuencian para los fines del diagnóstico prenatal no invasivo.

Cuando la capacidad de información relativa de una secuencia es mayor si se combina con los contextos parentales pertinentes, al maximizar el número de lecturas de secuencia que contienen un SNP para el que se conoce el contexto parental se puede maximizar la capacidad de información del conjunto de lecturas de secuencia de la muestra mezclada. En una realización, el número de lecturas de secuencia que contienen un SNP para el que se conocen los contextos parentales se puede aumentar utilizando la qPCR para amplificar preferentemente secuencias específicas. En una realización, el número de lecturas de secuencia que contienen un SNP para el que se conocen los contextos parentales se puede mejorar utilizando sondas circularizantes (por ejemplo, MIP) para amplificar preferentemente secuencias específicas. En una realización, el número de lecturas de secuencia que contienen un SNP para el que se conocen los contextos parentales se puede mejorar utilizando una captura mediante un método de hibridación (por ejemplo, SURESELECT) para amplificar preferentemente secuencias específicas. Se pueden utilizar métodos diferentes para mejorar el número de lecturas de secuencia que contienen un SNP para el que se conocen los contextos parentales. En una realización, la focalización se puede realizar mediante unión con extensión, unión sin extensión, captura por hibridación o PCR.

En una muestra de ADN genómico fragmentado, una fracción de las secuencias de ADN corresponde de forma exclusiva a cromosomas individuales; otras secuencias de ADN se pueden encontrar en diferentes cromosomas. Cabe señalar que el ADN que se encuentra en el plasma, sea de origen materno o fetal, está típicamente fragmentado, a menudo con longitudes inferiores a 500 pares de bases. En una muestra genómica típica, aproximadamente el 3,3% de las secuencias correlacionables se correlacionarán con el cromosoma 13; el 2,2% de las secuencias correlacionables se correlacionarán con el cromosoma 18; el 1,35% de las secuencias correlacionables se correlacionarán con el cromosoma 21; el 4,5% de las secuencias correlacionables se correlacionarán con el cromosoma X en una mujer; el 2,25% de las secuencias correlacionables se correlacionarán con el cromosoma X (en un hombre); y el 0,73% de las secuencias correlacionables se correlacionarán con el cromosoma Y (en un hombre). Estos son los cromosomas con más probabilidades de presentar estado de aneuploidía en un feto. Por otra parte, entre las secuencias cortas,

aproximadamente 1 de 20 secuencias contendrá un SNP, utilizando los SNP contenidos en dbSNP. La proporción podrá ser superior dado que hay muchos SNP que no se han descubierto.

En una realización de la presente divulgación, se pueden utilizar los métodos de focalización para mejorar la fracción de ADN en una muestra de ADN que corresponden a un cromosoma determinado de modo que la fracción significativamente supera los porcentajes que se muestran más arriba que son típicos para muestras genómicas. En una realización de la presente divulgación, se pueden utilizar métodos de focalización para mejorar la fracción de ADN de una muestra de ADN de forma que el porcentaje de secuencias que contienen SNP sea notablemente mayor que los porcentajes típicos para las muestras genómicas. En una realización de la presente divulgación, se pueden utilizar métodos de focalización para focalizar ADN de un cromosoma o de un conjunto de SNP en una mezcla de ADN materno y fetal para los fines del diagnóstico prenatal.

Cabe señalar que un método ha sido documentado (Patente USA 7.888.017) para determinar la aneuploidía fetal contando el número de lecturas que corresponden a un cromosoma sospechoso y comparándolo con el número de lecturas que corresponden a un cromosoma de referencia, y utilizando la suposición de que un exceso de lecturas del cromosoma sospechoso corresponde a una triploidía en el feto en ese cromosoma. Estos métodos para el diagnóstico prenatal no harían uso de ningún tipo de focalización ni describen el uso de la focalización para el diagnóstico prenatal.

Haciendo uso de planteamientos de focalización en la secuenciación de la muestra mezclada, se puede conseguir un determinado nivel de precisión con menos lecturas de secuencia. La precisión se puede referir a la sensibilidad, se puede referir a la especificidad o se puede referir a una combinación de estas. El nivel deseado de precisión puede ser entre 90 y 95%; puede ser entre 95% y 98%; puede ser entre 98% y 99%; puede ser entre 99% y 99,5%; puede ser entre 99,5% y 99,9%; puede ser entre 99,9% y 99,99%; puede ser entre 99,99% y 99,999%; puede ser entre 99,999% y 100%). Los niveles de precisión superiores al 95% se pueden denominar de alta precisión.

Existen diversos métodos publicados en la técnica que demuestran cómo se puede determinar el estado de ploidía de un feto a partir de una muestra mezclada de ADN materno y fetal, por ejemplo: G.J. W. Liao et al. Clinical Chemistry 2011; 57(1) pp. 92-101. Estos métodos se centran en miles de ubicaciones a lo largo de cada cromosoma. El número de ubicaciones a lo largo de un cromosoma que se pueden focalizar al tiempo que se consigue una determinación de alta precisión del estado de ploidía en un feto, para un determinado número de lecturas de secuencia con una muestra de ADN mezclado, es inesperadamente limitado. En una realización de la presente divulgación, una determinación precisa del estado de ploidía se puede realizar utilizando secuenciación focalizada, utilizando cualquier método de focalización, por ejemplo, qPCR, PCR mediada por ligando, otros métodos de PCR, captura por hibridación, o sondas circularizantes, donde el número de loci a lo largo de un cromosoma que necesita ser focalizado puede ser de entre 5000 y 2000 loci; puede ser de entre 2000 y 1000 loci; puede ser de entre 1000 y 500 loci; puede ser de entre 500 y 300 loci; puede ser de entre 300 y 200 loci; puede ser de entre 200 y 150 loci; puede ser de entre 150 y 100 loci; puede ser de entre 100 y 50 loci; puede ser de entre 50 y 20 loci; puede ser de entre 20 y 10 loci. Óptimamente, puede ser de entre 100 y 500 loci. El elevado nivel de precisión se puede conseguir focalizando un pequeño número de loci y ejecutando un número inesperadamente reducido de lecturas de secuencia. El número de lecturas puede ser entre 100 millones y 50 millones de lecturas; el número de lecturas puede ser entre 50 millones y 20 millones de lecturas; el número de lecturas puede ser entre 20 millones y 10 millones de lecturas; el número de lecturas puede ser entre 10 millones y 5 millones de lecturas; el número de lecturas puede ser entre 5 millones y 2 millones de lecturas; el número de lecturas puede ser entre 2 millones y 1 millón de lecturas; el número de lecturas puede ser entre 1 millón y 500.000 lecturas; el número de lecturas puede ser entre 500.000 y 200.000 lecturas; el número de lecturas puede ser entre 200.000 y 100.000 lecturas; el número de lecturas puede ser entre 100.000 y 50.000 lecturas; el número de lecturas puede ser entre 50.000 y 20.000 lecturas; el número de lecturas puede ser entre 20.000 y 10.000 lecturas; el número de lecturas puede ser inferior a 10.000. Cuanto mayor es la cantidad de ADN disponible, menos lecturas se necesitan.

En algunas realizaciones, hay una composición que comprende una mezcla de ADN de origen fetal, y ADN de origen materno, donde el porcentaje de secuencias que se corresponden de forma exclusiva con el cromosoma 13 es superior a un 4%, superior a un 5%, superior a un 6%, superior a un 7%, superior a un 8%, superior a un 9%, superior a un 10%, superior a un 12%, superior a un 15%, superior a un 20%, superior a un 25%, o superior a un 30%. En algunas realizaciones de la presente divulgación, hay una composición que comprende una mezcla de ADN de origen fetal, y ADN de origen materno, donde el porcentaje de secuencias que se corresponden de forma exclusiva con el cromosoma 18 es superior a un 3%, superior a un 4%, superior a un 5%, superior a un 6%, superior a un 7%, superior a un 8%, superior a un 9%, superior a un 10%, superior a un 12%, superior a un 15%, superior a un 20%, superior a un 25%, o superior a un 30%. En algunas realizaciones de la presente divulgación, hay una composición que comprende una mezcla de ADN de origen fetal, y ADN de origen materno, donde el porcentaje de secuencias que se corresponden de forma exclusiva con el cromosoma 21 es superior a un 2%, superior a un 3%, superior a un 4%, superior a un 5%, superior a un 6%, superior a un 7%, superior a un 8%, superior a un 9%, superior a un 10%, superior a un 12%, superior a un 15%, superior a un 20%, superior a un 25%, o superior a un 30%. En algunas realizaciones de la presente divulgación, hay una composición que comprende una mezcla de ADN de origen fetal, y ADN de origen materno, donde el porcentaje de secuencias que se corresponden de forma exclusiva con el cromosoma X es superior a un 6%, superior a un 7%, superior a un 8%, superior a un 9%, superior a un 10%, superior a un 12%, superior a un 15%, superior a un 20%, superior a un 25%, o superior a un 30%. En algunas realizaciones de la presente divulgación, hay una composición que comprende una mezcla de ADN de origen fetal, y ADN de origen materno, donde el porcentaje de secuencias que se corresponden de forma exclusiva con el cromosoma Y es superior a un 1%, superior a un 2%, superior a un 3%, superior a un 4%, superior a un 5%, superior a un 6%, superior a un 7%, superior a un 8%, superior a

un 9%, superior a un 10%, superior a un 12%, superior a un 15%, superior a un 20%, superior a un 25%, o superior a un 30%.

En algunas realizaciones, se describe una composición que comprende una mezcla de ADN de origen fetal, y ADN de origen materno, donde el porcentaje de secuencias que se corresponden de forma exclusiva con un cromosoma, y que contienen al menos un polimorfismo de un solo nucleótido es superior a un 0,2%, superior a un 0,3%, superior a un 0,4%, superior a un 0,5%, superior a un 0,6%, superior a un 0,7%, superior a un 0,8%, superior a un 0,9%, superior a un 1%, superior a un 1,2%, superior a un 1,4%, superior a un 1,6%, superior a un 1,8%, superior a un 2%, superior a un 2,5%, superior a un 3%, superior a un 4%, superior a un 5%, superior a un 6%, superior a un 7%, superior a un 8%, superior a un 9%, superior a un 10%, superior a un 12%, superior a un 15%, o superior a un 20%, y donde el cromosoma se toma del grupo 13, 18, 21, X, o Y. En algunas realizaciones de la presente divulgación, se describe una composición que comprende una mezcla de ADN de origen fetal y ADN de origen materno, donde el porcentaje de secuencias que corresponde exclusivamente a un cromosoma y que contiene al menos un polimorfismo de un único nucleótido de un conjunto de polimorfismos de un único nucleótido es superior al 0,15%, superior al 0,2%, superior al 0,3%, superior al 0,4%, superior al 0,5%, superior al 0,6%, superior al 0,7%, superior al 0,8%, superior al 0,9%, superior al 9%, superior al 1%, superior al 1,2%, superior al 1,4%, superior al 1,6%, superior al 1,8%, superior al 2%, superior al 2,5%, superior al 3%, superior al 4%, superior al 5%, superior al 6%, superior al 7%, superior al 8%, superior al 9%, superior al 10%, superior al 12%, superior al 15%, o superior al 20%, donde el cromosoma se toma del grupo de cromosomas 13, 18, 21, X e Y, y donde el número de polimorfismos de un único nucleótido del conjunto de polimorfismos de un único nucleótido está entre 1 y 10, entre 10 y 20, entre 20 y 50, entre 50 y 100, entre 100 y 200, entre 200 y 500, entre 500 y 1000, entre 1000 y 2000, entre 2000 y 5000, entre 5000 y 10.000, entre 10.000 y 20.000, entre 20.000 y 50.000, y entre 50.000 y 100.000.

En teoría, cada ciclo de la amplificación duplica la cantidad de ADN presente; sin embargo, en la práctica el grado de amplificación es ligeramente inferior a dos. En teoría, la amplificación, incluyendo la amplificación focalizada, producirá una amplificación libre de sesgos de una mezcla de ADN; sin embargo, en la práctica los diferentes alelos tienden a amplificarse en diferente medida que otros alelos. Cuando se amplifica ADN, normalmente el grado de sesgo alélico aumenta con el número de pasos de amplificación. En algunas realizaciones, los métodos descritos en el presente documento implican la amplificación de ADN con un bajo nivel de sesgo alélico. Dado que el sesgo alélico aumenta con cada ciclo adicional, se puede determinar el sesgo alélico por ciclo calculando la raíz n -ésima del sesgo total donde n es el logaritmo en base 2 del grado de enriquecimiento. En algunas realizaciones, hay una composición que comprende una segunda mezcla de ADN, donde la segunda mezcla de ADN se ha enriquecido preferentemente con una pluralidad de loci polimórficos) de una primera mezcla de ADN donde el grado de enriquecimiento es al menos 10, al menos 100, al menos 1.000, al menos 10.000, al menos 100.000 o al menos 1.000.000, y donde el ratio de los alelos en la segunda mezcla de ADN en cada locus difiere del ratio de los alelos en ese locus en la primera mezcla de ADN por un factor que es, de media, inferior a 1.000%, 500%>, 200%>, 100%>, 50%>, 20%>, 10%>, 5%, 2%, 1%, 0,5%, 0,2%, 0,1%, 0,05%, 0,02%, o 0,01%. En algunas realizaciones, se presenta una composición que comprende una segunda mezcla de ADN, donde la segunda mezcla de ADN ha sido enriquecida preferentemente en una pluralidad de loci polimórficos de una primera mezcla de ADN, donde el sesgo alélico por ciclo para la pluralidad de loci polimórficos es, de media, inferior al 10%, 5%, 2%, 1%, 0,5%, 0,2%, 0,1%, 0,05%, o 0,02%. En algunas realizaciones, la pluralidad de loci polimórficos comprende al menos 10 loci, al menos 20 loci, al menos 50 loci, al menos 100 loci, al menos 200 loci, al menos 500 loci, al menos 1.000 loci, al menos 2.000 loci, al menos 5.000 loci, al menos 10.000 loci, al menos 20.000 loci, o al menos 50.000 loci.

Algunas realizaciones

En algunas realizaciones, se divulga un método en el presente documento para generar un informe que divulga el estado de ploidía determinado de un cromosoma en un feto en gestación, el método consiste en lo siguiente: la obtención de una primera muestra que contiene ADN de la madre del feto y ADN del feto; la obtención de datos genotípicos de uno o ambos padres del feto; la preparación de la primera muestra aislando el ADN para obtener una muestra preparada; la medición del ADN en la muestra preparada en una pluralidad de loci polimórficos; el cálculo, en un ordenador, de probabilidades de recuento de alelos o de recuentos de alelos en la pluralidad de loci polimórficos de las mediciones de ADN realizadas en la muestra preparada; la creación, en un ordenador, de una pluralidad de hipótesis de ploidía sobre probabilidades del recuento de alelos previstas en la pluralidad de loci polimórficos en el cromosoma para diferentes estados de ploidía posibles del cromosoma; la creación, en un ordenador, de un modelo de distribución conjunto para la probabilidad del recuento de alelos de cada locus polimórfico en el cromosoma para cada hipótesis de ploidía utilizando datos genotípicos de uno o ambos progenitores del feto; la determinación, en un ordenador, de una probabilidad relativa de cada una de las hipótesis de ploidía utilizando el modelo de distribución conjunto y las probabilidades del recuento de alelos calculadas para la muestra preparada; la determinación del estado de ploidía del feto seleccionando el estado de ploidía correspondiente a la hipótesis con la probabilidad más elevada; y la generación de un informe que divulgue el estado de ploidía determinado.

En algunas realizaciones, el método se utiliza para determinar el estado de ploidía de una pluralidad de fetos en gestación en una pluralidad de madres respectivas, el método además comprende: la determinación del porcentaje de ADN que es de origen fetal en cada una de las muestras preparadas; y donde el paso de medir el ADN en la muestra preparada se realiza mediante la secuenciación de un número de moléculas de ADN en cada una de las muestras preparadas, donde se secuencian más moléculas de ADN de aquellas muestras preparadas que tienen una fracción más pequeña de ADN fetal que aquellas muestras preparadas que tienen una fracción más grande de ADN fetal.

En algunas realizaciones, el método se utiliza para determinar el estado de ploidía de una pluralidad de fetos en gestación en una pluralidad de madres respectivas, y donde se realiza la medición del ADN en la muestra preparada, para cada uno de los fetos, mediante la secuenciación de una primera fracción de la muestra preparada de ADN para dar un primer conjunto de mediciones, el método además comprende: la creación de una primera determinación de probabilidad relativa para cada una de las hipótesis de ploidía para cada uno de los fetos, dado el primer conjunto de mediciones de ADN; la resecuenciación de una segunda fracción de la muestra preparada de aquellos fetos donde la primera determinación de la probabilidad relativa para cada una de las hipótesis de ploidía indica que una hipótesis de ploidía correspondiente a un feto aneuploide tiene una probabilidad significativa pero no conclusiva, para dar un segundo conjunto de mediciones; la creación de una segunda determinación de la probabilidad relativa para la hipótesis de ploidía para los fetos utilizando el segundo conjunto de mediciones y, opcionalmente, también el primer conjunto de mediciones; y la determinación de los estados de ploidía de los fetos cuya segunda muestra se ha resecuenciado seleccionando el estado de ploidía correspondiente a la hipótesis con la probabilidad más elevada determinada mediante la segunda determinación de probabilidad relativa.

En algunas realizaciones, se divulga una composición de materia, la composición de materia comprende: una muestra de ADN preferentemente enriquecido, donde la muestra de ADN preferentemente enriquecido se ha enriquecido preferentemente en una pluralidad de loci polimórficos de una primera muestra de ADN, donde la primera muestra de ADN consistía en una mezcla de ADN materno y ADN fetal derivado del plasma materno, donde el grado de enriquecimiento es al menos un factor de 2, y donde el sesgo alélico entre la primera muestra y la muestra preferentemente enriquecida se selecciona, de media, del grupo compuesto por menos de un 2%, menos de un 1%, menos de un 0,5%, menos de un 0,2%, menos de un 0,1 %, menos de un 0,05%, menos de un 0,02%, y menos de un 0,01%. En algunas realizaciones, se describe un método para crear una muestra de dicho ADN preferentemente enriquecido.

En alguna realización, se divulga un método para determinar la presencia o ausencia de aneuploidía fetal en una muestra de tejido materno que comprende ADN genómico fetal y materno, donde el método consiste en lo siguiente: (a) la obtención de una mezcla de ADN genómico materno y fetal de dicha muestra de tejido materno; (b) el enriquecimiento selectivo de la mezcla de ADN materno y fetal en una pluralidad de alelos polimórficos; (c) la distribución de los fragmentos enriquecidos selectivamente de la mezcla de ADN genómico materno y fetal del paso a para proporcionar muestras de la reacción que comprende una única molécula de ADN genómico o productos de amplificación de una única molécula de ADN genómico; (d) la realización de una secuenciación de ADN masivamente paralela de los fragmentos enriquecidos selectivamente de ADN genómico en las muestras de la reacción del paso c) para determinar la secuencia de dichos fragmentos enriquecidos selectivamente; (e) la identificación de los cromosomas a los que pertenecen las secuencias obtenidas en el paso d); (f) el análisis de los datos del paso d) para determinar i) el número de fragmentos de ADN genómico del paso d) que pertenecen a al menos un primer cromosoma diana que se supone que es diploide tanto en la madre como en el feto, y ii) el número de fragmentos de ADN genómico del paso d) que pertenecen a un segundo cromosoma diana, donde se sospecha que dicho segundo cromosoma es aneuploide en el feto; (g) el cálculo de una distribución prevista del número de fragmentos de ADN genómico del paso d) para el segundo cromosoma diana si el segundo cromosoma diana es euploide, utilizando el número determinado en el paso f) parte i); (h) el cálculo de una distribución prevista del número de fragmentos de ADN genómico del paso d) para el segundo cromosoma diana si el segundo cromosoma diana es aneuploide, utilizando el primer número en el paso f) parte i) y una fracción estimada de ADN fetal que se encuentra en la mezcla del paso b); e (i) la utilización de una probabilidad máxima o un planteamiento a posteriori máximo para determinar si el número de fragmentos de ADN genómico en el paso f) parte ii) es más probable que sea parte de la distribución calculada en el paso g) o la distribución calculada en el paso h); indicando de este modo la presencia o ausencia de aneuploidía fetal.

Ejemplos de métodos de diagnóstico del cáncer

Cabe señalar que se ha demostrado que el ADN procedente del cáncer que vive en un huésped se puede encontrar en la sangre del huésped. Del mismo modo que se pueden realizar diagnósticos genéticos con la medición del ADN mezclado que se encuentra en la sangre materna, también se pueden realizar diagnósticos genéticos con la medición del ADN mezclado que se encuentra en la sangre del hospedador. Los diagnósticos genéticos pueden incluir estados de aneuploidía o mutaciones genéticas. Cualquier reivindicación de la divulgación instantánea relativa a la determinación del estado de ploidía o del estado genético de un feto a partir de las mediciones realizadas con la sangre materna se puede aplicar igualmente a la determinación del estado de ploidía o del estado genético de un cáncer a partir de las mediciones realizadas con la sangre del hospedador. En algunas realizaciones, un método de la presente divulgación permite determinar el estado de ploidía de un cáncer, el método incluye la obtención de una muestra mezclada que contiene material genético del hospedador, y material genético del cáncer; y la medición del ADN en la muestra mezclada; el cálculo de la fracción de ADN de origen cancerígeno en la muestra mezclada; y la determinación del estado de ploidía del cáncer utilizando las mediciones realizadas en la muestra mezclada y la fracción calculada. En algunas realizaciones, el método puede incluir también la administración de un terapéutico para el cáncer basándose en la determinación del estado de ploidía del cáncer. En algunas realizaciones, el método puede incluir también la administración de un terapéutico para el cáncer basado en la determinación del estado de ploidía del cáncer, donde el terapéutico se toma del grupo compuesto por un producto farmacéutico, un terapéutico biológico y una terapia con anticuerpos o una combinación de estos.

Ejemplos de métodos de implementación

Cualquiera de las realizaciones divulgadas en el presente documento se puede implementar en circuitos electrónicos digitales, circuitos integrados, ASIC (circuitos integrados de aplicación específica) especialmente diseñados, hardware informático, firmware, software o combinaciones de estos. El aparato de las realizaciones divulgadas en el presente documento se puede implementar en un producto de programa informático realizado de forma tangible en un dispositivo de almacenamiento legible electrónicamente para su ejecución por un procesador programable; y los pasos del método de las realizaciones divulgadas en el presente documento se pueden realizar mediante un procesador programable que ejecuta un programa de instrucciones para realizar funciones de las realizaciones divulgadas en el presente documento operando sobre los datos de entrada y generando datos de salida. Las realizaciones divulgadas en el presente documento se pueden implementar de forma ventajosa en uno o más programas informáticos que se puedan ejecutar y/o interpretar en un sistema programable

incluyendo al menos un procesador programable, que puede ser especial o para múltiples usos, conectado para recibir datos e instrucciones de un sistema de almacenamiento y para transmitir datos e instrucciones a dicho sistema de almacenamiento, al menos un dispositivo de entrada, y al menos un dispositivo de salida. Cada programa informático se puede implementar en un lenguaje de programación orientado al objeto o procedural de alto nivel o en un lenguaje ensamblador o lenguaje máquina si se desea; y en cualquier caso el lenguaje puede ser un lenguaje compilado o interpretado. Se puede desplegar un programa informático en cualquier forma, incluido como programa independiente, o como un módulo, componente, subrutina u otra unidad adecuada para su uso en un entorno informático. Se puede desplegar un programa informático para que sea ejecutado o interpretado en un ordenador o múltiples ordenadores en un punto, o distribuido en múltiples puntos e interconectado a través de una red de comunicación.

Los medios de almacenamiento legibles por ordenador, como los utilizados en el presente documento, se refieren a un almacenamiento físico o tangible (en oposición a las señales) e incluyen, entre otros, medios volátiles y no volátiles, extraíbles y no extraíbles, implementados en cualquier método o tecnología para el almacenamiento tangible de información como instrucciones legibles por ordenador, estructuras de datos, módulos de programas y otros datos. Los medios de almacenamiento legibles por ordenador incluyen, entre otros, RAM, ROM, EPROM, EEPROM, memoria flash u otra tecnología de memoria de estado sólido, CD-ROM, DVD, u otro almacenamiento óptico, cassettes magnéticos, cinta magnética, almacenamiento en disco magnético u otros dispositivos de almacenamiento magnético, o cualquier otro medio físico o material que se pueda utilizar para almacenar tangiblemente la información deseada o datos o instrucciones y al que se pueda acceder mediante un ordenador o procesador.

Cualquiera de los métodos descritos en el presente documento puede incluir la producción de datos en un formato físico, como la pantalla de un ordenador o una copia impresa en papel. En las explicaciones de cualquiera de las realizaciones que aparecen en otra parte de este documento, se debe entender que los métodos descritos se pueden combinar con la salida de los datos ejecutables en un formato con el que puede actuar un médico. Por otra parte, los métodos descritos se pueden combinar con la ejecución real de una decisión clínica que resulta en un tratamiento clínico o la ejecución de una decisión clínica de no emprender medidas. Algunas de las realizaciones descritas en el documento para determinar los datos genéticos pertenecientes a un individuo diana se pueden combinar con la decisión de seleccionar uno o más embriones para la transferencia en el contexto de la IVF, opcionalmente combinada con el proceso de transferencia del embrión al útero de la futura madre. Algunas de las realizaciones descritas en el documento para determinar los datos genéticos pertenecientes a un individuo diana se pueden combinar con la notificación de una potencial anomalía cromosómica, o la ausencia de esta, a un profesional médico, combinada opcionalmente con la decisión de abortar o no un feto en el contexto del diagnóstico prenatal. Algunas de las realizaciones descritas en el presente documento se pueden combinar con la producción de datos ejecutables y la ejecución de una decisión clínica que resulta en un tratamiento clínico o la ejecución de una decisión clínica de no emprender medidas.

Ejemplos de cuadros de diagnóstico

En una realización, la presente divulgación comprende un cuadro de diagnóstico que es capaz de realizar parcial o completamente cualquiera de los métodos descritos en esta divulgación. En una realización, el cuadro de diagnóstico puede estar ubicado en la consulta de un médico, el laboratorio de un hospital o cualquier lugar adecuado y relativamente próximo al punto de atención del paciente. El cuadro puede ser capaz de ejecutar el método completo de forma totalmente automatizada o puede que un técnico tenga que completar uno o varios pasos manualmente. En una realización, el cuadro puede ser capaz de analizar al menos los datos genotípicos medidos en el plasma materno. En una realización, el cuadro puede estar conectado a medios que transmiten los datos genotípicos medidos en el cuadro de diagnóstico a una instalación de cálculo externa que puede posteriormente analizar los datos genotípicos y posiblemente también generar un informe. El cuadro de diagnóstico puede incluir una unidad robótica que es capaz de transferir muestras acuosas o líquidas de un contenedor a otro. Puede comprender una serie de reactivos, tanto sólidos como líquidos. Puede comprender un secuenciador de alto rendimiento. Puede comprender un ordenador.

Sección experimental

Las realizaciones divulgadas en el presente documento se describen en los siguientes ejemplos, que se establecen para facilitar la comprensión de la divulgación. Los siguientes ejemplos se presentan para proporcionar al experto en la técnica una divulgación y descripción completa de cómo utilizar las realizaciones descritas. Se ha hecho todo lo posible para garantizar la precisión con respecto a los números utilizados (por ejemplo, cantidades, temperatura, etc.) pero se debe tener en cuenta la posibilidad de algunas desviaciones y errores experimentales. A menos que se indique lo contrario, las partes son partes por volumen y la temperatura se indica en grados centígrados. Se debe entender que las

variaciones en los métodos que se describen se pueden realizar sin cambiar los aspectos fundamentales que los experimentos pretenden ilustrar.

Experimento 1

El objetivo fue mostrar que un algoritmo de una estimación de la probabilidad máxima Bayesiana (MLE) que utiliza genotipos de los progenitores para calcular la fracción fetal mejora la precisión del diagnóstico de la trisomía prenatal no invasiva en comparación con los métodos publicados.

Los datos de secuenciación simulados para el ADN libre de células materno se crearon mediante lecturas de muestreo obtenidas en trisomía-21 y en las líneas celulares maternas respectivas. La tasa de determinaciones de disomía y trisomía correcta se determinó a partir de 500 simulaciones en varias fracciones fetales para un método publicado (Chiu et al. BMJ 2011;342:c7401) y nuestro algoritmo basado en MLE. Validamos las simulaciones mediante la obtención de 5 millones de lecturas por fuerza bruta de cuatro madres embarazadas y los respectivos padres recogidas bajo un protocolo aprobado por IRB. Los genotipos parentales se obtuvieron en un array de SNP de 290K. (Véase la **Figura 14**)

En simulaciones, el planteamiento basado en MLE obtuvo una precisión del 99,0% para fracciones fetales tan bajas como un 9% e informó de certezas que corresponden a una precisión global. Validamos estos resultados utilizando cuatro muestras reales donde obtuvimos todas las determinaciones correctas con una certeza calculada que superó el 99%. Por el contrario, nuestra implementación del algoritmo publicado para Chiu et al. requirió un 18% de fracción fetal para conseguir una precisión del 99,0%, y consiguió únicamente un 87,8% de precisión a un 9% de ADN fetal.

La determinación de la fracción fetal de genotipos parentales conjuntamente con el planteamiento basado en MLE consigue una mayor precisión que los algoritmos publicados en las fracciones fetales previstas durante el primer trimestre y a principios del segundo trimestre. Además, el método que se divulga en el presente documento produce una métrica de certeza que es crucial en la determinación de la fiabilidad del resultado, especialmente en fracciones fetales bajas donde la detección de la ploidía es más difícil. Los métodos publicados utilizan un método con un umbral menos preciso para determinar la ploidía basado en grandes conjuntos de datos de formación de la disomía, un planteamiento que predefine una tasa de falsos positivos. Además, sin una métrica de certeza, los métodos publicados corren el riesgo de informar de resultados falsos negativos cuando hay un ADN libre de células fetal insuficiente para realizar una determinación. En algunas realizaciones se calcula una estimación de certeza para el estado de ploidía determinado.

Experimento 2

El objetivo fue mejorar la detección no invasiva de la trisomía fetal 18, 21, y X especialmente en muestras compuestas de fracciones fetales bajas utilizando un planteamiento de secuenciación focalizada combinado con los genotipos de los progenitores y datos Hapmap en un algoritmo de estimación de la probabilidad máxima Bayesiana (MLE).

Las muestras maternas de cuatro embarazos euploides y dos con trisomía-positiva y las muestras paternas respectivas se obtuvieron bajo un protocolo aprobado por RB de pacientes en los que se conocía el cariotipo fetal. El ADN libre de células materno se extrajo del plasma y se obtuvieron aproximadamente 15 millones de lecturas de secuencia tras el enriquecimiento preferente que focalizó SNPs específicos. Las muestras del progenitor se secuenciaron de forma similar para obtener genotipos.

El algoritmo descrito determinó correctamente la disomía 18 y 21 del cromosoma para todas las muestras euploides y cromosomas normales de muestras aneuploides. Las determinaciones de trisomía 18 y 21 fueron correctas, como lo fueron los números de copia del cromosoma X en fetos masculinos y femeninos. La certeza producida por el algoritmo fue superior a un 98% en todos los casos.

El método descrito informó de forma precisa de la ploidía de todos los cromosomas que fueron sometidos a pruebas de seis muestras, incluyendo muestras compuestas de menos de un 12% de ADN fetal, lo que representa aproximadamente un 30% de las muestras del primer trimestre y de principios del segundo trimestre. La diferencia crucial entre el algoritmo MLE instantáneo y los métodos publicados es que utiliza los genotipos del progenitor y los datos Hapmap para mejorar la precisión y generar una métrica de certeza. A fracciones fetales bajas, todos los métodos se vuelven menos precisos; es importante identificar correctamente las muestras sin suficiente ADN libre de células fetal para realizar una determinación fiable. Otros han utilizado sondas específicas del cromosoma Y para estimar la fracción fetal de fetos masculinos, pero la determinación del genotipo parental concurrente permite la estimación de la fracción fetal para ambos sexos. Otra limitación inherente de los métodos publicados que utilizan una secuenciación por fuerza bruta no focalizada es que la precisión de la determinación de ploidía varía entre los cromosomas debido a diferencias en los factores como la riqueza de GC. El planteamiento de secuenciación focalizada instantánea depende en gran medida de dichas variaciones en la escala del cromosoma y produce un rendimiento más constante entre cromosomas.

Experimento 3

El objetivo era determinar si la trisomía es detectable con una alta seguridad en un feto triploide, utilizando un programa informático nuevo para analizar loci de SNP de ADN fetal flotante libre en plasma materno.

Se extrajeron 20 mL de sangre de una paciente embarazada tras ultrasonidos anómalos. Después de la centrifugación, se extrajo ADN materno de la capa leucocitaria (DNEASY, QIAGEN); se extrajo ADN libre de células del plasma (QIAAMP QIAGEN). Se aplicó una secuenciación focalizada a los loci de SNP en los cromosomas 2, 21, y X en ambas muestras de ADN. La estimación Bayesiana de probabilidad máxima seleccionó la hipótesis más probable del conjunto

de todos los estados de ploidía posibles. El método determina la fracción de ADN fetal, el estado de ploidía y las certezas explícitas en la determinación del estado de ploidía. No se realiza ninguna suposición sobre la ploidía de un cromosoma de referencia. El diagnóstico utiliza una estadística de ensayo que es independiente de los recuentos de lectura de la secuencia, que es el estado reciente de la técnica.

5 El método instantáneo diagnosticó con precisión la trisomía de los cromosomas 2 y 21. La fracción del niño se estimó en un 11,9% [CI 11,7-12,1], se descubrió que el feto tenía una copia materna y dos paternas de los cromosomas 2 y 21 con una certeza de efectivamente 1 (error-30 probabilidad $<10^{-30}$). Esto se consiguió con 92.600 y 258.100 lecturas en los cromosomas 2 y 21 respectivamente.

10 Esta es la primera demostración de diagnóstico prenatal no invasivo de cromosomas trisómicos de sangre materna donde el feto era triploide, según confirmó el cariotipo de metafase. Los métodos existentes de diagnosis no invasiva no detectarían aneuploidía en esta muestra. Los métodos actuales se basan en una sobreproducción de lecturas de secuencia en un cromosoma trisómico relacionado con cromosomas de referencia disómicos; aunque un feto triploide no tiene referencia disómica. Además, los métodos existentes no conseguirían de forma similar una determinación del estado de ploidía de alta certeza con esta fracción de ADN fetal y número de lecturas de secuencia. Es sencillo ampliar
15 el planteamiento a los 24 cromosomas.

Experimento 4

El siguiente protocolo se utilizó para la amplificación de 800-plex de ADN aislado del plasma materno de un embarazo euploide y también ADN genómico de una línea celular de triploidía 21 utilizando PCR estándar (es decir, no se utilizó el anidado). La preparación de la biblioteca y la amplificación supuso un extremo romo de un único tubo seguido de un
20 factor de cola A. La unión de adaptador se ejecutó utilizando el kit de unión que se encuentra en el kit AGILENT SURESELECT, y la PCR se ejecutó durante 7 ciclos. A continuación, 15 ciclos de STA (95 °C durante 30 s; 72 °C durante 1 min; 60 °C durante 4 min; 65 °C durante 1 min; 72 °C durante 30 s) utilizando 800 diferentes pares de cebadores focalizando SNPs en los cromosomas 2, 21 y X. La reacción se ejecutó con una concentración del cebador de 12,5 nM. A continuación, se secuenció el ADN con un secuenciador ILLUMINA IIGAX. El secuenciador produce 1,9
25 millones de lecturas, de las cuales un 92% se correlaciona con el genoma; de estas lecturas que se han correlacionado con el genoma, más del 99% se han correlacionado con una de las regiones focalizadas por los cebadores focalizados. Los números eran básicamente el mismo para tanto el ADN del plasma como el ADN genómico. La **Figura 15** muestra el ratio de los dos alelos para los -780 SNPs detectados por el secuenciador en el ADN genómico que se tomó de una línea celular con trisomía conocida en el cromosoma 21. Cabe señalar que los ratios de alelos se representan aquí
30 gráficamente para facilitar la visualización, ya que las distribuciones alélicas no son sencillas de leer visualmente. Los círculos representan SNPs en cromosomas disómicos, mientras que las estrellas representan SNPs en un cromosoma trisómico. La **Figura 16** es otra representación de los mismos datos que en la Figura X, donde el eje Y es el número relativo de A y B medido para cada SNP, y donde el eje X es el número SNP donde los SNPs están separados por cromosoma. En la **Figura 16**, del SNP 1 al 312 se encuentran en el cromosoma 2, del SNP 313 al 605 se encuentran en
35 el cromosoma 21 que es trisómico, y del SNP 606 al 800 están en el cromosoma X. Los datos de los cromosomas 2 y X muestran un cromosoma disómico, ya que la secuencia relativa se encuentra en tres agrupaciones: AA en la parte superior del gráfico, BB en la parte inferior del gráfico, y AB en la parte central del gráfico. Los datos del cromosoma 21, que es trisómico, muestran cuatro agrupaciones: AAA en la parte superior del gráfico, AAB alrededor de la línea 0,65 (2/3), ABB alrededor de la línea 0,35 (1/3), y BBB en la parte inferior del gráfico.

40 Las **Figuras 17A-D**: muestran datos para el mismo protocolo de 800-plex, pero medidas en ADN que se amplificó de cuatro muestras de plasma de mujeres embarazadas. Para estas cuatro muestras, esperamos ver siete agrupaciones de puntos: (1) a lo largo de la parte superior del gráfico están aquellos loci donde tanto la madre como el feto son AA, (2) ligeramente por debajo de la parte superior del gráfico están aquellos loci en los que la madre es AA y el feto es AB, (3) ligeramente por encima de la línea 0,5 están aquellos loci en los que la madre es AB y el feto es AA, (4) a lo largo de
45 la línea 0,5 están aquellos loci en los que la madre y el feto son ambos AB, (5) ligeramente por debajo de la línea 0,5 están aquellos loci en los que la madre es AB y el feto es BB, (6) ligeramente por encima de la parte inferior del gráfico están aquellos loci en los que la madre es BB y el feto es AB, (1) a lo largo de la parte inferior del gráfico están aquellos loci en los que tanto la madre como el feto son BB. Cuanto más pequeña sea la fracción fetal, menos será la separación entre las agrupaciones (1) y (2), entre las agrupaciones (3), (4) y (5), y entre las agrupaciones (6) y (7). La separación
50 está previsto que sea la mitad de la fracción de ADN que es de origen fetal. Por ejemplo, si el ADN es un 20% fetal y un 80% materno, se espera que del (1) al (7) estén centrados en 1,0, 0,9, 0,6, 0,5, 0,4, 0,1 y 0,0 respectivamente; véase, por ejemplo, la **Figura 17D**, POOL1_BC5_índice_ref. Si, en caso contrario el ADN es un 8% fetal, y un 92% materno, se espera que del (1) al (7) estén centrados en 1,00, 0,96, 0,54, 0,50, 0,46, 0,04 y 0,00 respectivamente; véase por ejemplo, la **Figura 17B**, POOL1_BC2_índice_ref. Si no se detecta ADN fetal, no se espera ver (2), (3), (5), o (6); de
55 forma alternativa se puede decir que la separación es cero y, por tanto, (1) y (2) están en la parte superior uno de otro, como lo están (3), (4) y (5), y también (6) y (7); véase por ejemplo, la **Figura 17C**, 15POOL1_BC7_índice_ref. Cabe señalar que la fracción fetal para la **Figura 17A**, Índice de ref. POOL1 BC1 es de un 25%.

Experimento 5

60 La mayoría de métodos de amplificación de ADN y medición producirán algún sesgo alélico, donde los dos alelos que de forma típica se encuentran en un locus se detectan con intensidades o recuentos que no son representativos de las cantidades reales de alelos en la muestra de ADN. Por ejemplo, para un solo individuo, en un locus heterocigoto esperamos ver un ratio 1:1 de los dos alelos, que es el ratio teórico que se espera para un locus heterocigoto; sin

embargo, debido al sesgo alélico, es posible que veamos uno de 55:45, o incluso 60:40. También cabe señalar que en el contexto de la secuenciación, si la profundidad de lectura es baja, un simple ruido estocástico podría dar como resultado un sesgo alélico significativo. En una realización, es posible modelar el comportamiento de cada SNP de modo que si se observa un sesgo constante para determinados alelos, este sesgo se puede corregir. La **Figura 18** muestra la fracción de datos que se puede explicar por la varianza binomial antes y después de la corrección del sesgo. En la **Figura 18**, las estrellas representan el sesgo alélico observado en los datos de secuencia brutos para el experimento de 800-plex; los círculos representan el sesgo alélico después de la corrección. Cabe señalar que si no hubiera un sesgo alélico en absoluto, se esperaría que los datos recayeran en la línea $x=y$. Un conjunto de datos similar del que se ha producido mediante la amplificación de ADN utilizando una amplificación focalizada de 150-plex produjo datos que se encontraron muy cerca de la línea 1:1 después de la corrección del sesgo.

Experimento 6

La amplificación universal de ADN utilizando adaptadores unidos con cebadores específicos para las etiquetas del adaptador, donde la reformación térmica del cebador y los tiempos de ampliación son limitados a unos pocos minutos tiene el efecto de enriquecer la proporción de hebras de ADN más cortas. La mayoría de protocolos de biblioteca diseñados para crear bibliotecas de ADN adecuadas para la secuenciación contienen dicho paso, y se publican protocolos de ejemplos que son bien conocidos para el experto en la técnica. En algunas realizaciones de la divulgación, los adaptadores con una etiqueta universal están unidos al ADN del plasma, y amplificados utilizando cebadores específicos a la etiqueta del adaptador. En algunas realizaciones, la etiqueta universal puede ser la misma etiqueta que se utiliza para la secuenciación, puede ser una etiqueta universal sólo para la amplificación por PCR, o puede ser un conjunto de etiquetas. Puesto que el ADN fetal es de forma típica corto, mientras que el ADN materno puede ser tanto de naturaleza corta como larga, este método tiene el efecto de enriquecer la proporción de ADN fetal en la mezcla. El ADN flotante libre, que se cree que es ADN de células apoptóticas, y que contiene tanto ADN materno como fetal, es corto, en su mayor parte por debajo de 200 pares de bases. El ADN celular liberado por la célula lisis, un fenómeno común después de la flebotomía, es de forma típica casi exclusivamente materno, y es también bastante largo, en su mayor parte por encima de 500 pares de bases. Por tanto, las muestras de sangre que han estado asentadas durante más de unos minutos contendrán una mezcla de ADN (fetal + materno) corto y (materno) más largo. La realización de una amplificación universal con tiempos de ampliación relativamente cortos en plasma materno seguida de una amplificación focalizada tenderá a aumentar la proporción relativa de ADN fetal al compararla con el plasma que se ha amplificado utilizando únicamente la amplificación focalizada. Esto se puede ver en la **Figura 19** que muestra el porcentaje fetal medido cuando el aporte es ADN del plasma (eje vertical) frente al porcentaje fetal medido cuando el ADN aportado es ADN del plasma que ha tenido una biblioteca preparada utilizando el protocolo de preparación de bibliotecas ILLUMINA GAIIx. Todos los puntos caen por debajo de la línea, lo que indica que el paso de preparación de la biblioteca enriquece la fracción de ADN que es de origen fetal. Dos muestras de plasma que eran rojas, indicando hemólisis y por tanto que habría un incremento en la cantidad de ADN materno largo presente de célula lisis, muestran un enriquecimiento especialmente significativo de la fracción fetal cuando la preparación de la biblioteca se realiza antes de la amplificación focalizada. El método divulgado en el presente documento es especialmente útil en los casos donde hay hemólisis o se ha producido cualquier otra situación donde las células que comprenden hebras relativamente largas de ADN contaminante se han lisado, contaminando la muestra mezclada de ADN corto con el ADN largo. De forma típica la reformación térmica relativamente corta y los tiempos de ampliación están entre 30 segundos y 2 minutos, aunque podrían ser tan cortos como 5 o 10 segundos o menos, o tan largos como 5 o 10 minutos.

Experimento 7

El siguiente protocolo se utilizó para la amplificación de 1.200-plex de ADN aislado del plasma materno de un embarazo euploide y también ADN genómico de una línea celular de triploidía 21 utilizando un protocolo de PCR directo, y también un planteamiento semianidado. La preparación de la biblioteca y la amplificación supuso un extremo romo de un único tubo seguido de un factor de cola A. La unión de adaptador se ejecutó utilizando una modificación del kit de unión que se encuentra en el kit AGILENT SURESELECT, y la PCR se ejecutó durante 7 ciclos. En el conjunto del cebador focalizado, había 550 ensayos para SNPs del cromosoma 21, y 325 ensayos para SNPs de cada uno de los cromosomas 1 y X. Ambos protocolos suponían 15 ciclos de STA (95 °C durante 30 s; 72 °C durante 1 min; 60 °C durante 4 min; 65 °C durante 30 s; 72 °C durante 30 s) utilizando una concentración del cebador de 16 nM. El protocolo de PCR semianidado suponía una segunda amplificación de 15 ciclos de STA (95 °C durante 30 s; 72 °C durante 1 min; 60 °C durante 4 min; 65 °C durante 30 s; 72 °C durante 30 s) utilizando una concentración de etiqueta directa interior de 29 nM, y una concentración de etiqueta inversa de 1 uM o 0,1 uM. A continuación, se secuenció el ADN con un secuenciador ILLUMINA IIGAX. Para el protocolo de PCR directa, un 73% de las lecturas corresponde al genoma; para el protocolo semianidado, un 97,2% de las lecturas de secuencia corresponde al genoma. Por tanto, el protocolo semianidado da como resultado aproximadamente un 30% más de información, presumiblemente debido en gran parte a la eliminación de los cebadores que es más probable que causen dímeros de cebadores.

La variabilidad en la profundidad de lectura tiende a ser más alta cuando se utiliza el protocolo semianidado que cuando se utiliza el protocolo de PCR directa (véase la **Figura 20**) donde los diamantes hacen referencia a la profundidad de lectura para loci ejecutados con el protocolo semianidado, y los cuadrados hacen referencia a la profundidad de lectura para loci ejecutados sin anidado. Los SNPs están dispuestos por profundidad de lectura para los diamantes, de modo que todos los diamantes caen en una línea curva, mientras que los cuadrados aparecen correlacionados holgadamente; la disposición de los SNPs es arbitraria, y es la altura del punto lo que denota la profundidad de lectura en lugar de su ubicación de izquierda a derecha.

En algunas realizaciones, los métodos que se describen en el presente documento pueden obtener unas variaciones en la profundidad de lectura (DOR) excelentes. Por ejemplo, en un versión de este experimento (**Figura 21**) utilizando una amplificación por PCR directa de 1.200-plex de ADN genómico, de los 1.200 ensayos: 1186 ensayos tuvieron una DOR superior a 10; la profundidad de lectura media fue de 400; 1063 ensayos (88,6%) tuvieron una profundidad de lectura de entre 200 y 800, y una ventana ideal donde el número de lecturas para cada alelo es lo suficientemente alta para dar datos significativos, mientras que el número de lecturas para cada alelo no es tan alto de modo que el uso marginal de esas lecturas fue especialmente pequeño. Sólo 12 alelos tenían una profundidad de lectura más alta siendo la máxima de 1035 lecturas. La desviación estándar de la DOR fue de 290, la DOR fue de 453, el coeficiente de varianza de la DOR fue de un 64%, hubo un total de 950.000 lecturas, y 63,1% de las lecturas se correspondieron con el genoma. En otro experimento (**Figura 22**) utilizando un protocolo semianidado de 1.200-plex, la DOR fue más alta. La desviación estándar de la DOR fue de 583, la DOR fue de 630, el coeficiente de varianza de la DOR fue de un 93%, hubo un total de 870.000 lecturas, y 96,3% de las lecturas se correspondieron con el genoma. Cabe señalar que en ambos casos, los SNPs están dispuestos por la profundidad de lectura para la madre, de modo que la línea curva representa la profundidad de lectura materna. La diferenciación entre niño y padre no es significativa; sólo es la tendencia lo que es significativo para los fines de esta explicación.

Experimento 8

En un experimento, el protocolo para PCR de 1.200-plex semianidado se utilizó para amplificar el ADN de una célula y de tres células. Este experimento es relevante para la prueba de aneuploidía prenatal utilizando células fetales aisladas de la sangre materna, o para el diagnóstico genético previo al implante utilizando blastómeros biopsiados o muestras de trofotodermo. Hubo 3 duplicados de 1 y 3 células de individuos (46 XY y 47 XX+21) por condición. Los ensayos focalizaron los cromosomas 1, 21 y X. Se utilizaron tres diferentes métodos de lisis. ARCTURUS, MPERv2 y lisis alcalina. La secuenciación se ejecutó multiplexando 48 muestras en una línea de secuenciación. El algoritmo devolvió determinaciones correctas del estado de ploidía para cada uno de los tres cromosomas, y para cada uno de los duplicados.

Experimento 9

En un experimento, se prepararon cuatro muestras de plasma materno y se amplificaron utilizando un protocolo de 9.600-plex hemianidado. Las muestras se prepararon del modo siguiente: Se centrifugaron hasta 40 mL de sangre materna para aislar la capa leucocitaria y el plasma. El ADN genómico en la muestra materna se preparó de la capa leucocitaria y el ADN paterno se preparó a partir de una muestra de sangre o una muestra de saliva. El ADN libre de células en el plasma materno se aisló utilizando el kit de ácido nucleico circulante QIAGEN y se eluyó en 45 uL de tampón TE en función de las instrucciones del fabricante. Se anexaron adaptadores de unión universales al extremo de cada molécula de 35 uL de ADN del plasma purificado y las bibliotecas se ampliaron durante 7 ciclos utilizando cebadores específicos del adaptador. Las bibliotecas se purificaron con perlas AGENCOURT AMPURE y se eluyeron en 50 uL de agua.

Se amplificaron 3 uL del ADN con 15 ciclos de STA (95 °C durante 10 min para la activación de la polimerasa inicial, a continuación, 15 ciclos de 95 °C durante 30 s; 72 °C durante 10 s; 65 °C durante 1 min; 60 °C durante 8 min; 65 °C durante 3 min; y 72 °C durante 30 s; y una ampliación final a 72 °C durante 2 min;) utilizando una concentración del cebador de 14,5 nM de 9600 cebadores inversos etiquetados específicos diana y un cebador directo específico del adaptador de biblioteca a 500 nM.

El protocolo para PCR hemianidado suponía una segunda amplificación de una dilución del primer producto de STAs durante 15 ciclos de STA (95 °C durante 10 min para la activación de la polimerasa inicial, a continuación, 15 ciclos de 95 °C durante 30 s; 65 °C durante 1 min; 60 °C durante 5 min; 65 °C durante 5 min y 72 °C durante 30 s; y una ampliación final a 72 °C durante 2 min) utilizando una concentración de etiqueta inversa de 1000 nM, y una concentración de 16,6 u nM para cada uno de los 9600 cebadores directos específicos diana.

A continuación, se amplificó una alícuota de productos de STA mediante PCR estándar durante 10 ciclos con 1 uM de cebadores directos específicos etiquetados y cebadores inversos codificados mediante código de barras para generar bibliotecas de secuenciación codificadas mediante código de barras. Se mezcló una alícuota de cada biblioteca con bibliotecas de diferentes códigos de barras y se purificó utilizando una columna de centrifugación.

De este modo, se utilizaron 9.600 cebadores en las reacciones de un único pocillo; los cebadores se diseñaron para focalizar SNPs encontrados en los cromosomas 1, 2, 13, 18, 21, X e Y. A continuación, se secuenciaron los amplicones utilizando un secuenciador ILLUMINA GAIIIX. Por muestra, el secuenciador generó aproximadamente 3,9 millones de lecturas, con 3,7 millones de lecturas correlacionadas con el genoma; (94%), y de esas, 2,9 millones de lecturas (74%) correlacionadas con los SNPs focalizados con una profundidad media de lectura de 344 y una mediana de profundidad de lectura de 255. Se descubrió que la fracción fetal para las cuatro muestras era del 9,9%, 18,9%, 16,3%, y 21,2%

Las muestras de ADN genómico materno y paterno relevantes se amplificaron utilizando un protocolo de 9600-plex semianidado y se secuenciaron. El protocolo semianidado es diferente en que aplica 9.600 cebadores directos exteriores y cebadores inversos etiquetados a 7,3 nM en la primera STA. Las condiciones termocíclicas y la composición de la segunda STA, y la PCR mediante codificación por medio de código de barras eran las mismas que para el protocolo hemianidado.

Los datos de secuenciación se analizaron utilizando métodos informáticos divulgados en el presente documento y se determinó el estado de ploidía en seis cromosomas para los fetos cuyo ADN estaba presente en las 4 muestras de plasma materno. Las determinaciones del estado de ploidía para los 28 cromosomas en el conjunto se determinaron correctamente con certezas superiores al 99,2% excepto por un cromosoma que se determinó correctamente, pero con una certeza del 83%.

La Figura 23 muestra la profundidad de lectura del planteamiento hemianidado de 9.600-plex junto con la profundidad de lectura del planteamiento semianidado de 1.200-plex descrito en el Experimento 7, aunque el número de SNPs con una profundidad de lectura superior a 100, superior a 200 y superior a 400 fue significativamente más alto que en el protocolo de 1.200-plex. El número de lecturas en el 90° percentil se puede dividir por el número de lecturas en el 10° percentil para dar una métrica sin dimensiones que es indicativa de la uniformidad de la profundidad de lectura; el número más pequeño, el más uniforme (estrecho) la profundidad de lectura. El 90° percentil medio/10° ratio de percentil es 11,5 para el método ejecutado en el Experimento 9, mientras que es 5,6 para el método ejecutado en el Experimento 7. Una profundidad de lectura más estrecha para la plexidad de un protocolo determinado es mejor para la eficiencia de la secuenciación, ya que son necesarias menos lecturas de secuencia para garantizar que un determinado porcentaje de lecturas está por encima de un umbral de número de lecturas.

Experimento 10

En un experimento, se prepararon cuatro muestras de plasma materno y se amplificaron utilizando un protocolo de 9.600-plex semianidado. Los detalles del Experimento 10 fueron muy similares al Experimento 9, siendo la excepción el protocolo de anidado, e incluyendo la identidad de las cuatro muestras. Las determinaciones del estado de ploidía para los 28 cromosomas en el conjunto se determinaron correctamente con certezas superiores al 99,7%. 7,6 millones (91%) de lecturas se correlacionaron con el genoma, y 6,3 millones (80%) de lecturas se correlacionaron con los SNPs focalizados. La profundidad de lectura media fue de 751, y la mediana de profundidad de lectura fue de 396.

Experimento 11

En un experimento, tres muestras de plasma materno se dividieron en cinco partes iguales, y cada parte se amplificó utilizando o bien 2.400 cebadores multiplexados (cuatro partes) o 1.200 cebadores multiplexados (una parte) y se amplificaron utilizando un protocolo semianidado, para un total de 10.800 cebadores. Después de la amplificación, las partes se agruparon para la secuenciación. Los detalles del Experimento 11 fueron muy similares al Experimento 9, siendo la excepción el protocolo de anidado, y el planteamiento de división y reagrupación. Las determinaciones del estado de ploidía para los 21 cromosomas en el conjunto se determinaron correctamente con certezas superiores al 99,1% excepto por una determinación fallida donde la certeza fue del 83%. 3,4 millones de lecturas se correlacionaron con los SNPs focalizados, la profundidad de lectura media fue de 404 y la mediana de profundidad de lectura fue de 258.

Experimento 12

En un experimento, cuatro muestras de plasma materno se dividieron en cuatro partes iguales, y cada parte se amplificó utilizando 2.400 cebadores multiplexados y se amplificó utilizando un protocolo semianidado, para un total de 9.600 cebadores. Después de la amplificación, las partes se agruparon para la secuenciación. Los detalles del Experimento 12 fueron muy similares al Experimento 9, siendo la excepción el protocolo de anidado, y el planteamiento de división y reagrupación. Las determinaciones del estado de ploidía para los 28 cromosomas en el conjunto se determinaron correctamente con certezas superiores al 97% excepto por una determinación fallida donde la certeza fue del 78%. 4,5 millones de lecturas se correlacionaron con los SNPs focalizados, la profundidad de lectura media fue de 535 y la mediana de profundidad de lectura fue de 412.

Experimento 13

En un experimento, se prepararon cuatro muestras de plasma materno y se amplificaron utilizando un protocolo triplemente hemianidado de 9.600-plex, para un total de 9.600 cebadores. Los detalles del Experimento 12 fueron muy similares al Experimento 9, siendo la excepción el protocolo de anidado que supuso tres rondas de amplificación; las tres rondas suponían 15, 10 y 15 ciclos de STA respectivamente. Las determinaciones del estado de ploidía para 27 de 28 cromosomas en el conjunto se determinaron correctamente con certezas superiores al 99,9% excepto por una que se determinó correctamente con un 94,6%, y una determinación fallida con una certeza de un 80,8%. 3,5 millones de lecturas se correlacionaron con los SNPs focalizados, la profundidad de lectura media fue de 414 y la mediana de profundidad de lectura fue de 249.

Experimento 14

En un experimento se amplificaron 45 conjuntos de células utilizando un protocolo semianidado de 1.200-plex, se secuenciaron, y se realizaron determinaciones de ploidía en tres cromosomas. Cabe señalar que este experimento pretende simular las condiciones de realizar un diagnóstico genético previo a la implantación en biopsias de una sola célula de embriones de 3 días, o biopsias de trofotodermo de embriones de 5 días. 15 células únicas individuales y 30 conjuntos de tres células se colocaron en 45 tubos de reacción individuales para un total de 45 reacciones donde cada reacción contenía células de una sola línea celular, pero las diferentes reacciones contenían células de diferentes líneas celulares. Las células se prepararon en 5 ul de tampón de lavado y se lisaron añadiendo 5 ul de tampón de lisis ARCTURUS PICOPURE (APPLIED BIOSYSTEMS) y se incubaron a 56 °C durante 20 min, 95 °C durante 10 min.

Se amplificó el ADN de la célula individual/las tres células con 25 ciclos de STA (95 °C durante 10 min para la activación de la polimerasa inicial, a continuación, 25 ciclos de 95 °C durante 30 s; 72 °C durante 10 s; 65 °C durante 1 min; 60 °C durante 8 min; 65 °C durante 3 min; y 72 °C durante 30 s; y una ampliación final a 72 °C durante 2 min;) utilizando una concentración del cebador de 50 nM de 1200 cebadores directos específicos diana y cebadores inversos etiquetados.

El protocolo para PCR semianidada suponía la segunda amplificación de tres paralelas de una dilución del primer producto de STAs durante 20 ciclos de STA (95 °C durante 10 min para la activación de la polimerasa inicial, a continuación, 15 ciclos de 95 °C durante 30 s; 65 °C durante 1 min; 60 °C durante 5 min; 65 °C durante 5 min; y 72 °C durante 30 s; y una ampliación final a 72 °C durante 2 min) utilizando una concentración de cebador específico de etiqueta inversa de 1000 nM, y una concentración de 60 nM para cada uno de los 400 cebadores directos anidados específicos diana. De este modo, en las tres reacciones paralelas de 400-plex se amplificó el total de 1200 dianas amplificadas en la primera STA. A continuación, se amplificó una alícuota de productos STA mediante PCR estándar durante 15 ciclos con 1 uM de cebadores directos específicos etiquetados y cebadores inversos codificados mediante código de barras para generar bibliotecas de secuenciación codificadas mediante código de barras. Se mezcló una alícuota de cada biblioteca con bibliotecas de diferentes códigos de barras y se purificó utilizando una columna de centrifugación.

De este modo, se utilizaron 1.200 cebadores en las reacciones de una sola célula; los cebadores se diseñaron para focalizar SNPs encontrados en los cromosomas 1, 21 y X. A continuación, se secuenciaron los amplicones utilizando un secuenciador ILLUMINA GAIIIX. Por muestra, el secuenciador generó aproximadamente 3,9 millones de lecturas, con de 500.000 a 800.000 millones de lecturas que se correspondieron con el genoma; (de un 74% a un 94% de todas las lecturas por muestra).

Las muestras relevantes de ADN genómico materno y paterno de las líneas celulares se analizaron utilizando el mismo conjunto de ensayo semianidado de 1200-plex con un protocolo similar con menos ciclos y una segunda STA de 1200-plex, y se secuenciaron.

Los datos de secuenciación se analizaron utilizando métodos informáticos divulgados en el presente documento y se determinó el estado de ploidía en los tres cromosomas para las muestras.

La Figura 24 muestra ratios de profundidad de lectura normalizados (eje vertical) para seis muestras en tres cromosomas (1 = crom 1; 2 = crom 21; 3 = crom X). Los ratios se establecieron para que fueran iguales al número de lecturas que se correspondían con ese cromosoma, normalizado, y dividido por el número de lecturas que se correlacionan con ese cromosoma promediado en tres pocillos cada uno de los cuales contiene tres células 46XY. Los tres conjuntos de puntos de datos correspondientes a las reacciones 46XY está previsto que tengan ratios de 1:1. Los tres conjuntos de puntos de datos correspondientes a las células 47XX+21 está previsto que tengan ratios de 1:1 para el cromosoma 1, 1.5:1 para el cromosoma 21, y 2:1 para el cromosoma X.

La Figura 25 muestra ratios de alelos representados gráficamente para tres cromosomas (1, 21, X) para tres reacciones. La reacción en la parte inferior izquierda muestra una reacción en tres células 46XY. La región izquierda representa los ratios de alelos para el cromosoma 1, la región central son los ratios de alelos para el cromosoma 21, y la región derecha son los ratios de alelos para el cromosoma X. Para las células 46XY, para el cromosoma 1 se prevé ver ratios de 1, 0,5 y 0, correspondientes a los genotipos AA, AB y BB SNP. Para las células 46XY, para el cromosoma 21 se prevé ver ratios de 1, 0,5 y 0, correspondientes a los genotipos AA, AB y BB SNP. Para las células 46XY, para el cromosoma X se prevé ver ratios de 1 y 0, correspondientes a los genotipos A, y B SNP. La reacción en la parte inferior derecha muestra una reacción en tres células 47XX+21. Los ratios de alelos son aislados por el cromosoma como en el gráfico de la parte inferior izquierda. Para las células 47XX+21, para el cromosoma 1 se prevé ver ratios de 1, 0,5 y 0, correspondientes a los genotipos AA, AB y BB SNP. Para las células 47XX+21, para el cromosoma 21 se prevé ver ratios de 1, 0,67, 0,33 y 0, correspondientes a los genotipos AAA, AAB, ABB y BBB SNP. Para las células 47XX+21, para el cromosoma X se prevé ver ratios de 1, 0,5 y 0, correspondientes a los genotipos AA, AB y BB SNP. El gráfico en la parte superior derecha se realizó en una reacción que comprende 1 ng de ADN genómico de la línea celular 47XX+21. La **Figura 26** muestra los mismos gráficos que la Figura 25, pero para reacciones realizadas en una sola célula. El gráfico de la izquierda fue una reacción que contenía una célula 47XX+21, y el gráfico de la derecha fue una reacción que contenía una célula 46XX.

De los gráficos que se muestran en la Figura 25 y en la Figura 26, es visualmente aparente que hay dos agrupaciones de puntos para los cromosomas donde se prevé ver ratios de 1 y 0; tres agrupaciones de puntos para los cromosomas donde se prevé ver ratios de 1, 0,5, y 0, y cuatro agrupaciones de puntos para los cromosomas donde se prevé ver ratios de 1, 0,67, 0,33 y 0. El algoritmo de Parental Support pudo realizar determinaciones correctas en los tres cromosomas para todas las 45 reacciones.

Experimento 15

En un experimento, se prepararon muestras de plasma materno y se amplificaron utilizando un protocolo de 19.488-plex hemianidado. Las muestras se prepararon del modo siguiente: se centrifugaron hasta 20 mL de sangre materna para aislar la capa leucocitaria y el plasma. El ADN genómico en la muestra materna se preparó de la capa leucocitaria y el ADN paterno se preparó a partir de una muestra de sangre o una muestra de saliva. El ADN libre de células en el plasma materno se aisló utilizando el kit de ácido nucleico circulante QIAGEN y se eluyó en 50 uL de tampón de TE en función de las instrucciones del fabricante. Se anexaron adaptadores de unión universales al extremo de cada molécula de 40 uL de ADN del plasma purificado y las bibliotecas se ampliaron durante 9 ciclos utilizando cebadores específicos

del adaptador. Las bibliotecas se purificaron con perlas AGENCOURT AMPURE y se eluyeron en 50 µl de tampón de suspensión de ADN. Se amplificaron 6 µl del ADN con 15 ciclos de STAR 1 (95 °C durante 10 min_ para la activación de la polimerasa inicial, a continuación, 15 ciclos de 96 ° C durante 30 s; 65 °C durante 1 min; 58 °C durante 6 min; 60 °C durante 8 min; 65 °C durante 4 min; y 72 °C durante 30 s; y una ampliación final a 72 °C durante 2 min) utilizando una concentración del cebador de 7,5 nM de 19.488 cebadores inversos etiquetados específicos diana y un cebador directo específico del adaptador de biblioteca a 500 nM.

El protocolo para PCR hemianidada suponía una segunda amplificación de una dilución del producto STAR 1 durante 15 ciclos (STAR 2) (95 °C durante 10 min para la activación de la polimerasa inicial, a continuación, 15 ciclos de 95 °C durante 30 s; 65 °C durante 1 min; 60 °C durante 5 min; 65 °C durante 5 min y 72 °C durante 30 s; y una ampliación final a 72 °C durante 2 min) utilizando una concentración de etiqueta inversa de 1000 nM, y una concentración de 20 nM para cada uno de los 19.488 cebadores directos específicos diana.

A continuación, se amplificó una alícuota de productos STAR 2 mediante PCR estándar durante 12 ciclos con 1 µM de cebadores directos específicos etiquetados y cebadores inversos codificados mediante código de barras para generar bibliotecas de secuenciación codificadas mediante código de barras. Se mezcló una alícuota de cada biblioteca con bibliotecas de diferentes códigos de barras y se purificó utilizando una columna de centrifugación.

De este modo, se utilizaron 19.488 cebadores en las reacciones de un único pocillo; los cebadores se diseñaron para focalizar SNPs encontrados en los cromosomas 1, 2, 13, 18, 21, X e Y. A continuación, se secuenciaron los amplicones utilizando un secuenciador ILLUMINA GAIIIX. Para las muestras de plasma, el secuenciador generó aproximadamente 10 millones de lecturas, con 9,4-9,6 millones de lecturas correlacionadas con el genoma; (94-96 %), y de esas, 99,95 % correlacionadas con los SNPs focalizados con una profundidad media de lectura de 460 y una mediana de profundidad de lectura de 350. Con fines comparativos, una distribución perfectamente uniforme sería: 10M lecturas / 19.488 dianas = 513 lecturas/diana. Para los dímeros de cebadores, 30.000 lecturas fueron de dímeros de cebadores secuenciados (un 0,3% de las lecturas se generaron mediante el secuenciador). Para muestras genómicas, 99,4-99,7% de las lecturas se correlacionaron con el genoma, de aquellas, 99,99% de las lecturas se correlacionaron con los SNPs focalizados, y 0,1 % de las lecturas generadas por el secuenciador fueron dímeros de cebadores.

Para las muestras de plasma con 10 millones de lecturas de secuenciación, de forma típica al menos 19.350 de los 19.488 SNPs focalizados (99,3 %) se amplifican y secuencian. Para las muestras de plasma con 2 millones de lecturas de secuenciación, de forma típica al menos 19.000 SNPs focalizados (97,5%) se amplifican y secuencian. El número más bajo puede ser debido a ruido en el muestreo ya que el número de lecturas es inferior y el secuenciador omite algunos de los productos amplificados. Si se desea, se puede aumentar el número de lecturas de secuenciación para aumentar el número de SNPs focalizados que se amplifican y secuencian.

Las muestras de ADN genómico materno y paterno relevantes se amplificaron utilizando 19.488 cebadores directos exteriores y cebadores inversos etiquetados a 7,5 nM en el STAR 1. Las condiciones termocíclicas y la composición de STAR 2, y la PCR mediante codificación por medio de código de barras eran las mismas que para el protocolo hemianidado.

Se descubrió que la fracción fetal media para 407 muestras era del 14,8%. Los datos de secuenciación se analizaron utilizando métodos informáticos divulgados en el presente documento y se determinó el estado de ploidía en cuatro cromosomas (13, 18, 21, Y) para los fetos cuyo ADN estaba presente en 378 de las 407 muestras de plasma materno, y en el cromosoma X en 375 de las 407 muestras de plasma materno. Las determinaciones del estado de ploidía para los 1.887 cromosomas en el conjunto se determinaron correctamente con certezas superiores al 90%. 1882 de las 1887 determinaciones eran superiores al 95%; y 1.862 de las 1.887 determinaciones se determinaron con una certeza superior al 99%.

Un experimento de control similar se llevó a cabo utilizando agua en lugar de ADN extraído del plasma en el protocolo de PCR del plasma. En base a seis de dichos ensayos de un experimento, 5-6% de las lecturas secuenciadas fueron dímeros de cebadores. Otras lecturas secuenciadas fueron debidas a ruido de fondo. Este experimento demuestra que incluso en ausencia de una muestra de ácido nucleico con loci diana para que los cebadores se hibriden (en lugar de hibridarse en otros cebadores y formar dímeros de cebadores amplificados) se forman pocos dímeros de cebadores.

Experimento 16

El siguiente experimento ilustra un ejemplo de método para diseñar y seleccionar una biblioteca de cebadores que se pueden utilizar en cualquiera de los métodos de PCR multiplexada de la invención.

El objetivo es seleccionar cebadores de una biblioteca de cebadores candidatos inicial que se puede utilizar para amplificar simultáneamente un gran número de loci diana (o un subconjunto de loci diana) en una sola reacción. Para un conjunto inicial de loci diana candidatos, los cebadores no tenían que diseñarse o seleccionarse para cada locus diana. Los cebadores se diseñan y seleccionan preferiblemente para una gran parte de los loci diana más convenientes.

Paso 1

Un conjunto de loci diana candidatos (como los SNP) se seleccionaron en base a información disponible públicamente sobre parámetros deseados para los loci diana, como frecuencia de los SNPs dentro de una población diana o la tasa de heterocigosidad de los SNPs (página web ncbi.nlm.nih.gov/projects/SNP/; Sherry ST, Ward MH, Kholodov M, et al.

dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 1 de enero de 2001;29(1):308-11). Para cada locus candidato, uno o más pares de cebadores para PCR se diseñaron utilizando el programa Primer3 (página web primer3.sourceforge.net; libprimer3 versión 2.2.3). Si no había diseños factibles para los cebadores para PCR para un locus diana determinado, ese locus diana se eliminaba de cualquier consideración ulterior.

- 5 Si se desea, se puede calcular una "puntuación de locus diana" (puntuación más alta que representa una mayor deseabilidad) para la mayoría o todos los loci diana, como una puntuación de locus diana calculada en base a una media ponderada de distintos parámetros deseados para los loci diana. A los parámetros se les puede asignar diferentes pesos en base a su importancia para la aplicación particular para la que se utilizarán los cebadores. Entre los ejemplos de parámetros se incluye la tasa de heterocigosidad del locus diana, el predominio de la enfermedad asociado a una secuencia (por ejemplo, un polimorfismo) en el locus diana, la penetrancia de la enfermedad asociada con una secuencia (por ejemplo, un polimorfismo) en el locus diana, la especificidad del cebador o cebadores candidatos utilizados para amplificar el locus diana, el tamaño del cebador o cebadores candidatos utilizado para amplificar el locus diana, y el tamaño del amplicón diana.

Paso 2

- 15 Se calculó una puntuación de interacción termodinámica entre cada cebador y todos los cebadores para todos los otros loci diana del paso 1 (véase, por ejemplo, Allawi, H. T. & SantaLucia, J., Jr. (1998), "Thermodynamics of Internal C-T Mismatches in DNA", Nucleic Acids Res. 26, 2694-2701; Peyret, N., Seneviratne, P. A., Allawi, H. T. & SantaLucia, J., Jr. (1999), "Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A-A, C-C, G-G, and T-T Mismatches", Biochemistry 38, 3468-3477; Allawi, H. T. & SantaLucia, J., Jr. (1998), "Nearest-Neighbor Thermodynamics of Internal A-C Mismatches in DNA: Sequence Dependence and pH Effects", Biochemistry 37, 9435-9444.; Allawi, H. T. & SantaLucia, J., Jr. (1998), "Nearest Neighbor Thermodynamic Parameters for Internal G-A Mismatches in DNA", Biochemistry 37, 2170-2179; & Allawi, H. T. & SantaLucia, J., Jr. (1997), "Thermodynamics and NMR of Internal G-T Mismatches in DNA", Biochemistry 36, 10581- 10594; MultiPLX 2.1 (Kaplinski L, Andreson R, Puurand T, Remm M. MultiPLX: automatic grouping and evaluation of PCR primers. Bioinformatics. 15 de abril de 2005;21(8): 1701-2). Este paso dio como resultado una matriz en 2D de puntuaciones de interacción. La puntuación de interacción predijo la probabilidad de dímeros de cebadores en la que estaban implicados los dos cebadores que interactuaban. La puntuación se calculó del siguiente modo:

$$\text{Puntuación_interacción} = \text{máx}(-\text{deltaG_2}, 0.8 * (-\text{deltaG_1}))$$

donde

- 30 deltaG_2 = energía Gibbs (energía necesaria para romper el dímero) para un dímero que es extensible mediante PCR en ambos extremos, es decir, el extremo 3' de cada cebador se reforma térmicamente en el otro cebador; y deltaG_1 = energía Gibbs para un dímero que es extensible mediante PCR en al menos un extremo.

Paso 3:

- 35 Para cada locus diana, si había más de un diseño de par de cebadores, se seleccionó un diseño utilizando el método siguiente:

- 1 Para cada diseño de par de cebadores para el locus, encuentre la peor puntuación de interacción (la más alta) para los cebadores en ese diseño y todos los cebadores de todos los diseños para todos los otros loci diana.
- 2 Seleccione el diseño con la mejor puntuación de interacción (la más baja).

Paso 4

- 40 Se creó un gráfico de modo que cada nodo representaba un locus y su diseño de par de cebadores asociado (por ejemplo, un problema Maximal Clique). Se creó un borde entre cada par de nodos. Se asignó un peso a cada borde igual a la peor puntuación de interacción (más alta) entre los cebadores asociados a los dos nodos conectados por el borde.

Paso 5

- 45 Si se desea, para cada par de diseños para dos loci diana diferentes donde uno de los cebadores de un diseño y uno de los cebadores del otro diseño se reformarían térmicamente en regiones diana con solapamiento, se añadió un borde adicional entre los nodos para los dos diseños.

El peso de estos bordes se estableció igual al peso más alto asignado en el paso 4. Por tanto, el paso 5 evita que la biblioteca tenga cebadores que se reformarían térmicamente en regiones diana con solapamiento y, de este modo,

- 50 interfieran entre sí durante una reacción por PCR multiplexada.

Paso 6

Se calculó un umbral de puntuación de interacción inicial del siguiente modo: $\text{umbral_peso} = \text{máx}(\text{peso_borde}) - 0,05 * (\text{máx}(\text{peso_borde}) - \text{mín}(\text{peso_borde}))$ donde

máx(peso_borde) es el peso del borde máximo en el gráfico; y mín(peso_borde) es el peso del borde mínimo en el gráfico. Los límites iniciales para el umbral se establecieron del modo siguiente: máx_umbral_peso = máx(peso_borde)
mín_umbral_peso = mín(peso_borde)

Paso 7

- 5 Se creó un nuevo gráfico compuesto por el mismo conjunto de nodos que el gráfico del paso 5, que sólo incluía bordes con pesos que superaban el umbral_peso. Por tanto, el paso ignora las interacciones con puntuaciones iguales o inferiores al umbral de peso.

Paso 8

- 10 Los nodos (y todos los bordes conectados a los nodos extraídos) se extrajeron del gráfico del paso 7 hasta que no quedaron bordes. Los nodos se extrajeron aplicando el siguiente procedimiento repetidamente:

- 1 Hallar el nodo con el grado más alto (número de bordes más alto). Si hay más de uno entonces seleccionar uno arbitrariamente.
- 2 Definir el conjunto de nodos compuesto del nodo seleccionado anteriormente y de todos los nodos conectados a éste, pero excluyendo cualquier nodo que tenga un grado inferior al nodo seleccionado anteriormente.
- 15 3 Elegir el nodo del conjunto que tiene la puntuación de locus diana más baja (una puntuación más baja representa una menor deseabilidad) del paso 1. Extraer el nodo del gráfico.

Paso 9

Si el número de nodos que quedan en el gráfico satisface el número requerido de loci diana para el conjunto de PCR multiplexada (dentro de una tolerancia aceptable), entonces se continuó con el método en el paso 10.

- 20 Si había muchos o pocos nodos que quedaban en el gráfico, se llevó a cabo una búsqueda binaria para determinar qué valores de umbral darían como resultado en el número deseado de nodos que quedase en los gráficos. Si había demasiados nodos en el gráfico, los límites del umbral de peso se ajustaron del modo siguiente:

máx_umbral_peso = umbral_peso

- 25 En caso contrario (si hay demasiado pocos nodos en el gráfico, los límites del umbral de peso se ajustaron del modo siguiente:

mín_umbral_peso = umbral_peso

A continuación, el umbral de peso se ajustó del modo siguiente:

umbral_peso = (máx_umbral_peso + mín_umbral_peso) / Se repitieron 2 pasos.

Paso 10

- 30 Los diseños de par de cebadores asociados con los nodos que quedan en el gráfico se seleccionaron para la biblioteca de cebadores. Esta biblioteca de cebadores se puede utilizar en cualquiera de los métodos de la invención.

Si se desea, este método de diseñar y seleccionar cebadores se puede realizar para bibliotecas de cebadores en las que se utiliza un solo cebador (en lugar de un par de cebadores) para la amplificación de un locus diana. En este caso, un nodo presenta un cebador por locus diana (en lugar de un par de cebadores).

35 Experimento 17

La Figura 27 es un gráfico en el que se comparan dos bibliotecas de cebadores diseñadas utilizando los métodos de la invención. Este gráfico muestra el número de loci con una determinada frecuencia de alelos menor que son focalizados por cada biblioteca de cebadores. Durante la selección de la biblioteca del "nuevo conjunto", se retuvieron más cebadores. Esta biblioteca permite la amplificación de más loci diana, especialmente loci diana con frecuencias alélicas menores relativamente grandes (que son los alelos más informativos para algún método de la invención como, por ejemplo, para la detección de anomalías cromosómicas fetales).

- 40 Estas bibliotecas de cebadores se utilizaron en el siguiente método de PCR multiplexada. Se recogió sangre (20- 40 mL) de cada sujeto en de dos a cuatro tubos de ADN CELL-FREE™ (Streck). Se aisló plasma (un mínimo de 7 mL) de cada muestra a través de un protocolo de doble centrifugación de 2.000 g durante 20 min, seguido de 3.220 g durante 30 min, con una transferencia de sobrenadante después de la primera centrifugación. El ADN libre de células se aisló del plasma de 7-20 mL utilizando el kit ácido nucleico circulante QIAGEN QIAamp y se eluyó en 45 uL del tampón de TE. El ADN genómico materno puro se aisló de la capa leucocitaria obtenida después de la primera centrifugación, y el ADN genómico paterno puro se preparó de forma similar a partir de una muestra de sangre, saliva o bucal.

- 50 Las muestras de ADN libre de células materno, ADN genómico materno, y ADN genómico paterno se preamplificaron durante 15 ciclos utilizando 11.000 ensayos diana específicos y se transfirió una alícuota a una segunda reacción por PCR de 15 ciclos utilizando cebadores anidados. Finalmente, se prepararon muestras para la secuenciación añadiendo etiquetas codificadas mediante código de barras en una tercera ronda de 12 ciclos de PCR. Por tanto, se amplificaron 11.000 dianas en una sola reacción; las dianas incluían SNPs encontrados en los cromosomas 13, 18, 21, X, e Y. A

continuación, se secuenciaron los amplicones utilizando un secuenciador ILLUMINA GAIIX o HISEQ. Los genotipos parentales se secuenciaron a una profundidad de lectura más baja (~20% de la profundidad de lectura del ADN libre de células) que los genotipos fetales.

Experimento 18

5 Si se desea, el tamaño y cantidad de los productos de PCR se puede analizar utilizando métodos estándar, como el uso del Bioanalizador 2100 de Agilent Technologies (Figura 28A-M). Por ejemplo, los métodos de PCR directa que se describe en el presente documento sin anidado se utilizaron en experimentos de 2.400-plex (Figuras 28B-28G) y 19.488-plex (Figuras 28H a 28M). La cantidad de cebador fue 10 nM para las Figuras 28B-28D y 28H a 28 J. La cantidad de cebador fue 1 nM para las Figuras 28E-28G y 28K a 28M. La cantidad de ADN aportado fue de 24 ng para las Figuras 28B, 28E, 28H, y 28K; 80 ng para las Figuras 28C, 28½ f, 28I, y 28L; y 250 ng para las Figuras 28D, 28G, 28 J, y 28M. Más ADN aportado dio como resultado una mayor proporción del producto de 180 pares de bases deseado. El pico a 140 pares de bases es un producto de dímeros de cebadores.

Experimento 19

15 Un estudio de prueba de principio demostró la detección de T13, T18, T21, 45,X, y 47,XXY con precisiones igualmente altas en todos los cromosomas.

Pacientes

Las parejas embarazadas se inscribieron en centros específicos de atención prenatal bajo protocolos aprobados por una Junta de Revisión Institucional en conformidad con las leyes locales. Los criterios de inclusión fueron tener al menos 18 años de edad, una edad gestacional de al menos nueve semanas, embarazos únicos, y la firma de un consentimiento por parte del paciente sobre los beneficios y riesgos involucrados. Se extrajeron muestras de sangre de madres embarazadas, y una muestra de sangre o bucal del padre. Se seleccionaron muestras de 2 embarazos con T13 (síndrome de Patau), 2 con T18 (síndrome de Edwards), 2 con T21 (síndrome de Down), 2 con 45,X, 2 con 47,XXY, y 90 embarazos normales antes de realizar una prueba de una cohorte de 500 mujeres para determinar qué anomalías cromosómicas detecta el método. El cariotipo fetal normal se confirmó mediante el establecimiento del cariotipo molecular para las muestras donde el tejido del niño posterior al nacimiento estaba disponible. La muestra euploide se extrajo antes de la prueba invasiva de mujeres de bajo riesgo. Se extrajeron muestras aneuploides al menos 7 días después de realizar la prueba invasiva y se confirmó la aneuploidía mediante el establecimiento del cariotipo citogenético o hibridación in situ de fluorescencia en laboratorios independientes.

Preparación de la muestra y PCR multiplexada

30 Para los datos en las Figuras 30A-E, 30G, 30H, y 31A-31G, se realizó la preparación de la muestra y PCR de 19.488-plex como se describe en el Experimento 15. Para los datos de la **Figura 30F**, la preparación de la muestra y la PCR de 11.000 plex se llevaron a cabo como se describe en el Experimento 17.

Metodología y análisis de datos

35 El algoritmo considera los genotipos parentales y los datos de frecuencia de cruces (como datos de la base de datos HapMap) para calcular las distribuciones alélicas previstas para 19.488 loci polimórficos para un número muy grande de estados de ploidía fetal posibles, y a distintas fracciones de ADN libre de células fetales (Figuras 29A-29C). A diferencia de los métodos basados en el ratio de alelos, también tiene en cuenta el desequilibrio del enlace y utiliza modelos de datos no gaussianos para describir la distribución prevista de mediciones alélicas en un SNP dadas las características de la plataforma y los sesgos de amplificación observados. A continuación, compara las distintas distribuciones alélicas previstas con las distribuciones alélicas reales medidas en la muestra de ADN libre de células (Figura 29C), y calcula la probabilidad de cada hipótesis (monosomía, disomía, o trisomía, para la que hay numerosas hipótesis en base a los distintos cruces potenciales) en base a los datos de secuenciación. El algoritmo suma las probabilidades de cada hipótesis de monosomía, disomía, o trisomía individual (Figura 29D), y determina el estado de ploidía con la probabilidad global máxima como el número de copia y fracción fetal (Figura 29E). Aunque el cariotipo de muestra no fue ciego para los investigadores de laboratorio, el algoritmo determinó los estados de ploidía sin intervención humana y fue ciego con respecto a la veracidad.

Interpretación de datos

Representaciones gráficas de los datos generados

50 Para determinar el estado de ploidía de un cromosoma de interés, el algoritmo considera la distribución de los recuentos de secuencia de cada uno de dos alelos posibles de 3.000 a 4.000 SNPs por cromosoma. Es importante señalar que el algoritmo realiza determinaciones del estado de ploidía utilizando un planteamiento que no se presta a la visualización. Por tanto, con fines ilustrativos, los datos se muestran aquí de forma simplificada como ratios de los dos alelos más probables, etiquetados como A y B de modo que las tendencias relevantes se puedan visualizar más fácilmente. Esta ilustración simplificada no tiene en cuenta algunas de las características del algoritmo. Por ejemplo, dos aspectos importantes del algoritmo que no es posible ilustrar con un método de visualización que muestra ratios de alelos son: 1) la capacidad de aprovechar el desequilibrio del enlace, es decir, la influencia que una medición en un SNP tiene en la identidad probable de un SNP contiguo, y 2) el uso de modelos de datos no gaussianos que describen la distribución prevista de mediciones alélicas en un SNP dadas las características de la plataforma y los sesgos de amplificación.

También cabe señalar que el algoritmo sólo considera los dos alelos más comunes en cada SNP, omitiendo otros posibles alelos.

Las representaciones gráficas en la Figura 30A-30H incluyen muestras para las que dos, uno, o tres cromosomas fetales están presentes. Generalmente, estas indican euploidía (Figuras 30A-30C) monosomía (Figura 30D), y trisomía (Figuras 30E-30H), respectivamente. En todos los gráficos, cada punto representa un único SNP, donde los SNPs focalizados se representan gráficamente de forma secuencial de izquierda a derecha para un cromosoma a lo largo de los ejes horizontales. Los ejes verticales indican el número de lecturas para el alelo A como una fracción del número total de lecturas para los alelos A y B para ese SNP. Cabe señalar que las mediciones se realizan en ADN libre de células total aislado de la sangre materna, y el ADN libre de células incluye tanto ADN libre de células materno como fetal; por tanto, cada punto representa la combinación de la contribución del ADN materno y fetal para ese SNP. Por tanto, aumentar la proporción de ADN libre de células materno de un 0% a un 100% cambiará gradualmente algunos puntos arriba o abajo dentro de los gráficos, dependiendo del genotipo materno y fetal. Esto se describe con mayor detalle a continuación con los gráficos correspondientes. Si se desea facilitar la visualización, los puntos se pueden codificar por colores en función del genotipo materno, ya que el genotipo materno contribuye más a la localización de cada punto y la mayoría de trisomías se heredan de la madre; esto contribuye a la visualización de los estados de ploidía. Específicamente, los SNPs para los cuales el genotipo materno es AA se pueden indicar en rojo, aquellos para los que el genotipo materno es AB se pueden indicar en verde, y aquellos para los que el genotipo materno es BB se pueden indicar en azul.

En todos los casos, los SNPs que son homocigotos para el alelo A (AA) tanto en la madre como en el feto se encuentran estrechamente asociados con el límite superior de los gráficos, ya que la fracción de lecturas de alelo A es alta porque no debería haber alelos B presentes. Por el contrario, los SNPs que son homocigotos para el alelo B tanto en la madre como en el feto se encuentran estrechamente asociados con el límite inferior de los gráficos, ya que la fracción de lecturas de alelo A es baja porque debería haber sólo alelos B. Los puntos que no están estrechamente asociados con los límites superior e inferior de los gráficos representan SNPs para los que la madre, el feto, o ambos son heterocigotos; estos puntos son útiles para identificar la ploidía fetal, pero también pueden ser útiles para determinar la herencia paterna frente a la materna. Estos puntos se aíslan en base a los genotipos maternos y fetales y la fracción fetal, y como tal la posición precisa de cada punto individual a lo largo del eje y depende de la estequiometría y la fracción fetal. Por ejemplo, los loci donde la madre es AA y el feto es AB se prevé que tengan una fracción diferente de lecturas de alelo A y, por tanto, una posición diferente a lo largo del eje y, en función de la fracción fetal.

Dos cromosomas presentes

Las Figuras 30A-30C muestran datos que indican la presencia de dos cromosomas cuando la muestra es completamente materna (no hay ADN libre de células fetal presente, Figura 30A), contiene una fracción de ADN libre de células fetal moderada (Figura 30B), o contiene una fracción de ADN libre de células fetal alta (Figura 30C).

La Figura 30A muestra datos obtenidos de ADN libre de células aislado de la sangre de una mujer no embarazada. Cuando hay un número de ADN libre de células fetal presente y la muestra contiene sólo ADN libre de células materno, los gráficos representan puramente el genotipo materno euploide; el patrón distintivo incluye "agrupaciones" de puntos: una agrupación roja estrechamente asociada con la parte superior del gráfico (SNPs donde el genotipo materno es AA), una agrupación azul estrechamente asociada con la parte inferior del gráfico (SNPs donde el genotipo materno es BB), y una agrupación verde central única (SNPs donde el genotipo materno es AB) (no se muestra el color). Cuando el ADN libre de células fetal está presente, la ubicación de los puntos cambia de modo que las agrupaciones se aíslan en "bandas" discretas. Cabe señalar que para las muestras con una fracción fetal de 0%, se hace referencia a las agrupaciones de puntos como "grupos" (como en la Figura 30A), y para todas las muestras con una fracción fetal de >0%, se hace referencia a las agrupaciones de puntos como "bandas" (como en las Figuras 30B- 30J). Si la fracción fetal es lo suficientemente alta, estas bandas discretas serán fácilmente visibles.

Específicamente, las Figuras 30B y 30C demuestran el patrón característico asociado con dos cromosomas fetales presentes en fracciones fetales moderadas y altas, respectivamente. Este patrón incluye tres bandas verdes centrales que corresponden a SNPs que son heterocigotos en la madre, y dos bandas "periféricas" cada una de ellas tanto en la parte superior (roja) como en la parte inferior (azul) de los gráficos que corresponden a SNPs que son homocigotos en la madre (no se muestra el color).

La Figura 30B muestra datos obtenidos de ADN libre de células aislado a partir de una muestra de plasma de una mujer portadora de un feto euploide y con un 12% de fracción de ADN libre de células fetal. Aquí, las agrupaciones de puntos estrechamente asociados con la parte superior y la parte inferior del gráfico se aíslan en dos bandas discretas cada una: una banda periférica externa roja y una azul que permanecen estrechamente asociadas con el límite superior o inferior de los gráficos, y una banda periférica interna roja y una azul que se ha separado de los límites de los gráficos (no se muestra el color). Estas bandas periféricas internas, centradas alrededor de 0,92 y 0,08, representan SNPs para los que el genotipo materno es AA y el genotipo fetal es AB (indicado en rojo), y SNPs para los que el genotipo materno es BB y el genotipo fetal es AB (indicado en azul), respectivamente. La agrupación central de puntos verdes se ensancha, pero en esta fracción fetal la separación en bandas diferentes no es fácilmente visible.

A una fracción de ADN libre de células fetal alta, el patrón típico que indica la presencia de dos cromosomas (un trío de bandas verdes así como dos bandas rojas y dos azules periféricas) se puede ver rápidamente (no se muestra el color).

La Figura 30C muestra datos obtenidos de una muestra de plasma de una mujer portadora de un feto euploide y con una fracción de ADN libre de células fetal del 26%. Aquí, las bandas periféricas se han separado de tal manera que la banda interna se ha trasladado hacia el centro del gráfico debido a los niveles alterados de alelos B de la fracción de ADN libre de células fetal que ha aumentado. Significativamente, a fracciones fetales más altas, la separación de la agrupación verde central en tres bandas diferentes ahora se puede ver rápidamente. Este trío central de bandas, en este caso agrupándose alrededor de 0,37, 0,50 y 0,63, corresponde a aquellos SNPs donde el genotipo materno es AB, y el genotipo fetal es AA (parte superior), AB (parte central) y BB (parte inferior). Estos patrones distintivos, concretamente las tres bandas verdes y las cuatro bandas periféricas (dos rojas y dos azules), indican la presencia de dos cromosomas, como en la euploidia autosómica o para el cromosoma X en un feto (XX) de mujer.

Un cromosoma presente

Cuando el feto sólo hereda un único cromosoma y, por tanto, sólo hereda un único alelo, la heterocigosidad del feto no es posible. Como tal, las únicas identidades de SNP fetales posibles son A o B. Por tanto, los cromosomas monosómicos heredados de la madre tienen un patrón característico de dos bandas verdes centrales que representan SNPs para los que la madre es heterocigota, y sólo tiene bandas rojas y azules periféricas individuales que representan SNPs para los que la madre es homocigota, y que permanecen estrechamente asociados con los límites superior e inferior de los gráficos (1 y 0), respectivamente (**Figura 30D**) (no se muestra el color). Cabe señalar la ausencia de bandas periféricas internas. Este patrón indica la presencia de un cromosoma, como en la monosomía autosómica heredada de la madre, o para el cromosoma X en un feto (XY) varón.

Tres cromosomas presentes

Los cromosomas trisómicos tienen tres patrones característicos. El primer patrón indica trisomía mitótica heredada de la madre, un error meiótico donde el feto ha heredado dos cromosomas no idénticos homólogos, de la madre (Figura 30E); este patrón incluye dos bandas verdes centrales cada una de las cuales con dos bandas rojas y dos azules periféricas (no se muestra el color). El segundo patrón indica trisomía mitótica heredada del padre, donde el feto ha heredado dos cromosomas no idénticos homólogos del padre (Figura 30F); este patrón incluye cuatro bandas verdes centrales y cada una de las cuales con tres bandas rojas y azules periféricas (no se muestra el color). El tercer patrón indica una trisomía mitótica o bien heredada de la madre (Figura 30G) o heredada del padre (Figura 30H), un error mitótico donde el feto ha heredado dos cromosomas idénticos de la madre o del padre; este patrón incluye cuatro bandas verdes en el centro cada una de las cuales con dos bandas rojas y azules periféricas. Las trisomías mitóticas heredadas del padre y de la madre se pueden distinguir por la colocación de las bandas rojas y azules de flanco, de modo que las bandas rojas y azules periféricas internas (aquellas no asociadas con los límites de los gráficos) están más cerca del centro en la trisomía mitótica heredada del padre (no se muestra el color). Esto es debido a la contribución paterna de cromosomas idénticos. Cabe señalar que nuestros resultados anteriores indican que en la etapa de blastómero, un 66,7% de las trisomías heredadas de la madre son meióticas, y que sólo un 10,2% de las trisomías son heredadas del padre.

Para el cromosoma Y, el método PS considera un conjunto diferente de hipótesis: cero, uno o dos cromosomas presentes. Puesto que no hay contribución materna a las lecturas de secuencia en cada locus y puesto que los loci heterocigotos no son posibles (los casos de dos cromosomas Y necesariamente implican dos cromosomas idénticos), las bandas se mantienen estrechamente asociadas con la parte superior (alelos A) o la parte inferior (alelos B) del gráfico (no se muestran los datos), y el análisis se ve simplificado en gran medida, dependiendo de los datos del recuento de alelos cuantitativo. Cabe señalar que puesto que el método interroga a los SNPs, utiliza SNPs no recombinantes homólogos del cromosoma Y, obteniendo así datos tanto en X como en Y para un par de sondas.

Identificación de aneuploidías

La identificación de las aneuploidías autosómicas utilizando este método de visualización basado en gráficos es sencilla dada una fracción fetal suficiente, y requiere únicamente identificar gráficos para los que hay un número anómalo de cromosomas presente, como se ha descrito anteriormente. La combinación del conocimiento del número de copia de los cromosomas X e Y identifica si hay aneuploidías cromosómicas sexuales presentes. Específicamente, los gráficos que representan un feto con un genotipo 47,XXX tendrán de forma típica un patrón de “tres cromosomas”, y los gráficos que representan un feto con un genotipo 47,XXY tendrán de forma típica el patrón de “dos cromosomas” para el cromosoma X, pero también tendrán lecturas de alelos que indiquen la presencia de un cromosoma Y. El método de forma similar puede determinar 47,XYY, donde un patrón de “un cromosoma” indica la presencia de un solo cromosoma X, y las lecturas de alelos indican la presencia de dos cromosomas Y. Un feto con un genotipo 45,X tendrá el típico patrón de “un cromosoma” para el cromosoma X, y datos que indiquen cero cromosomas Y.

Efectos de la fracción fetal

Como se ha expuesto anteriormente, el número de lecturas de secuencia del feto contribuye a la ubicación precisa de cada punto a lo largo del eje y en los gráficos. Una fracción fetal afectará a la proporción de lecturas que se originan del feto y de la madre, también afectará la posición de cada punto. A una fracción de ADN libre de células fetal alta (por lo general superior a -20%), como en las Figuras 30C-30E y las Figuras 30G y 30H, se puede ver rápidamente que aunque los puntos se agrupan en base principalmente al genotipo materno, la presencia de ADN fetal de alelos cuyo genotipo es diferente del genotipo materno cambia las agrupaciones en múltiples bandas diferentes. Sin embargo, puesto que la fracción fetal disminuye (como en las Figuras 30B y 30F), los puntos retroceden hacia los polos y el centro del gráfico, lo que da como resultado agrupaciones más estrechas. Específicamente, el conjunto de bandas rojas periféricas, donde el genotipo materno es AA, retrocede hacia la parte superior del gráfico; el conjunto de bandas azules periféricas, donde el

genotipo materno es BB, retrocede hacia la parte inferior; el conjunto de bandas verdes centrales, donde la madre es heterocigota, se condensa en una sola agrupación en el centro del gráfico (compare las Figuras 30B y 30C) (no se muestra el color). Aunque la aneuploidía no se puede ver rápidamente a simple vista utilizando esta técnica de visualización para casos de fracciones fetales bajas, el algoritmo puede identificar estados de ploidía con una fracción fetal muy baja, como una fracción fetal del 3%. Puede hacer esto ya que la técnica estadística compara los datos observados con modelos de datos muy precisos que predicen las distribuciones alélicas para un conjunto de parámetros de muestra determinado (incluido el número de copia, los genotipos parentales, y la fracción fetal, por ejemplo). La precisión del modelo de datos es crítica en casos de fracciones fetales bajas, ya que las diferencias entre las distribuciones alélicas para diferentes estados de ploidía son proporcionales a la fracción fetal. Además, el algoritmo puede determinar cuándo un conjunto de datos no contiene datos suficientes para realizar una determinación del estado de ploidía fetal con certeza.

Resultados

Las lecturas de secuenciación que se correlacionaron con los SNPs focalizados se consideraron informativas y se utilizaron mediante el algoritmo. Se observaron más del 95% de loci focalizados en los resultados de secuenciación. Los gráficos para visualizar determinaciones del estado de ploidía clave se muestran en la Figura 31A-31G. La Figura 31A indica una muestra euploide. Aquí, los cromosomas 13, 18, y 21 tienen el típico patrón de “dos cromosomas” (como se describe en el presente documento). Esto incluye un trío de bandas verdes centrales, y dos bandas periféricas rojas y dos azules. Esto, junto con las dos bandas verdes centrales para el cromosoma X y la presencia de bandas de cromosoma Y a lo largo de las periferias de los gráficos, indican un genotipo XY euploide (no se muestra el color).

Las trisomías autosómicas más prevalentes, T13, T18, y T21, se indican mediante los gráficos en las Figuras 31B, 31C, y 31D, respectivamente. Específicamente, la Figura 31B representa una muestra T13. Aquí, los cromosomas 18 y 21 muestran de forma típica el patrón de “dos cromosomas”, el cromosoma X muestra de forma típica el patrón de “un cromosoma”, y hay lecturas del cromosoma Y.

Conjuntamente, esto indica disomía en los cromosomas 18 y 21, e identifica un genotipo XY fetal. Sin embargo, el cromosoma 13 muestra un típico patrón de “tres cromosomas”, específicamente. De forma similar, la Figura 31C representa una muestra T18, y la Figura 31D representa una muestra T21.

El método también puede detectar aneuploidías cromosómicas sexuales, incluidas 45,X (Figura 31E), 47,XXY (Figura 31F), y 47,XYY (Figura 31G). Cabe señalar que el método determina el número de copia en los cromosomas 13, 18, 21, X, e Y; se informa del número de cromosoma total asumiendo la disomía en los cromosomas restantes. Las regiones del cromosoma X del gráfico que representan una muestra 45,X revela la presencia de un solo cromosoma. Sin embargo, la ausencia de lecturas del cromosoma Y, junto con el patrón de “dos cromosomas” para los cromosomas 13, 18, y 21, indica un genotipo 45,X. Por el contrario, las muestras 47,XXY generan un gráfico que revela la presencia de dos cromosomas X. Los datos también revelaron lecturas para alelos del cromosoma Y. Junto con la presencia de dos copias de cromosomas 13, 18 y 21, esto indica un genotipo 47,XXY. Un genotipo 47,XYY se indica mediante la presencia de un patrón de “un cromosoma” para el cromosoma X, y lecturas que indican la presencia de dos cromosomas Y.

Exposición

Este método detectó T13, T18, T21, 45,X, 47,XXY, y 47,XYY de forma no invasiva de sangre materna. Este método interroga al ADN libre de células del plasma materno mediante amplificación por PCR multiplexada focalizada y una secuenciación de alto rendimiento de 19.488 SNPs. Esto, junto con los sofisticados análisis informáticos del método que tienen en cuenta la información genética parental y numerosos parámetros de muestra, incluida la fracción fetal y la calidad del ADN, detecta de un modo más robusto la señal fetal y hace determinaciones del estado de ploidía altamente precisas en los cinco cromosomas implicados en los siete tipos más comunes de aneuploidía en el nacimiento (T13, T18, T21, 45,X, 47,XXX, 47,XXY, y 47,XYY). Este método ofrece una serie de ventajas clínicas sobre métodos anteriores, incluida y más significativamente una mayor cobertura clínica y precisiones calculadas específicas de la muestra (análogas a una puntuación de riesgo personalizada).

Mayor cobertura clínica

Este método ofrece aproximadamente un aumento doble en la cobertura de la aneuploidía en comparación con las metodologías NIPT clínicamente disponibles, dada su capacidad de detectar de forma precisa trisomías y aneuploidías cromosómicas sexuales. El método que se presenta aquí es la única prueba no invasiva que determina la ploidía en los cromosomas sexuales con una alta precisión. Los experimentos de mezcla de ADN anteriores y las muestras de plasma separadas analizadas en nuestros ensayos experimentales sugieren que este método detectará una cohorte mayor de anomalías de cromosomas sexuales, incluido 47,XXX. El método que aquí se presenta también detecta aneuploidías en los cromosomas 13, 18, y 21 con altas sensibilidades y especificidades, y con un diseño de cebadores apropiado se prevé que pueda detectar también el número de copia en los cromosomas restantes.

Precisiones calculadas específicas de la muestra

Significativamente, este método calcula una precisión específica de la muestra para las determinaciones del estado de ploidía en cada cromosoma de cada muestra. Las precisiones calculadas por este método se prevé que desciendan significativamente la tasa de determinaciones incorrectas identificando y marcando muestras individuales que tienen

ADN de baja calidad o fracciones fetales bajas que es probable que den un resultado de la prueba con una precisión baja. Por el contrario, los métodos basados en una secuenciación por fuerza bruta masivamente paralela (MPSS) producen una determinación positiva o negativa utilizando una prueba de rechazo de una única hipótesis, y su estimación de la precisión se basa en una cohorte de estudios publicados en lugar de en las características de la muestra individual, que se supone que tienen la misma precisión que el grupo. Sin embargo, las precisiones individuales para las muestras con parámetros en la cola de la distribución de cohorte pueden diferir significativamente. Esto se ve exacerbado en las fracciones fetales bajas, como en la edad gestacional temprana, o para muestras con una calidad de ADN baja. Estas muestras por lo general no están identificadas y marcadas para el seguimiento, lo que puede dar como resultado determinaciones fallidas. Sin embargo, el presente método, tiene en cuenta muchos parámetros, incluida la fracción fetal y un número de métricas de calidad del ADN, para realizar cada determinación del número de copia del cromosoma, calculando una precisión específica de la muestra para esa determinación. Esto permite que el método identifique muestras individuales con una precisión baja y las marque para su seguimiento. Esto está previsto que elimine las determinaciones fallidas, especialmente en las primeras etapas del embarazo cuando las fracciones fetales, de forma típica, son bajas. La suposición es que es preferible una no determinación que una determinación fallida, ya que una no determinación simplemente requiere un nuevo trazado y un nuevo análisis.

Conversión de precisiones calculadas en puntuaciones de riesgo tradicionales

Este método puede ofrecer un riesgo ajustado de aneuploidía para mujeres embarazadas de alto riesgo, donde el riesgo ajustado tiene en cuenta un riesgo a priori (Benn P, Cuckle H, Pergament E. Non-invasive prenatal diagnosis for Down syndrome: the paradigm will shift, but slowly. *Ultrasound Obstet Gynecol* 2012;39:127-130). Aunque el método presente ofrece a cada paciente una precisión calculada personalizada, para uso clínico estas precisiones se pueden convertir en puntuaciones de riesgo tradicionales, lo que también denota el riesgo de un embarazo aneuploide pero se expresan como fracciones. Las puntuaciones de riesgo tradicionales tienen en cuenta distintos parámetros, incluidos los riesgos relacionados con la edad y los niveles de suero de los marcadores bioquímicos, para ofrecer una puntuación de riesgo por encima de la cual una madre se considera de alto riesgo y para la que se recomiendan procedimientos de diagnóstico invasivo de seguimiento. Este método significativamente refina esta puntuación de riesgo, reduciendo así tanto las tasas de falsos positivos como las de falsos negativos, y ofrece una evaluación más precisa del riesgo materno individual. Una precisión calculada como la que se utiliza aquí es la probabilidad de que la determinación del estado de ploidía es correcta y se expresa como un porcentaje, pero las precisiones calculadas utilizadas en el Experimento 19 no incluyen un riesgo relacionado con la edad. Puesto que el cálculo de una puntuación de riesgo de forma típica incluye un riesgo relacionado con la edad, las precisiones calculadas y las puntuaciones de riesgo tradicionales no son intercambiables; se deben combinar para convertirse en una puntuación de riesgo tradicional. La fórmula para combinar el riesgo relacionado con la edad con la precisión calculada es:

$$\frac{R_1 R_2}{R_1 R_2 + [1 - R_1][1 - R_2]}$$

donde R1 es la puntuación de riesgo calculada por el presente método y R2 es la puntuación de riesgo calculada por el análisis del primer trimestre.

Problemas de invalidación de los métodos basados en SNP con variación de amplificación

Una desventaja inherente a los métodos de recuento utilizada por otros métodos es que determinan el estado de ploidía fetal midiendo el ratio del número de lecturas que se correlacionan con el cromosoma de interés (por ejemplo, el cromosoma 21) con aquellos que se correlacionan con un cromosoma de referencia. Los cromosomas con un contenido alto o bajo de GC, incluidos los cromosomas 13, X, e Y, amplifican con una alta variabilidad. Esto puede dar como resultado una variación de señal que es comparable en magnitud a la señal del ADN libre de células fetal, que puede confundir las determinaciones del número de copia alterando el ratio de lecturas de alelos del cromosoma de interés con aquellos del cromosoma de referencia. Esto puede dar como resultado una precisión baja para los cromosomas 13, X, e Y. Significativamente, este problema se ve exacerbado en las fracciones de ADN libre de células fetal, como suele ser el caso en edades gestacionales tempranas.

Por el contrario, los métodos basados en SNP no se basan en niveles de amplificación constantes entre los cromosomas y, por tanto, se espera que proporcionen resultados que sean igual de precisos en todos los cromosomas. Puesto que el método presente mira, en parte, a recuentos relativos de alelos diferentes en loci polimórficos, que por definición difieren sólo por un único nucleótido, no requiere el uso de cromosomas de referencia, y esto obvia los problemas con la variación de amplificación cromosoma a cromosoma que son inherentes en los métodos que se basan en cuantificar los recuentos de lectura. A diferencia de los métodos cuantitativos que requieren cromosomas de referencia que son euploides, el presente método se espera que pueda detectar la triploidía así como las anomalías neutrales de número de copia como la disomía uniparental.

La importancia de una detección prematura

Significativamente, el predominio de aneuploidías cromosómicas sexuales combinadas en el nacimiento es mayor que el de la mayoría de aneuploidías autosómicas comunes (**Figura 32**). Sin embargo, no hay actualmente métodos de análisis no invasivo rutinarios que detecten de manera fiable las anomalías cromosómicas sexuales. Por tanto, las anomalías de los cromosomas sexuales por lo general se detectan prenatalmente como un efecto secundario de las pruebas rutinarias para el síndrome de Down u otras aneuploidías autosómicas; una gran proporción de casos son

completamente fallidos. La detección prematura y precisa es crucial para muchos de estos trastornos donde una intervención terapéutica prematura mejora los resultados clínicos. Por ejemplo, el síndrome de Turner a menudo no se diagnostica hasta la adolescencia, aunque su predominio total en el nacimiento es de 1 en 2.500 mujeres. Se conoce que la terapia de la hormona del crecimiento evita la baja estatura que da como resultado este trastorno, pero los tratamientos son significativamente más eficaces cuando se inician antes de los 4 años de edad. Además, la terapia de sustitución de estrógenos puede estimular las características sexuales secundarias en pacientes con el síndrome de Turner, aunque, de nuevo, la terapia se debe iniciar antes de la adolescencia, antes de que se detecte por lo general el síndrome. Conjuntamente, esto subraya la importancia de una detección prematura, rutinaria y segura de las aneuploidías cromosómicas sexuales. Este método ofrece un primer planteamiento con el potencial para servir como un análisis rutinario para las anomalías de los cromosomas sexuales.

Aplicaciones adicionales

Puesto que este método utiliza la amplificación focalizada, se encuentra en una posición exclusiva para detectar anomalías submicroscópicas, como microdeleciones y microduplicaciones. Aunque se ha demostrado que los métodos no focalizados como MPSS detectan el síndrome de microdelección de DiGeorge, esto ha requerido un nivel suficientemente alto de cobertura genómica para hacer que el planteamiento sea inviable. Esto es porque la amplificación no focalizada será varios órdenes de magnitud menos eficiente en regiones submicroscópicas, ya que una fracción muy pequeña de las lecturas de secuenciación será informativa. Además, el hecho de que los métodos disponibles actualmente tienen dificultades para identificar de forma precisa el estado de ploidía para los cromosomas sexuales sugiere que también encontrarán problemas de amplificación variable en los segmentos cromosómicos más pequeños.

De forma similar, los métodos basados en SNP pueden detectar trastornos de UPD, que son anomalías neutrales de número de copia que no se detectarán o bien mediante los métodos no invasivos actuales que se basan en el recuento o los métodos invasivos tradicionales como la amniocentesis y CVS que se basan en el establecimiento del cariotipo citogenético y/o la hibridación in situ de fluorescencia. Esto es porque los métodos basados en SNP pueden distinguir de forma exclusiva haplotipos individuales, mientras que los métodos focalizados y basados en MPSS disponibles clínicamente amplifican loci no polimórficos y, por tanto, no pueden determinar, por ejemplo, si los cromosomas de interés se originan a partir del mismo progenitor. Esto significa que estas microdeleciones/microduplicaciones y síndromes de UPD, incluyendo los síndromes de Prader-Willi, Angelman, y Beckwith-Wiedemann, por lo general, no se diagnostican prenatalmente y, a menudo, inicialmente se diagnostican erróneamente postnatalmente. Esto retrasa significativamente la intervención terapéutica. Además, puesto que este método focaliza SNPs, este método también facilitará la reconstrucción del haplotipo parental, lo que permite la detección de la herencia fetal de loci vinculados a enfermedad individual (Kitzman JO, Snyder MW, Ventura M, et al. Noninvasive whole-genome sequencing of a human fetus. *Sci Transl Med* 2012;4:137ra76). Los resultados que aquí se presentan confirman el ámbito ampliado de este método para identificar la aneuploidía prenatal. Específicamente, al amplificar y secuenciar 19.488 SNPs, este método puede determinar el número de copia en los cromosomas 13, 18, 21, X, e Y, y está previsto que detecte de forma exclusiva otras anomalías cromosómicas, como triploidía y UPD, que no se detectan mediante ningún otro método no invasivo disponible clínicamente. La mayor cobertura clínica y las potentes precisiones calculadas específicas de la muestra sugieren que este método puede ser un complemento viable a la realización de pruebas invasivas para la detección de aneuploidías cromosómicas fetales.

Por ejemplo, cualquiera de los métodos que se divulgan en el presente documento para el ADN se puede adaptar fácilmente para ARN mediante la inclusión de un paso de transcripción inversa para convertir el ARN en ADN. Los ejemplos ilustrativos que utilizan loci polimórficos se pueden adaptar rápidamente para la amplificación de loci no polimórficos si se desea.

REIVINDICACIONES

1.Un método de amplificar loci diana en una muestra de ácido nucleico, el método comprende:

5 a)poner en contacto una muestra de ácido nucleico que comprende loci diana con una biblioteca de cebadores de prueba que comprende al menos 1.000 diferentes pares de cebadores para producir una mezcla de reacción en un volumen de reacción; donde cada par de cebadores incluye un cebador de prueba directo y un cebador de prueba inverso que se hibridan en el mismo locus diana, y donde los cebadores no incluyen sondas de inversión molecular (MIPs);

10 b)someter la mezcla de reacción a las condiciones de reacción en cadena de la polimerasa (PCR) para producir productos amplificados que incluyen amplicones diana; donde la concentración de cada cebador de prueba es inferior a 10 nM; donde la longitud del paso de reformación térmica de las condiciones de reacción es superior a 10 minutos; donde al menos 1.000 loci diana diferentes se amplifican simultáneamente; y donde (i) menos de un 20% de los productos amplificados son dímeros del cebador de prueba, (ii) al menos un 80% de los productos amplificados son amplicones diana, y (iii) al menos un 80% de los loci diana están amplificados; y

15 c)secuenciar los productos amplificados.

2.El método de la reivindicación 1, donde al menos 5.000 loci diana diferentes se amplifican.

20 3.El método de la reivindicación 1, donde la longitud de los amplicones diana es inferior a 100 nucleótidos.

4.El método de la reivindicación 1, donde los cebadores de prueba se seleccionan de una biblioteca de cebadores candidatos basándose al menos en parte en la capacidad de los cebadores candidatos de formar dímeros de cebadores.

25 5.El método de la reivindicación 1, donde la muestra comprende ADN materno de la madre embarazada de un feto y ADN fetal, y donde el método comprende la determinación de la presencia o ausencia de una anomalía cromosómica fetal de los datos de secuenciación.

6.El método de la reivindicación 1, donde los loci diana están presentes en el genoma humano, o donde los loci diana comprenden polimorfismos de un solo nucleótido humanos.

7.El método de la reivindicación 1, donde la muestra de ácido nucleico comprende ADN de un tumor, trasplante, o feto.

30 8.El método de la reivindicación 1, donde la muestra comprende ADN de una sola célula.

9.Un método según la reivindicación 1, donde la biblioteca de cebadores de prueba se selecciona de una biblioteca de cebadores candidatos mediante un método que comprende:

35 a)calcular en un ordenador una puntuación no deseable para al menos un 90% de las combinaciones posibles de dos cebadores candidatos de la biblioteca, donde cada puntuación no deseable se basa, al menos en parte, en la probabilidad de la formación de dímeros entre los dos cebadores candidatos;

b)extraer de la biblioteca de cebadores candidatos el cebador candidato que es parte del mayor número de combinaciones de dos cebadores candidatos con una puntuación no deseable por encima de un primer umbral mínimo;

40 c)si el cebador candidato que se ha extraído en el paso (b) es un miembro de un par de cebadores, extraer, a continuación, el otro miembro del par de cebadores de la biblioteca de cebadores candidatos; y

d)opcionalmente repetir los pasos (b) y (c), con el fin de seleccionar una biblioteca de cebadores de prueba.

10.El método de la reivindicación 9, que comprende o bien

45 i)reducir todavía más el número de cebadores candidatos que quedan en la biblioteca disminuyendo el primer umbral mínimo utilizado en el paso (b) a un segundo umbral mínimo inferior y repetir los pasos (b) y (c) hasta que las puntuaciones no deseables para las combinaciones del cebador candidato restantes en la biblioteca son todas iguales o por debajo del segundo umbral mínimo, o hasta que el número de cebadores candidato restantes en la biblioteca se reduce a un número deseado; o

50 ii)aumentar el número de cebadores candidatos restantes en la biblioteca aumentando el primer umbral mínimo utilizado en el paso (b) a un segundo umbral mínimo más alto y repetir los pasos (b) y (c) hasta que las puntuaciones no deseables para las combinaciones del cebador candidato restantes en la biblioteca son todas iguales o por debajo del segundo umbral mínimo, o hasta que el número de cebadores candidato restantes en la biblioteca se reduce a un número deseado.

55 11.El método de la reivindicación 9, además comprende, antes del paso (b), quitar un par de cebadores de la biblioteca que produce un amplicón diana que se solapa con un amplicón diana producido por otro par de cebadores.

12.El método de la reivindicación 9, donde las puntuaciones no deseables se basan al menos en parte en uno o más parámetros seleccionados del grupo compuesto por una tasa de heterocigosidad de un locus diana, el predominio de la enfermedad asociada a un polimorfismo o mutación en un locus diana, la penetrancia de la enfermedad asociada a un polimorfismo o mutación en un locus diana, la especificidad del cebador candidato para un locus diana, el tamaño del

cebador candidato, la temperatura de fusión de un amplicón diana, el contenido de guanina-citosina (GC) de un amplicón diana, la eficacia de amplificación de un amplicón diana, y el tamaño de un amplicón diana.

5 13.El método de la reivindicación 9, que comprende la selección de una biblioteca de al menos 1.000 cebadores de prueba diferentes y la utilización de al menos 1.000 de los cebadores de prueba seleccionados para amplificar simultáneamente al menos 1.000 loci diana diferentes.

14.El método de la reivindicación 1, donde las temperaturas de fusión de los cebadores de los pares de cebadores son de 55 a 65 °C.

15.El método de la reivindicación 1, donde la temperatura de reformación térmica comprende una temperatura que se eleva.

10 16.El método de la reivindicación 1, donde las temperaturas de fusión de los cebadores de los pares de cebadores son de 57 a 60,5 °C.

17.El método de la reivindicación 1, donde la PCR es PCR anidada.

18.El método de la reivindicación 1, donde el rango de la longitud de los amplicones diana está entre 50 y 100 nucleótidos.

15 19.El método de 1, donde los pares de cebadores están diseñados para amplificar loci de SNP potencialmente heterocigóticos en los cromosomas X, Y, 13, 18, y 21 y estos ensayos se utilizan en una sola reacción para amplificar ADN libre de células obtenido de una muestra de plasma materno.

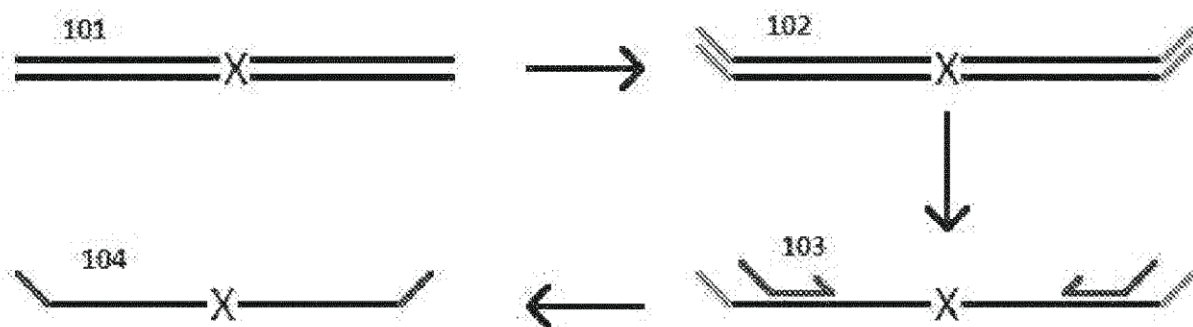


Fig 1

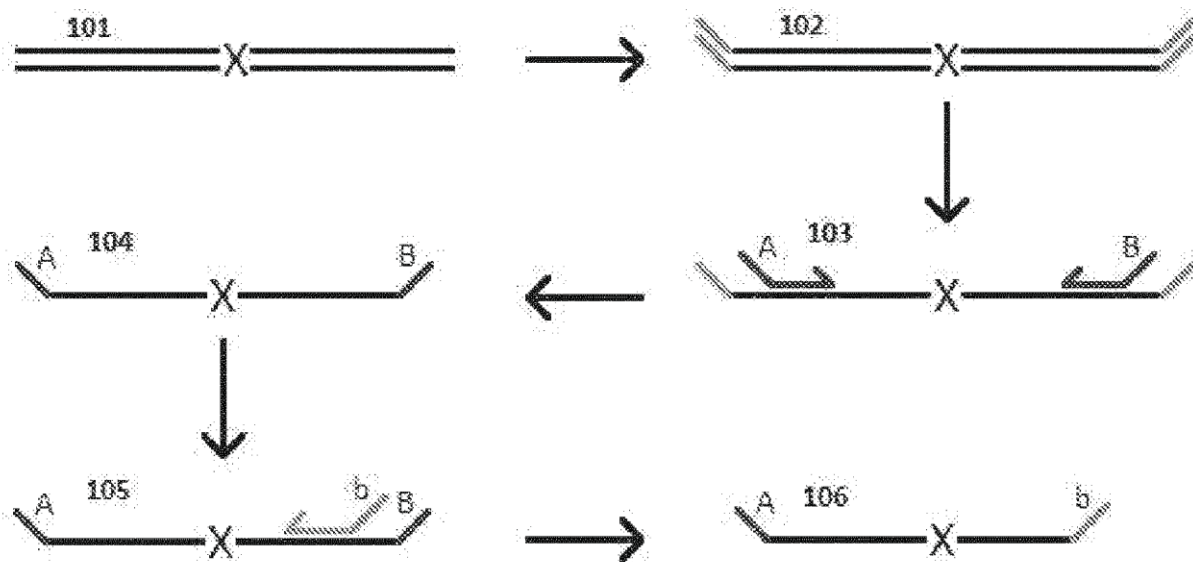


Fig 2

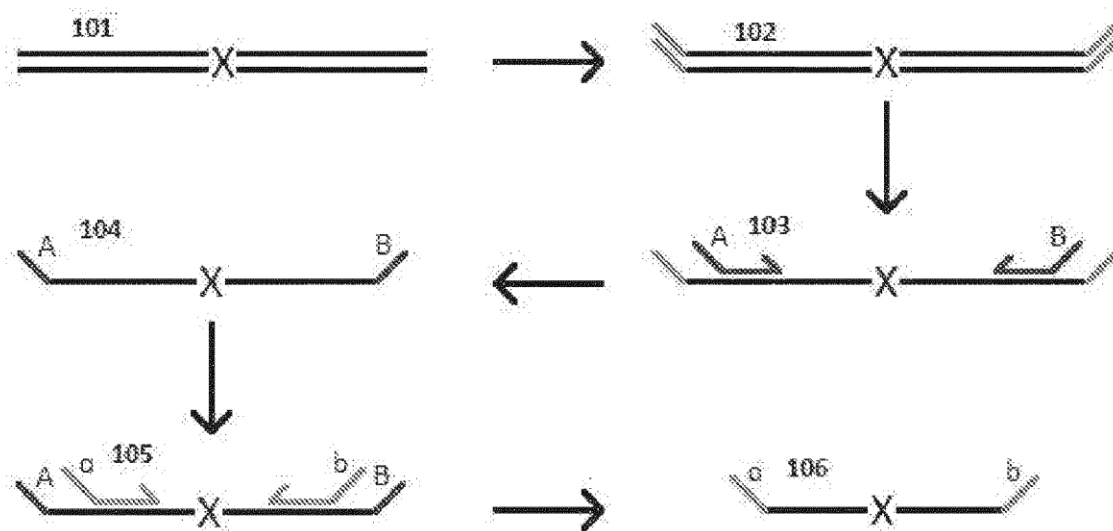


Fig 3

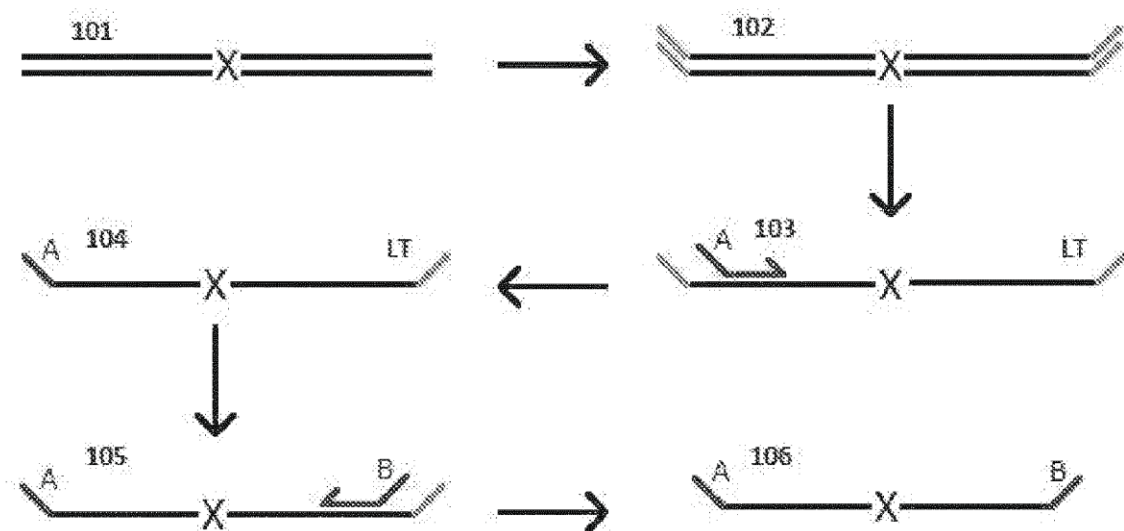


Fig 4

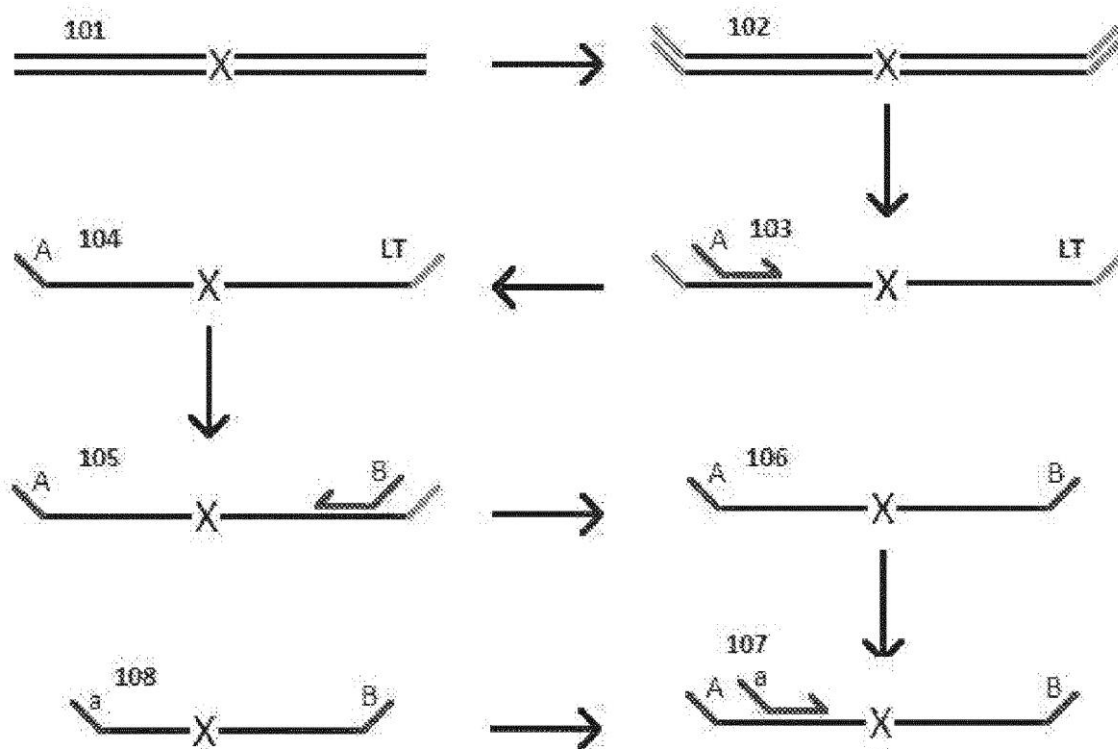


Fig 5

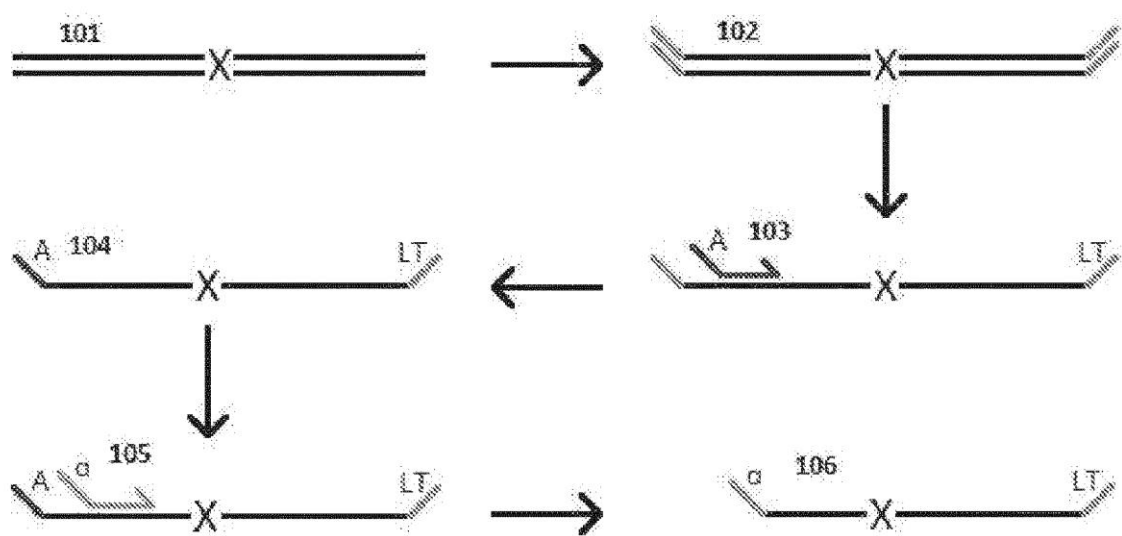


Fig 6

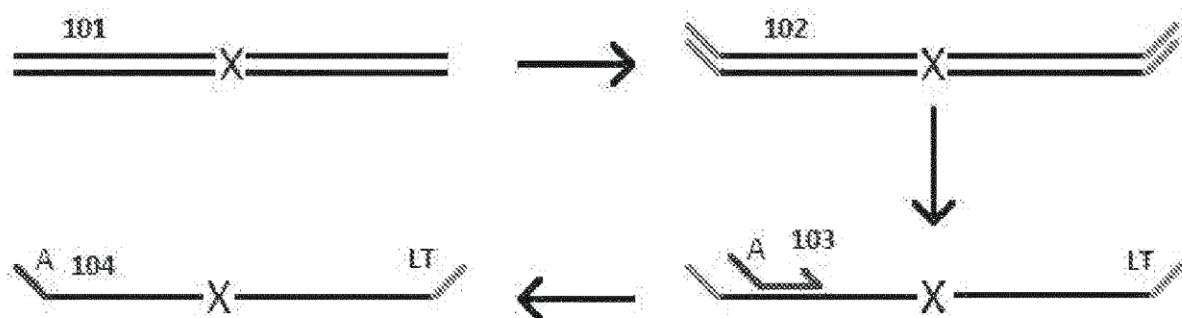


Fig 7

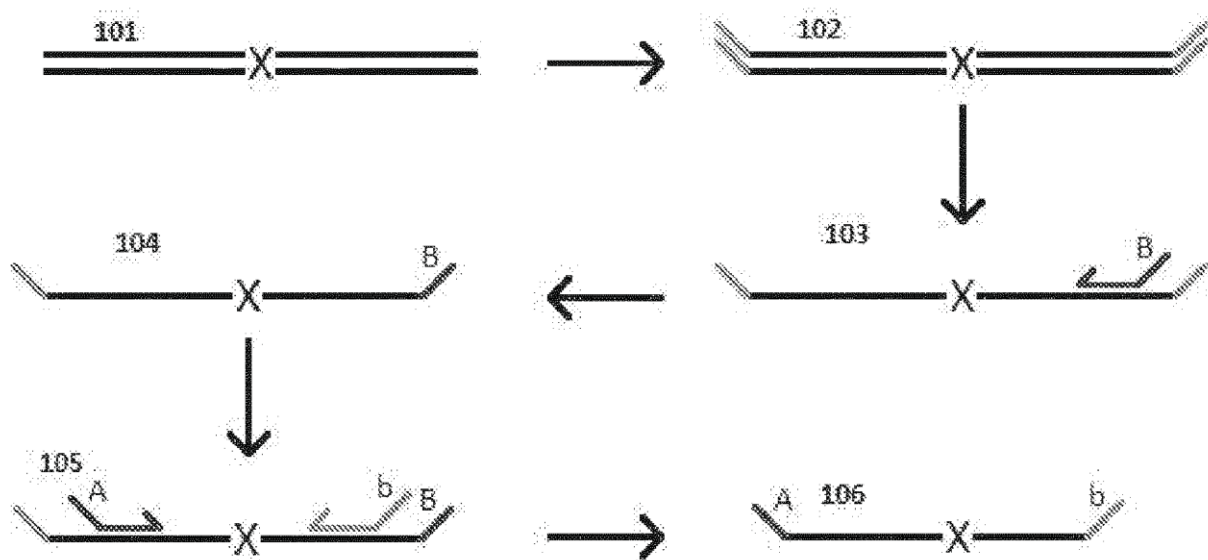


Fig 8

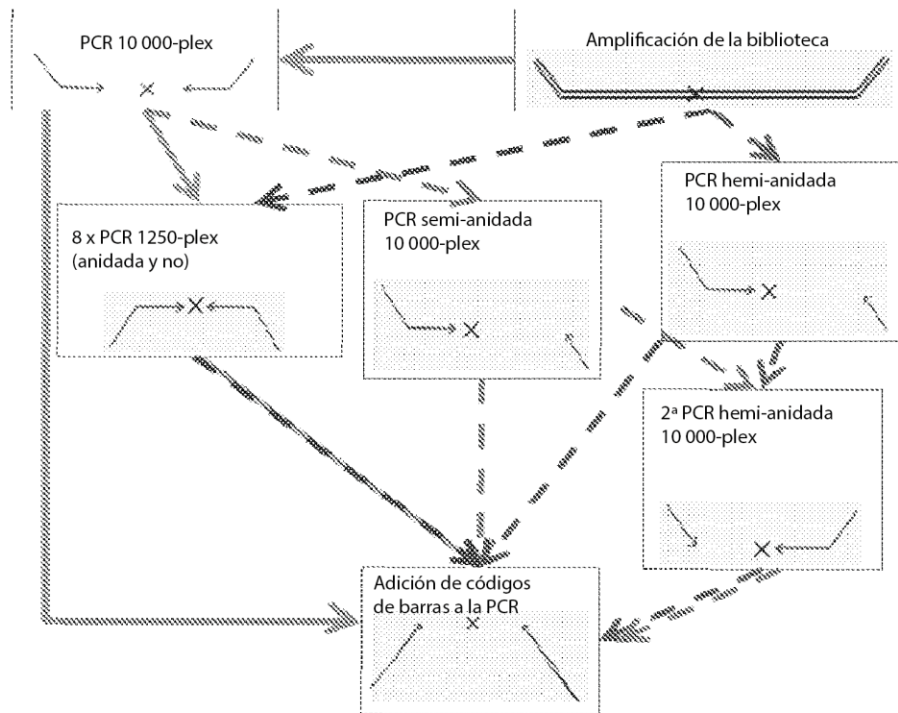


Fig 9

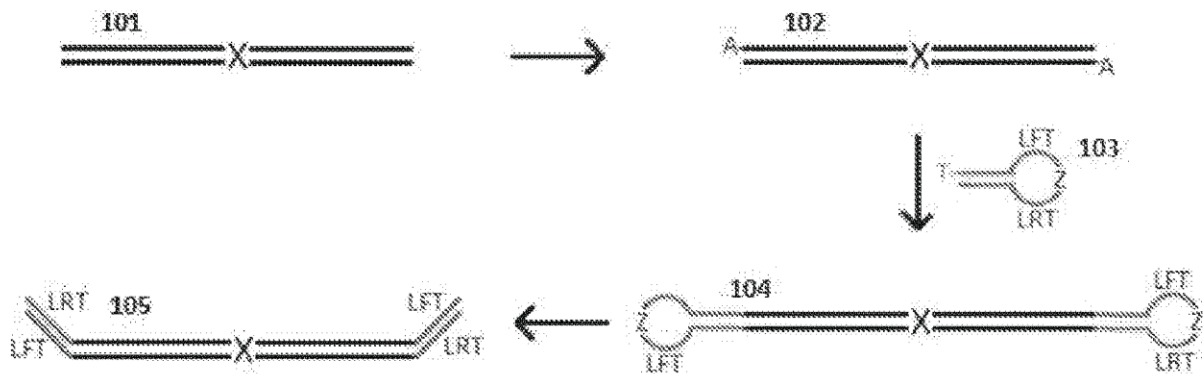


Fig 10

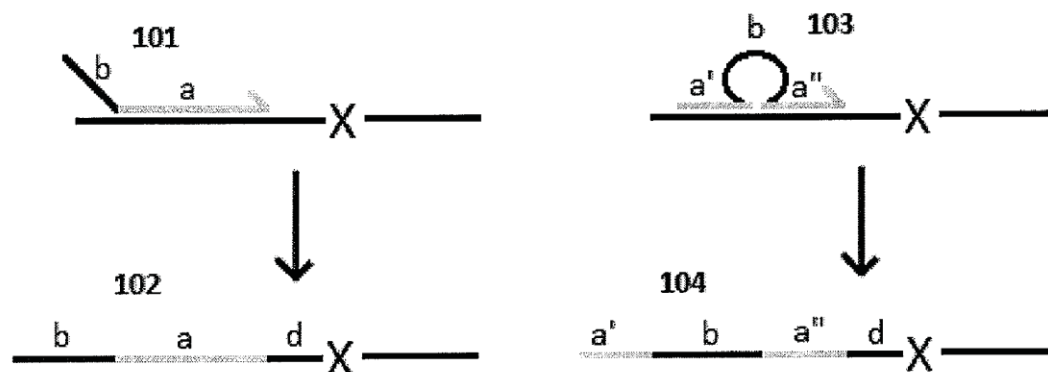


Fig 11

La secuencia del adaptador de secuenciación se encuentra dentro de la secuencia del cebador y flanqueada por una secuencia específica de diana a ambos lados. 10 bases son específicas de diana en el extremo 3' de cada cebador. Los cebadores se probaron con éxito en una PCR en tiempo real. Para la secuenciación, esto reduce el número de bases del cebador que es necesario secuenciar.			
Etiqueta int.			Id. de sec.
rs8130564	1.10	AACTCACATAGC ACACGACGCTCTCCGATCT TGCAAGCACA	1
rs2832093	2.10	TCCTCTGTG ACACGACGCTCTCCGATCT CCCTGCTCTT	2
rs12011281	3.10	tcctctctctACACGACGCTCTCCGATCTcGGGCTGTCA	3
rs6719561	4.10	TACATCCTTGAGACACGACGCTCTCCGATCTGCTGTGCAGT	4
rs10187018	5.10	tttgcttgagctACACGACGCTCTCCGATCTcgggagtttc	5
rs10460481	6.10	gtcttatggtggACACGACGCTCTCCGATCTcaaagccagt	6

Fig 12

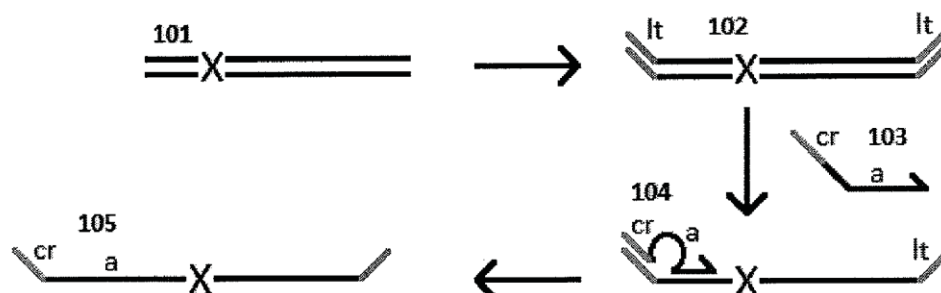


Fig 13

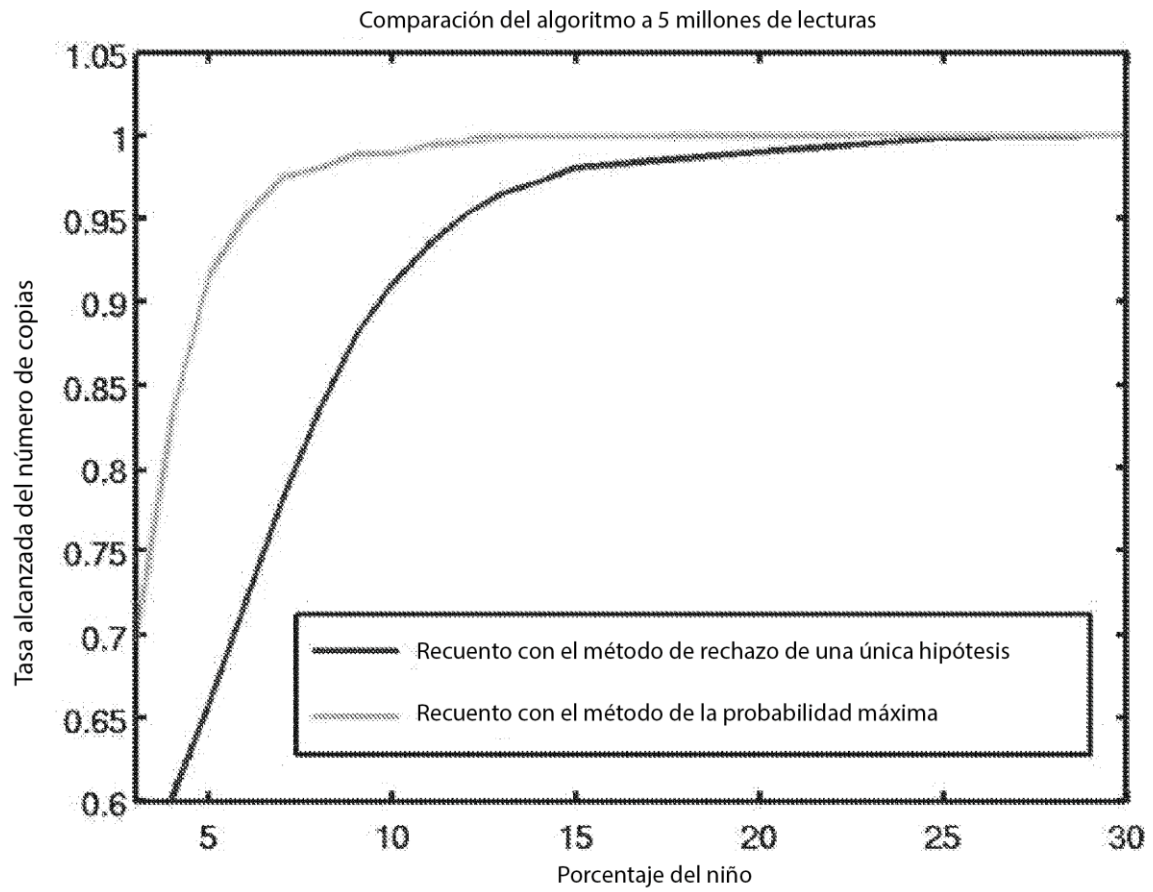


Fig 14

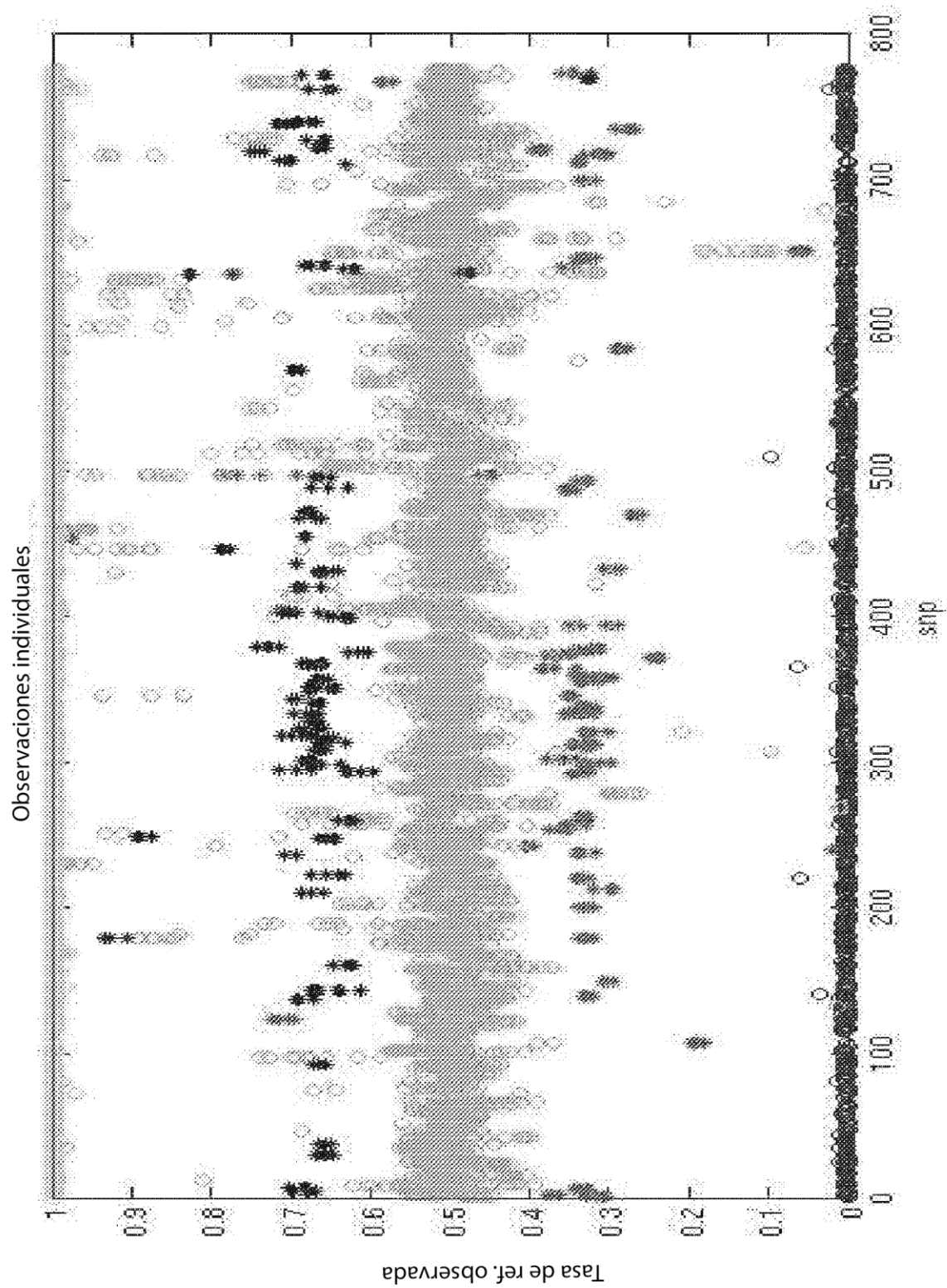


Fig 15

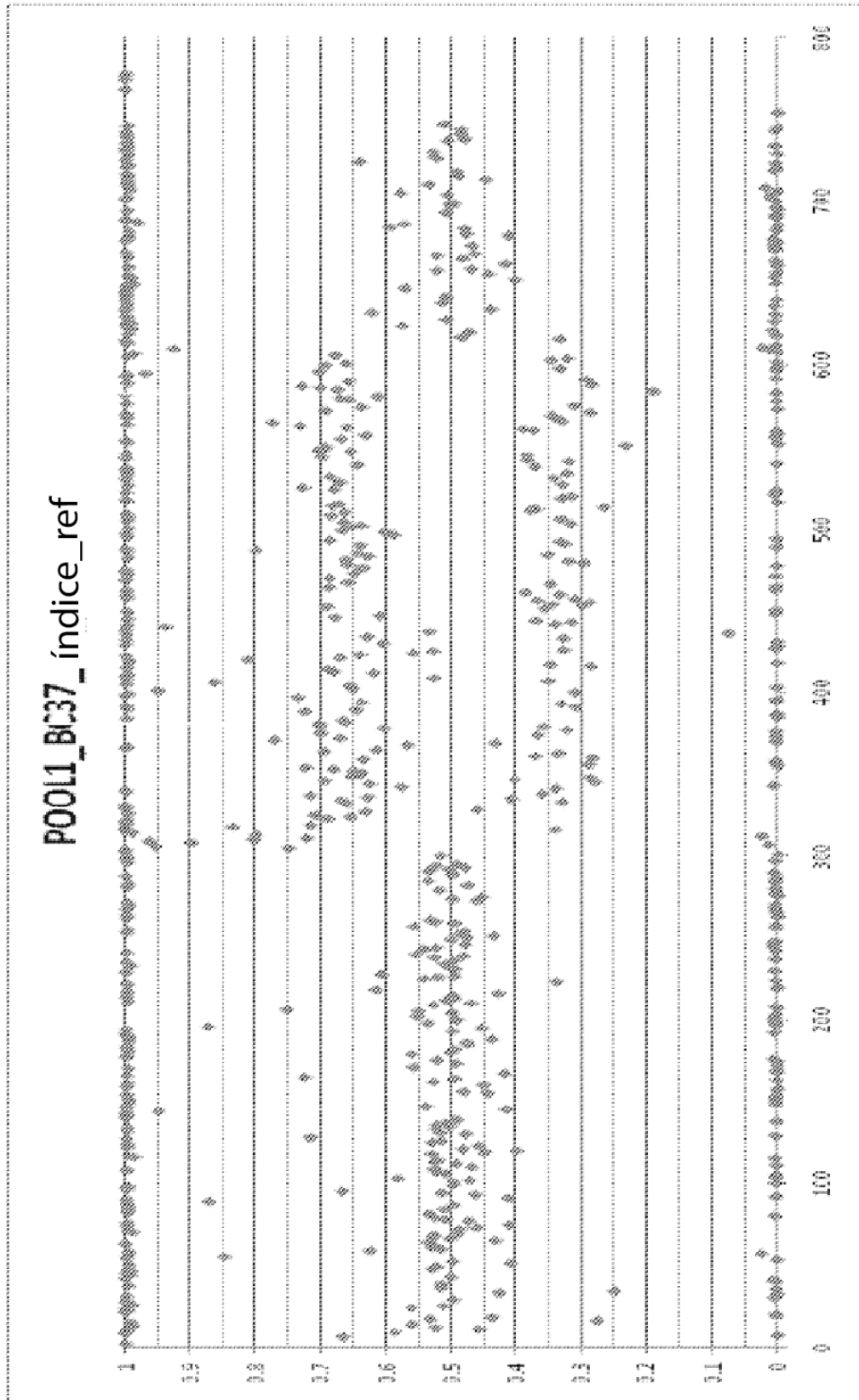


Fig 16

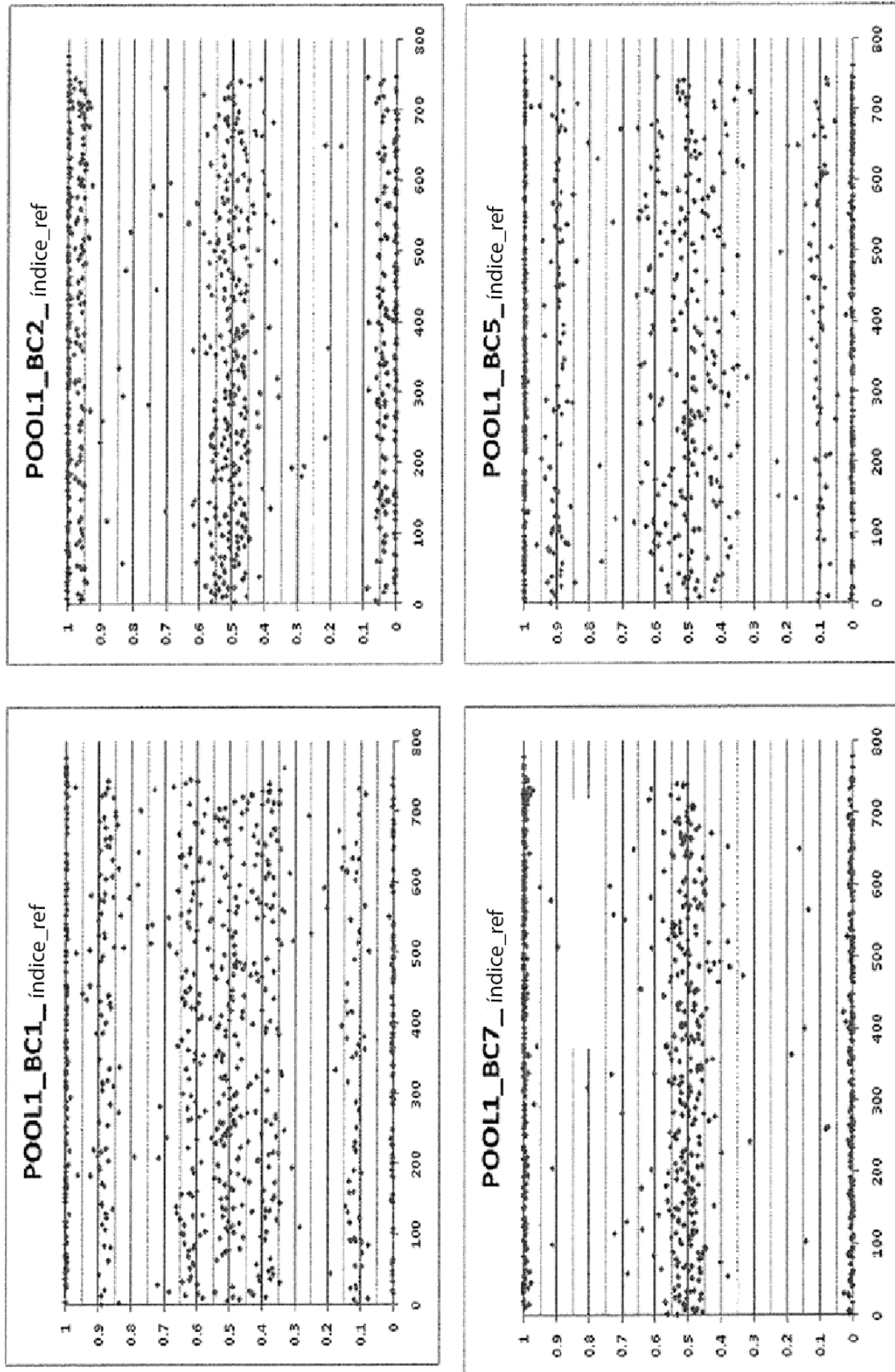


Fig 17

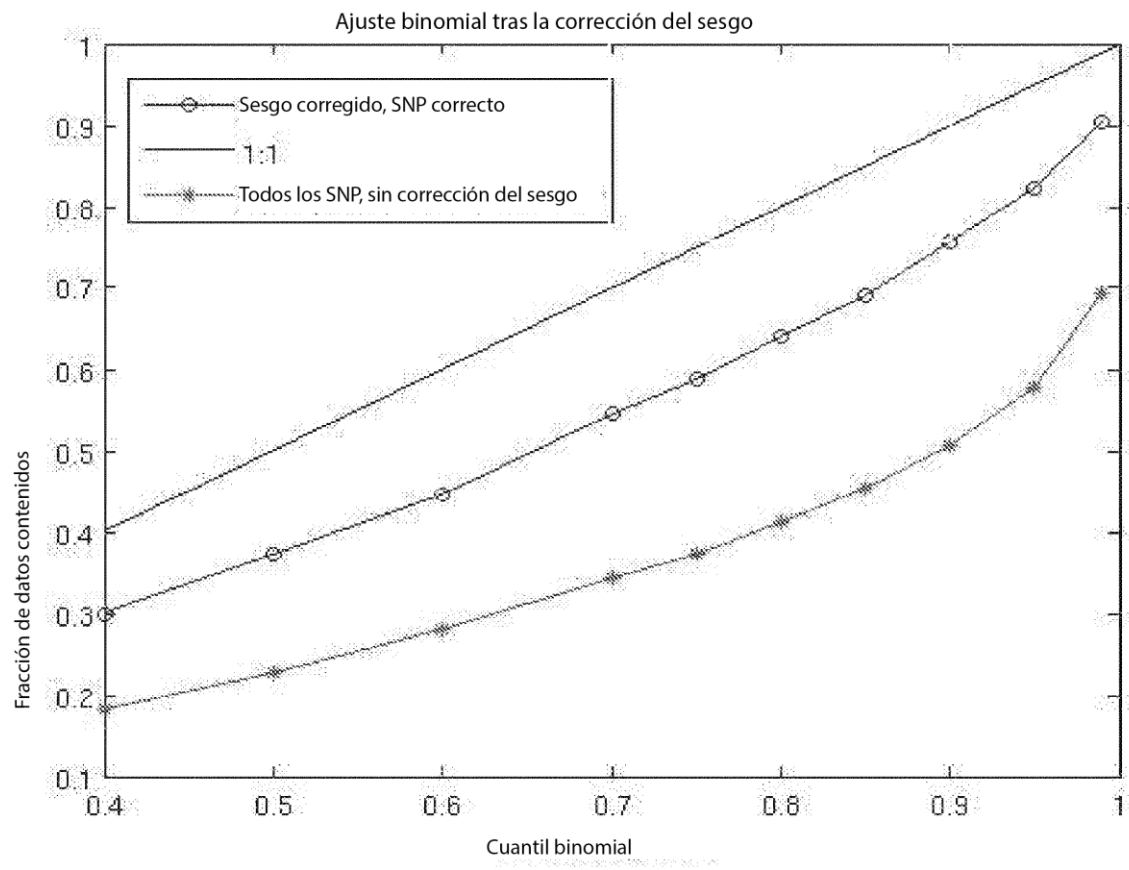


Fig 18

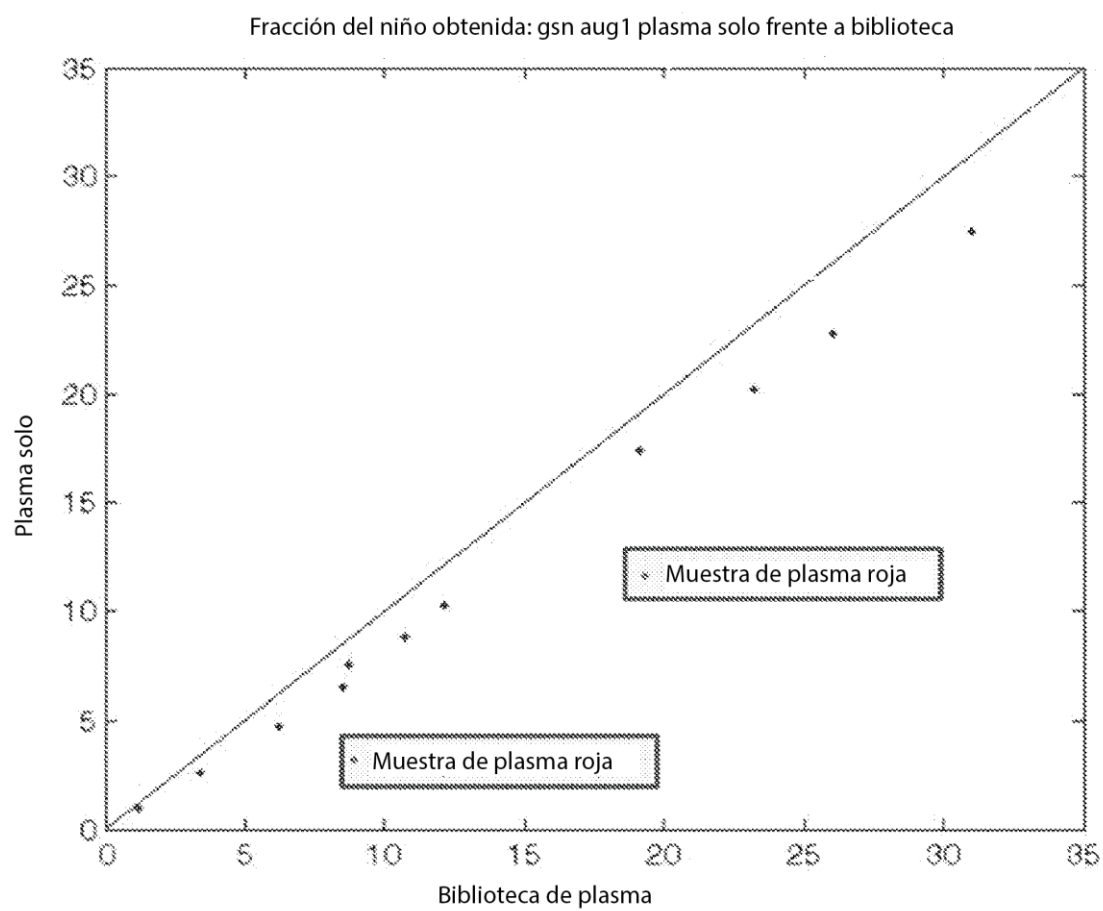


Fig 19

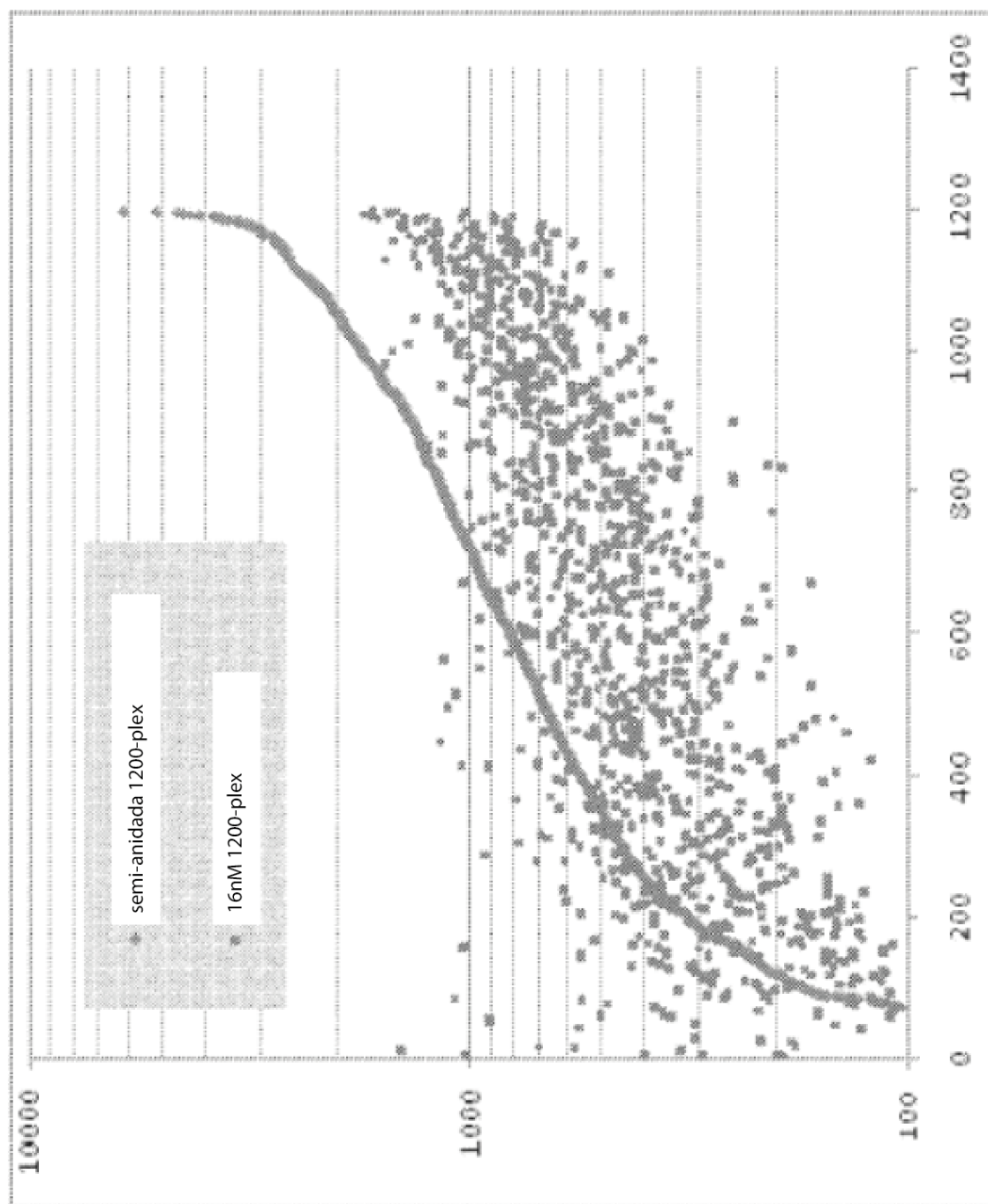


Fig 20

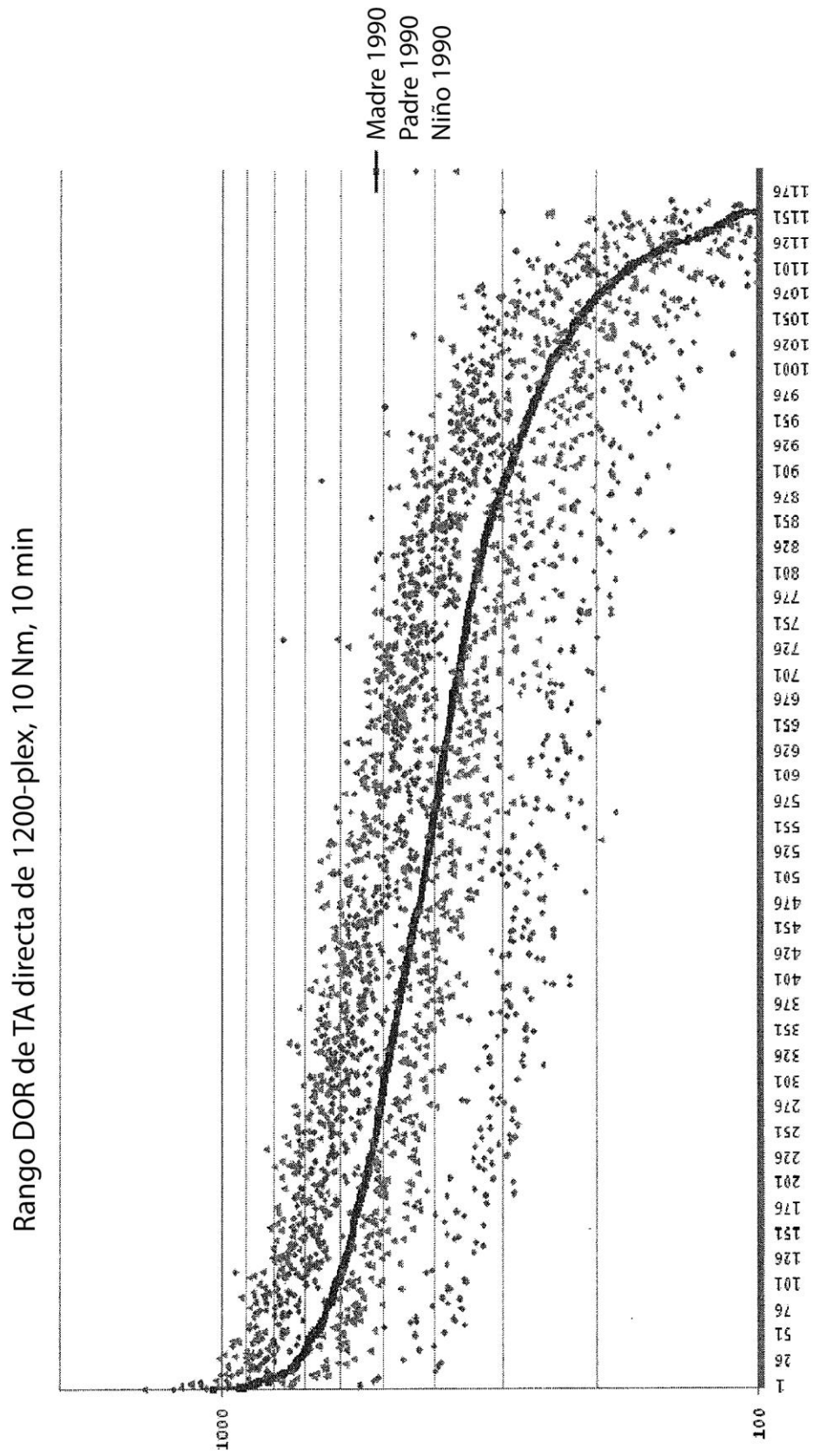


Fig 21

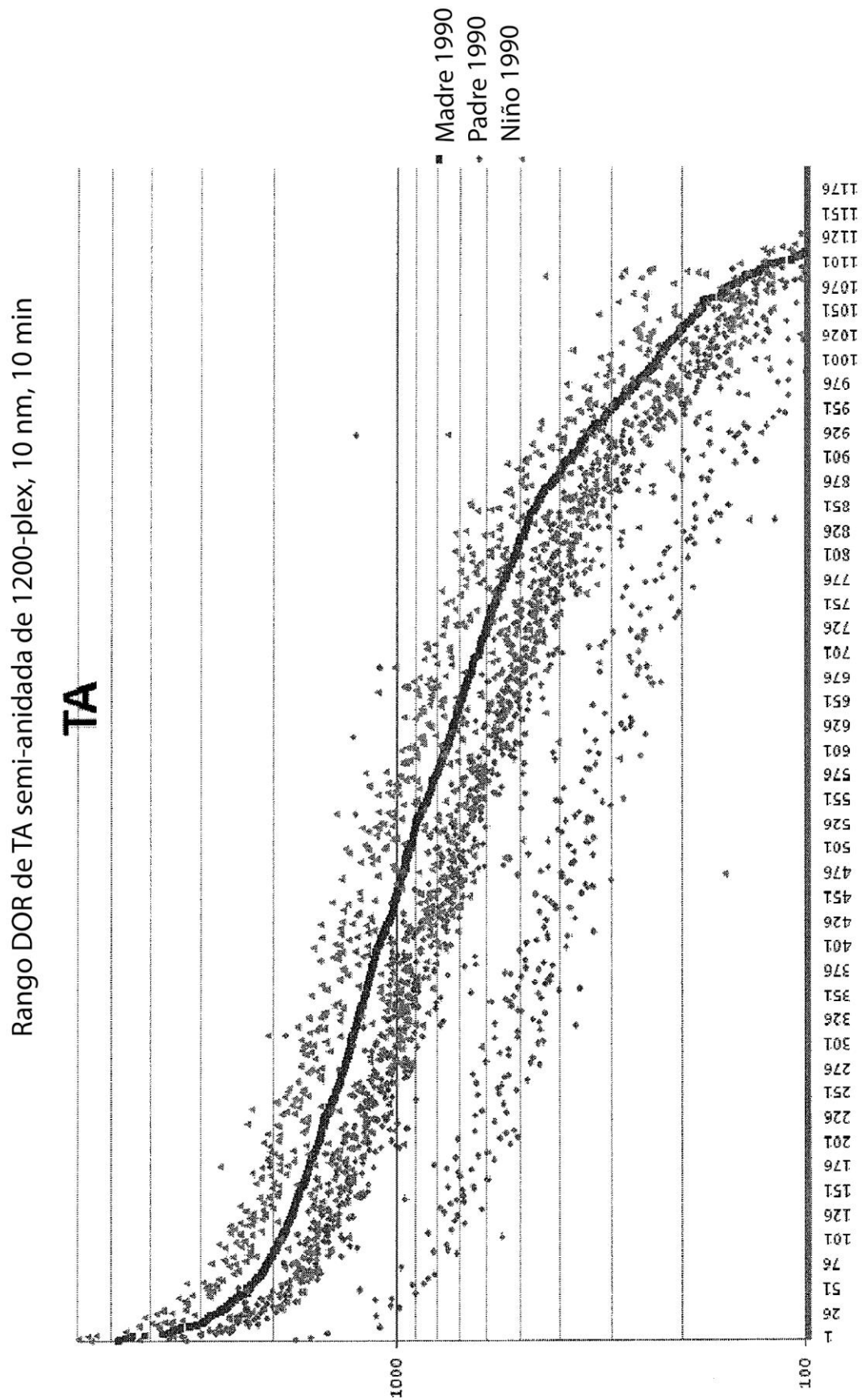


Fig 22

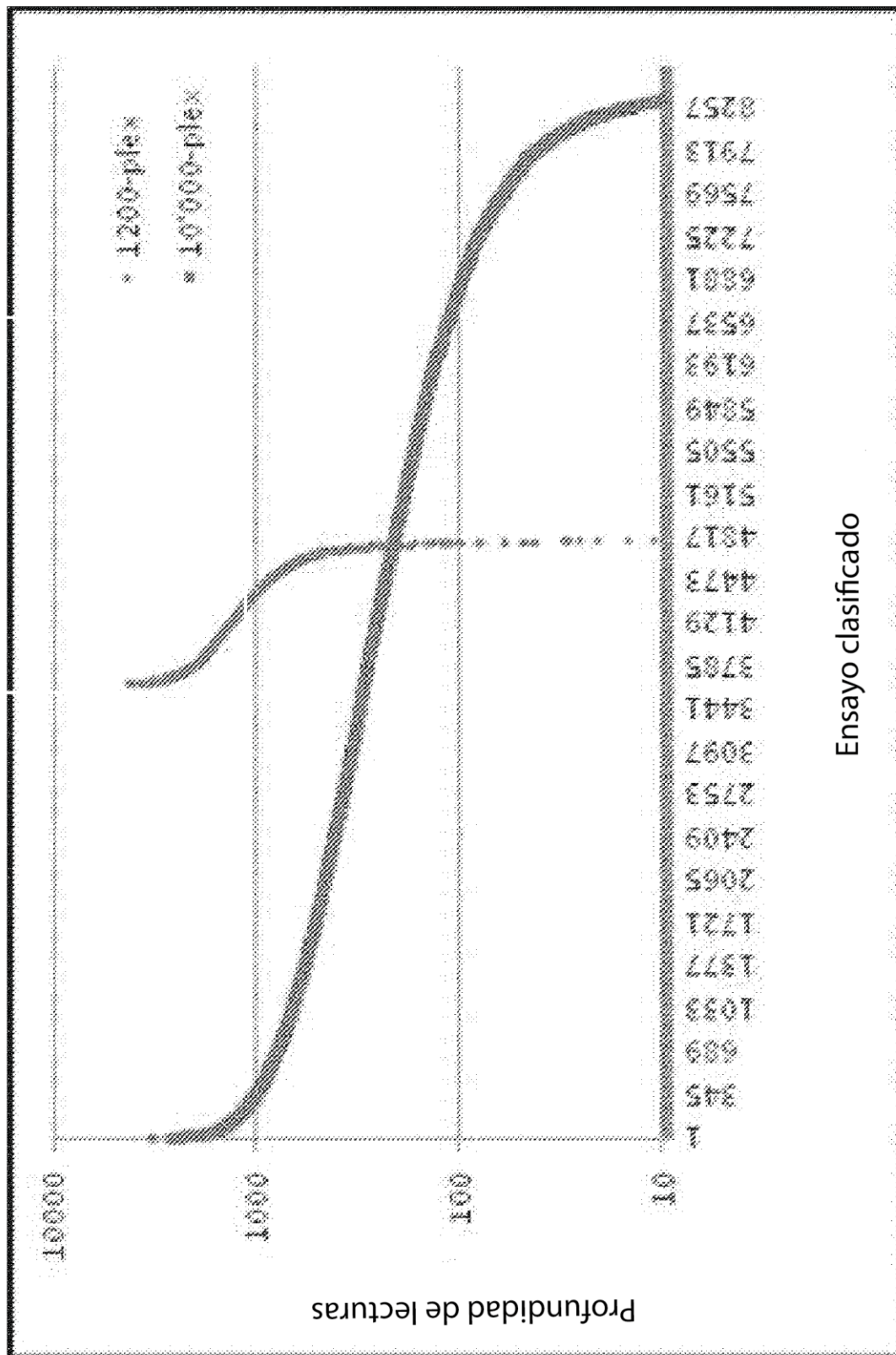


Fig : 23

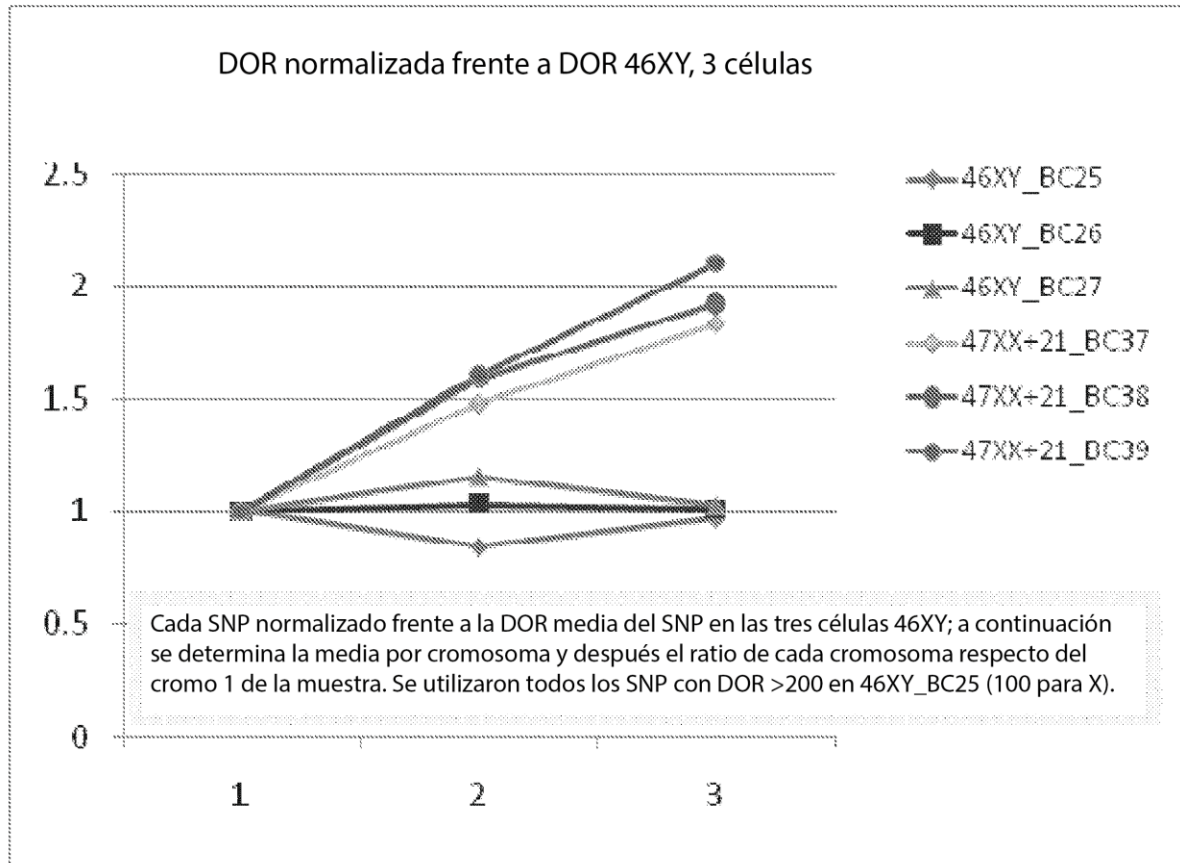


Fig 24

- Datos de 3 células con Arcturus: la mini-PCR puede detectar trisomía 21 de 3 células
- Izquierda: cromosoma 1; medio: cromosoma 21; y derecha: cromosoma X
- Clasificados de izquierda a derecha por la DOR para cada cromosoma

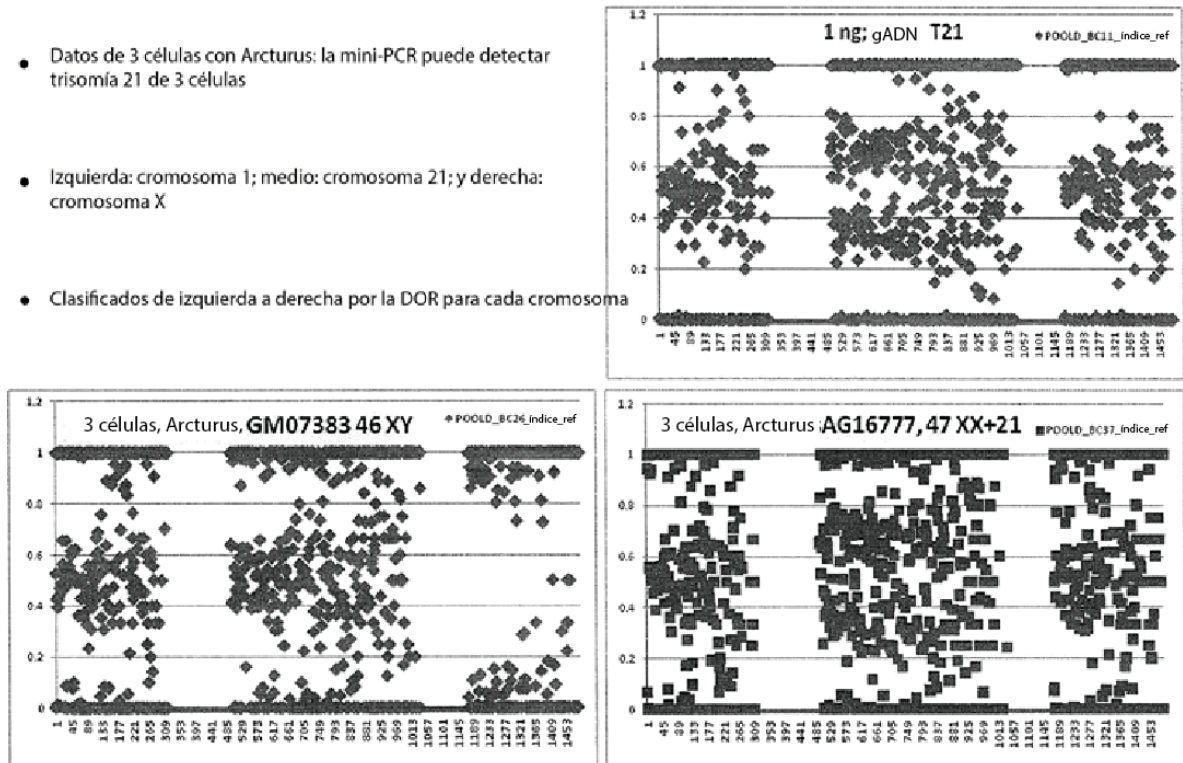


Fig 25

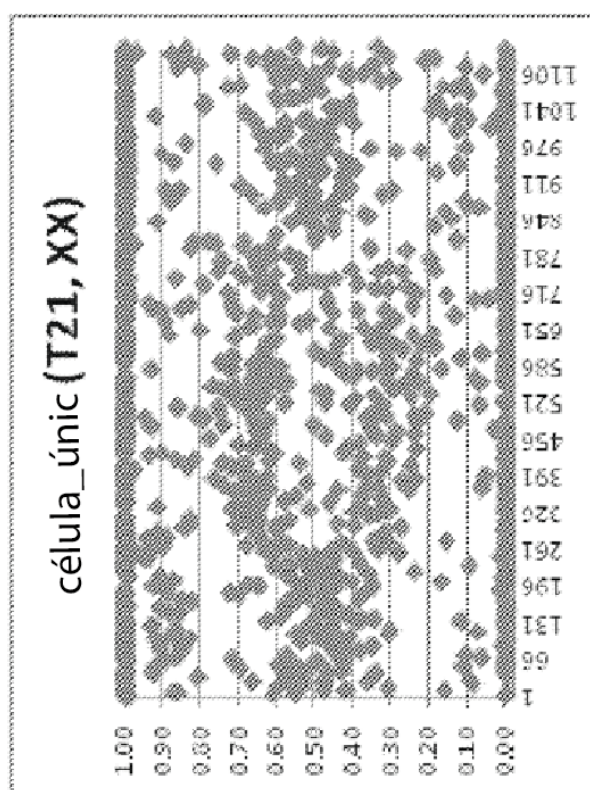
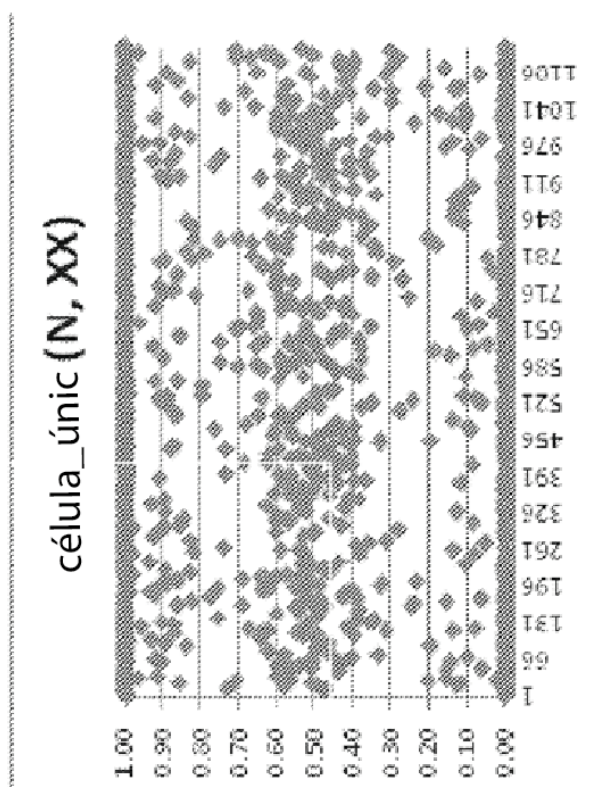


Fig 26

Fig 27

Frecuencias alélicas menores para NIPD dianas, Cromosoma 21

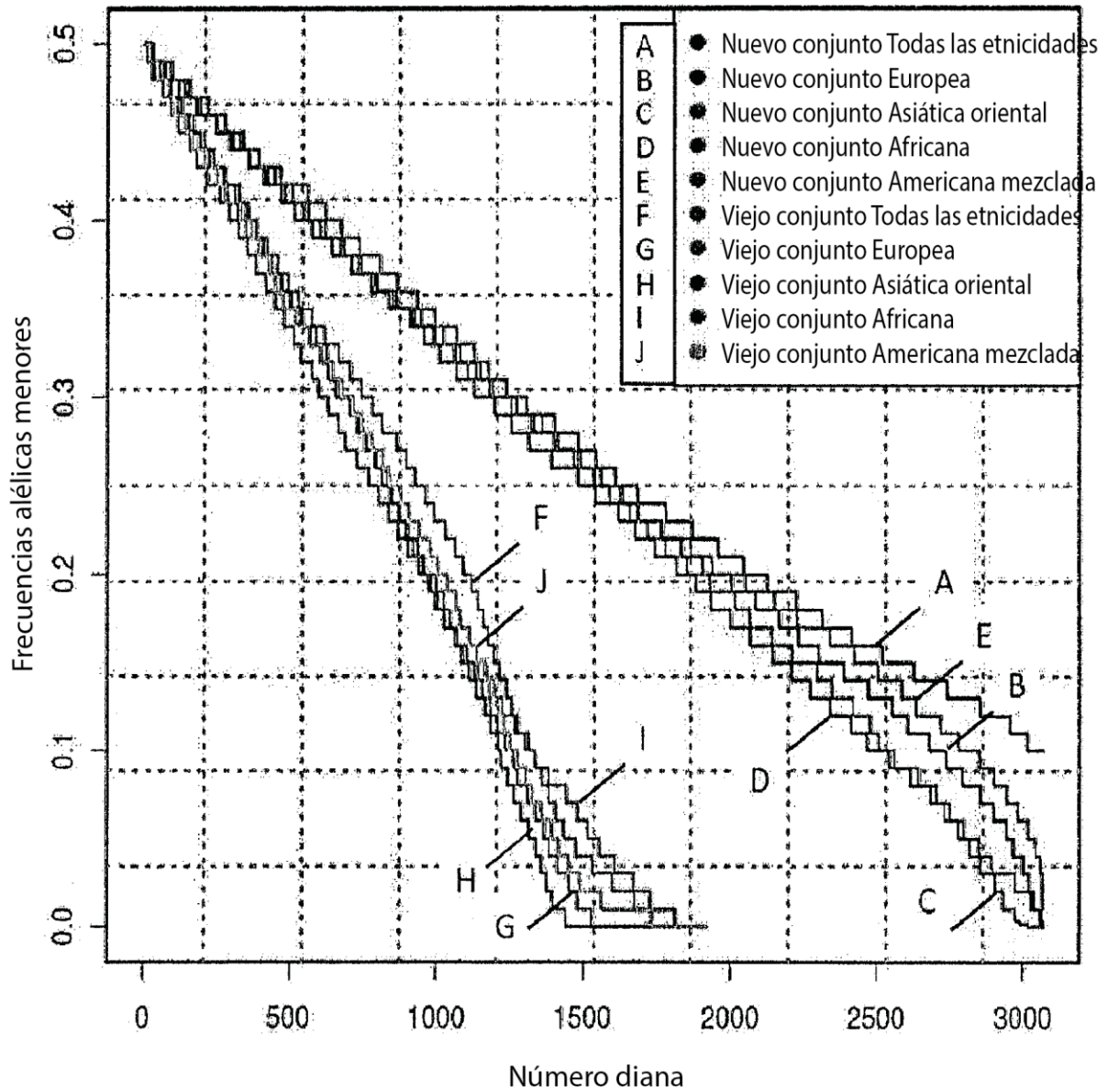
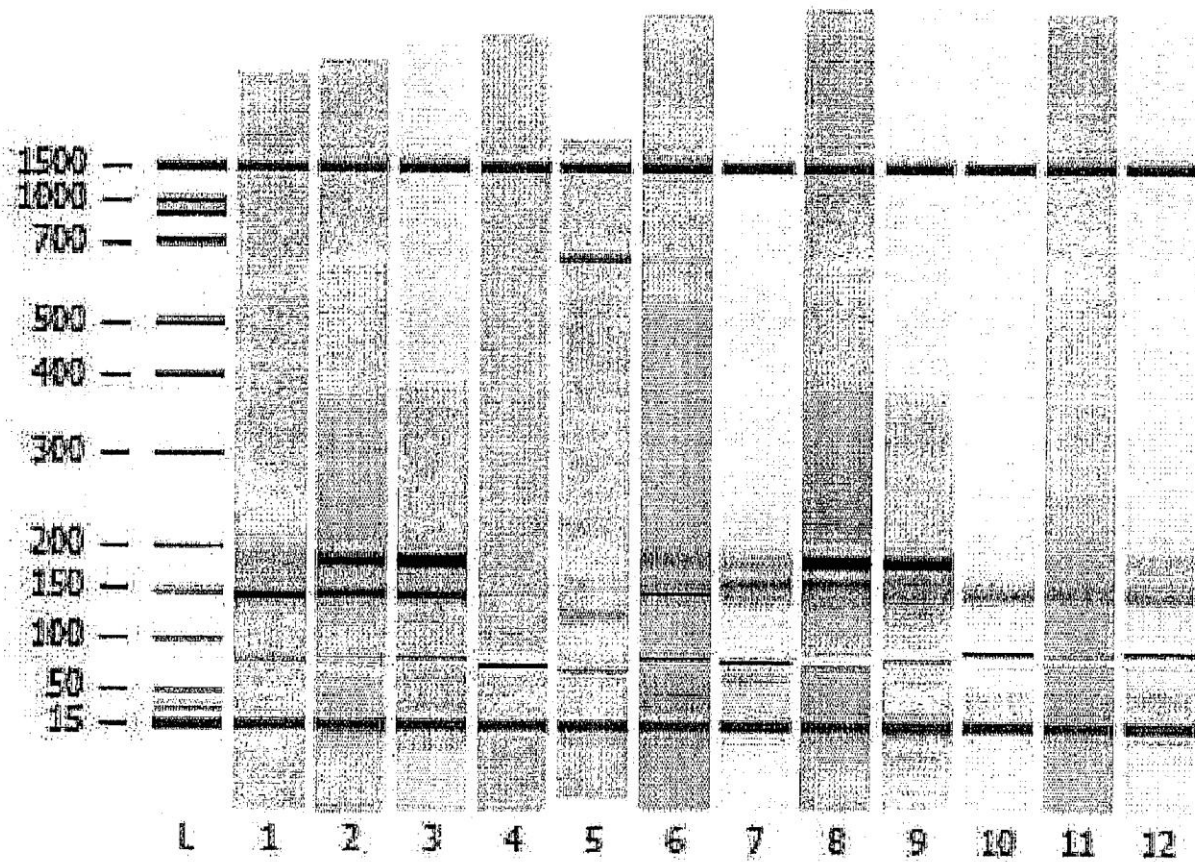


Fig 28A



PCR directa **2.4k 24ng 10nM**

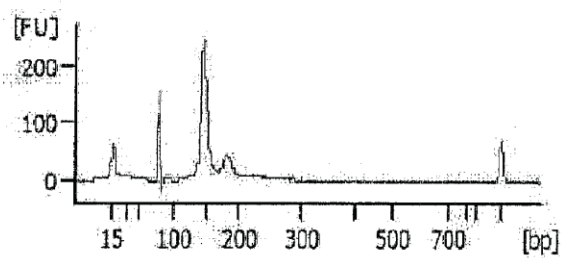


Fig 28B

PCR directa **2.4k 80ng 10nM**

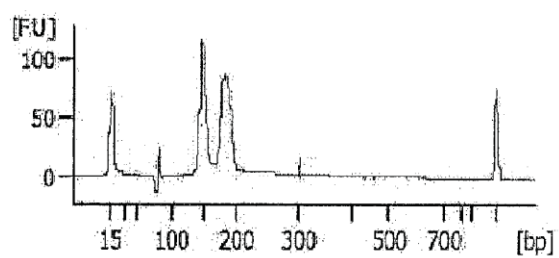


Fig 28C

PCR directa **2.4k 250ng 10nM**

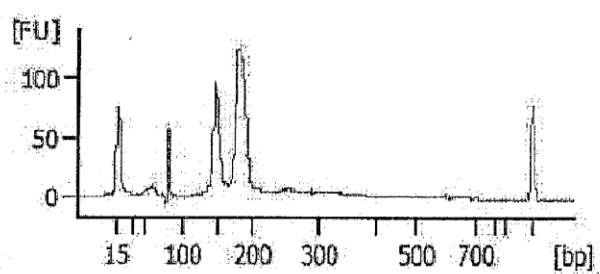


Fig 28D

PCR directa **2.4k 24ng 1nM**

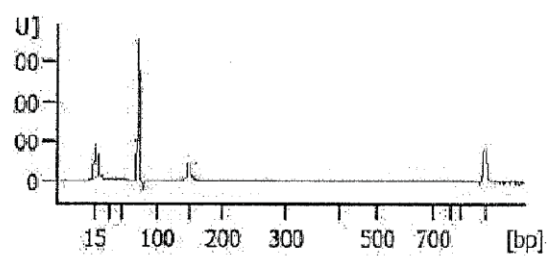


Fig 28E

PCR directa **2.4k 80ng 1nM**

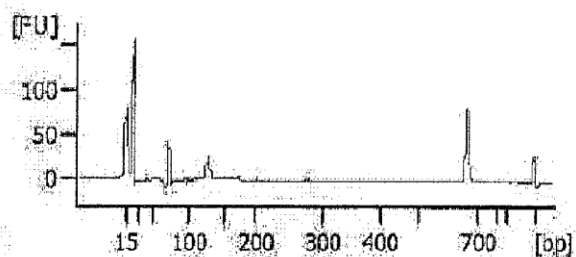


Fig 28F

PCR directa **2.4k 250ng 1nM**

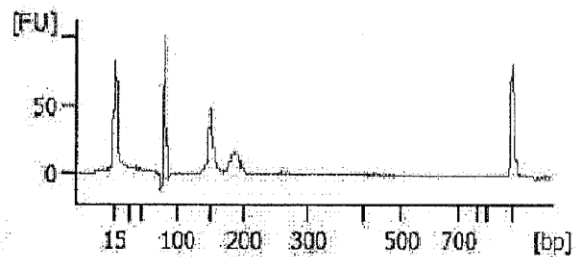


Fig 28G

PCR directa **20k 24ng 10nM**

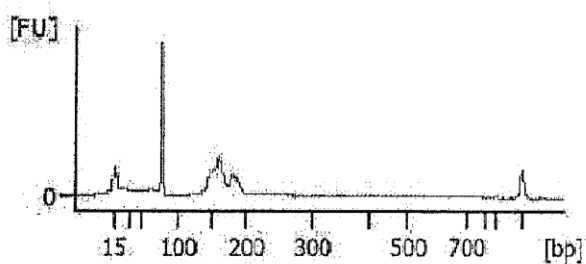


Fig 28H

PCR directa **20k 80ng 10nM**

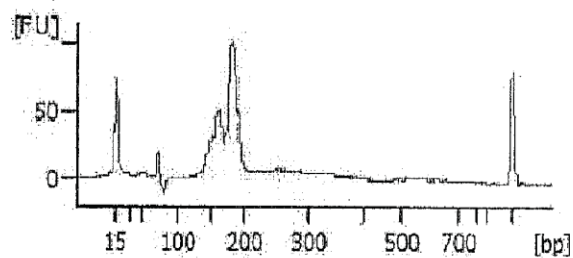


Fig 28I

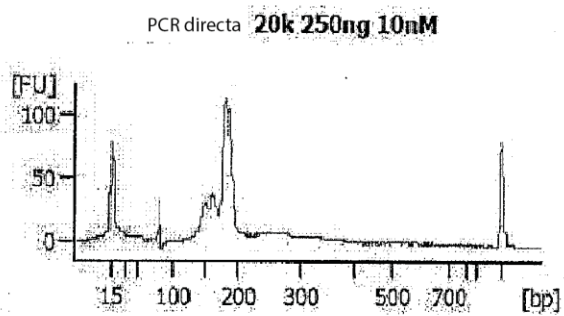


Fig 28J

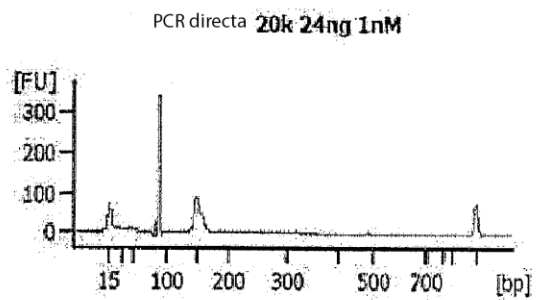


Fig 28K

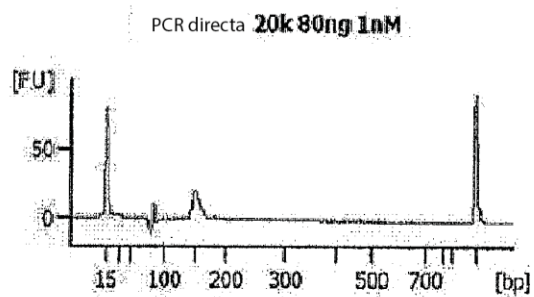


Fig 28L

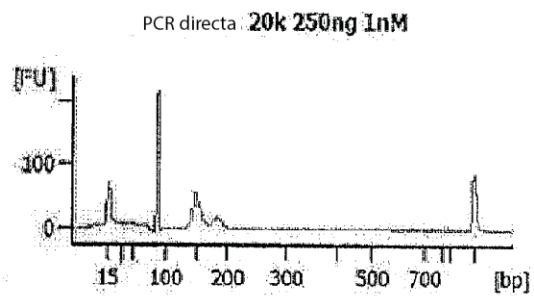


Fig 28M

Fig 29A-29E

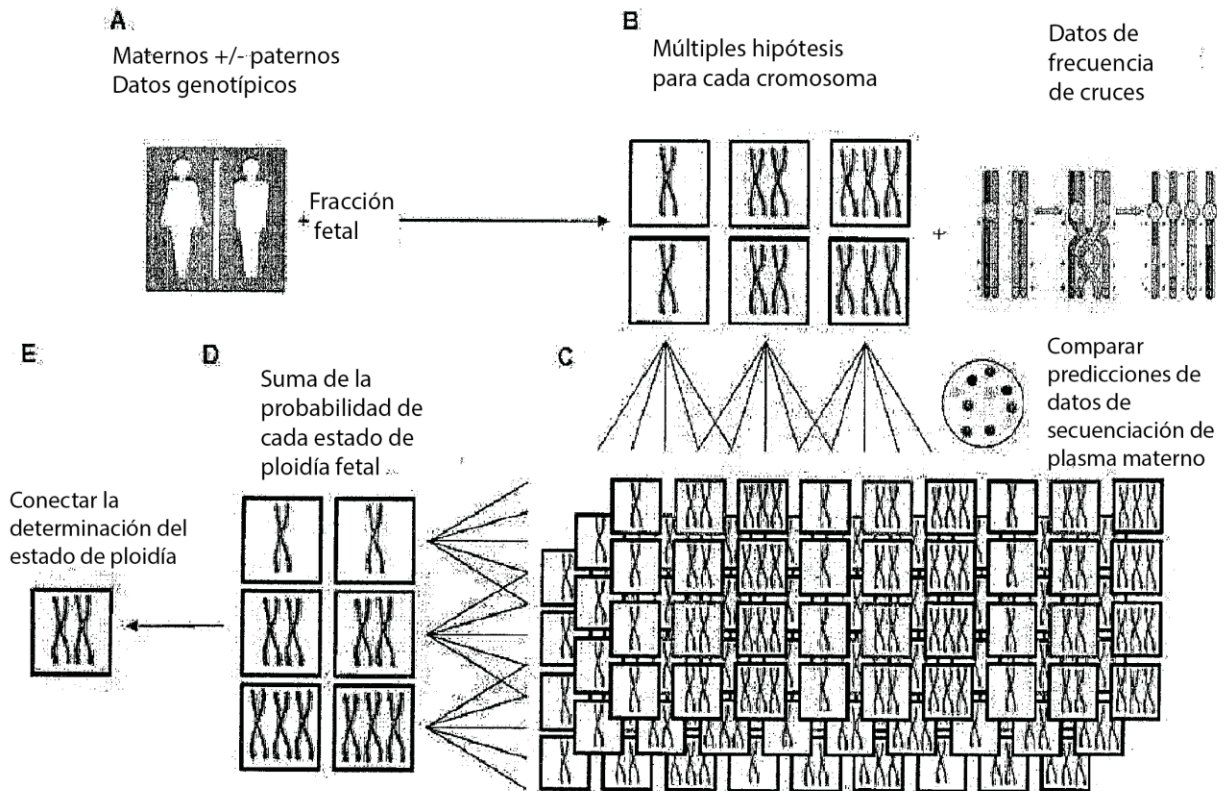


Fig 30A-30C

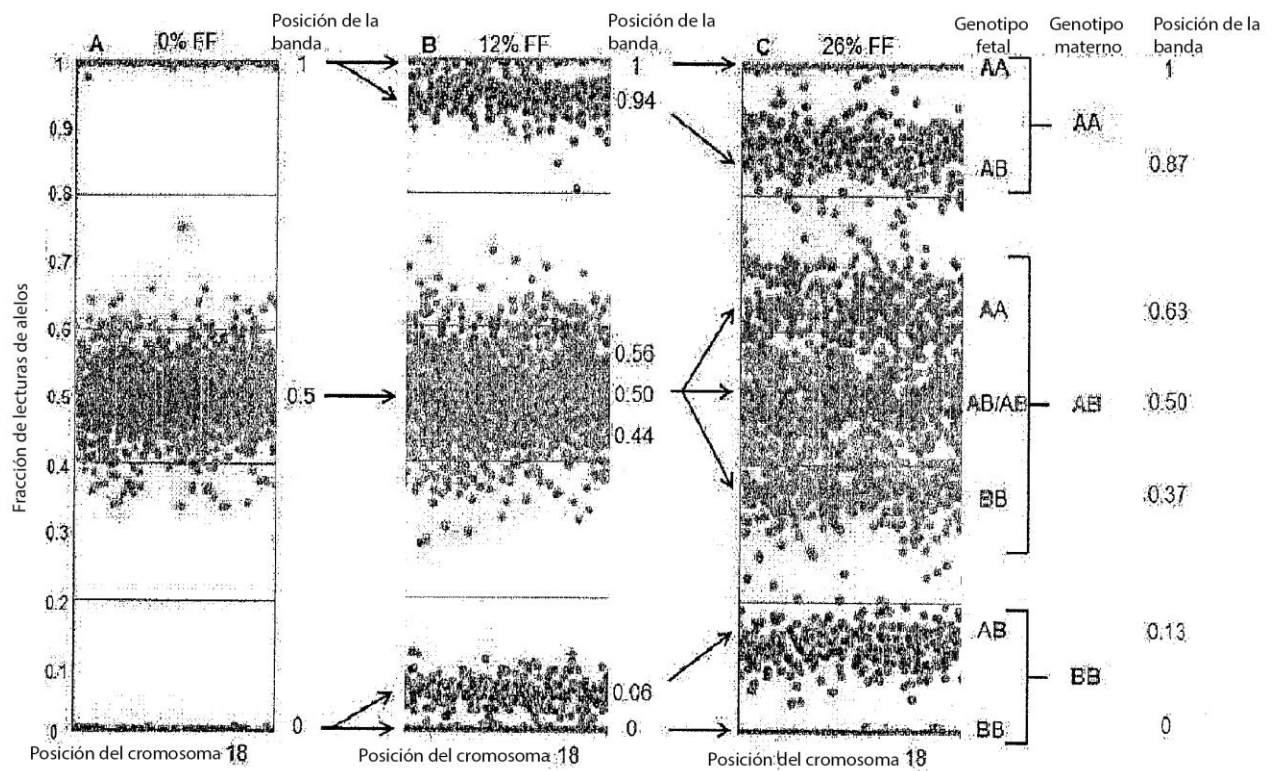


Fig 30D y 30E

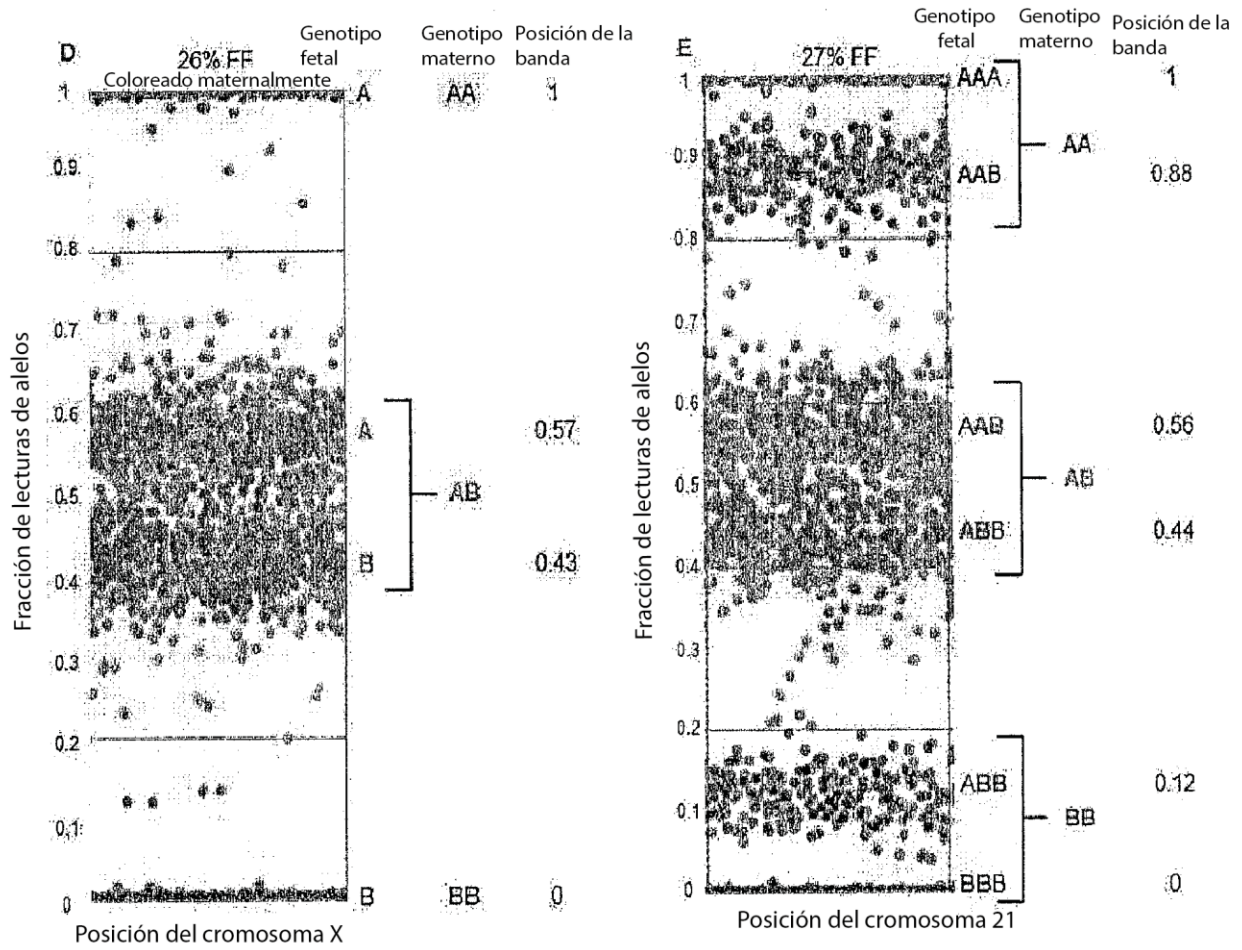


Fig 30F y 30G

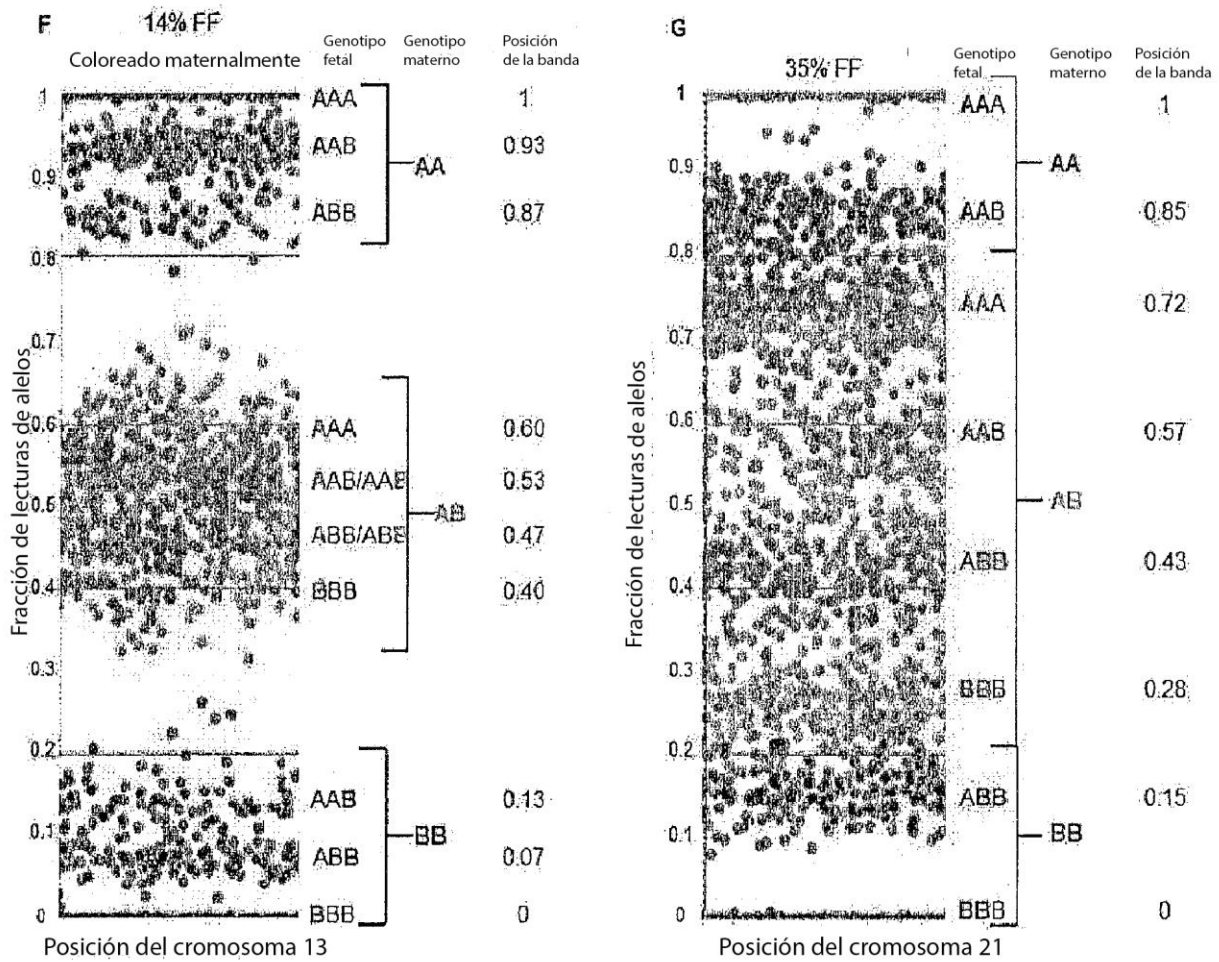


Fig 30H

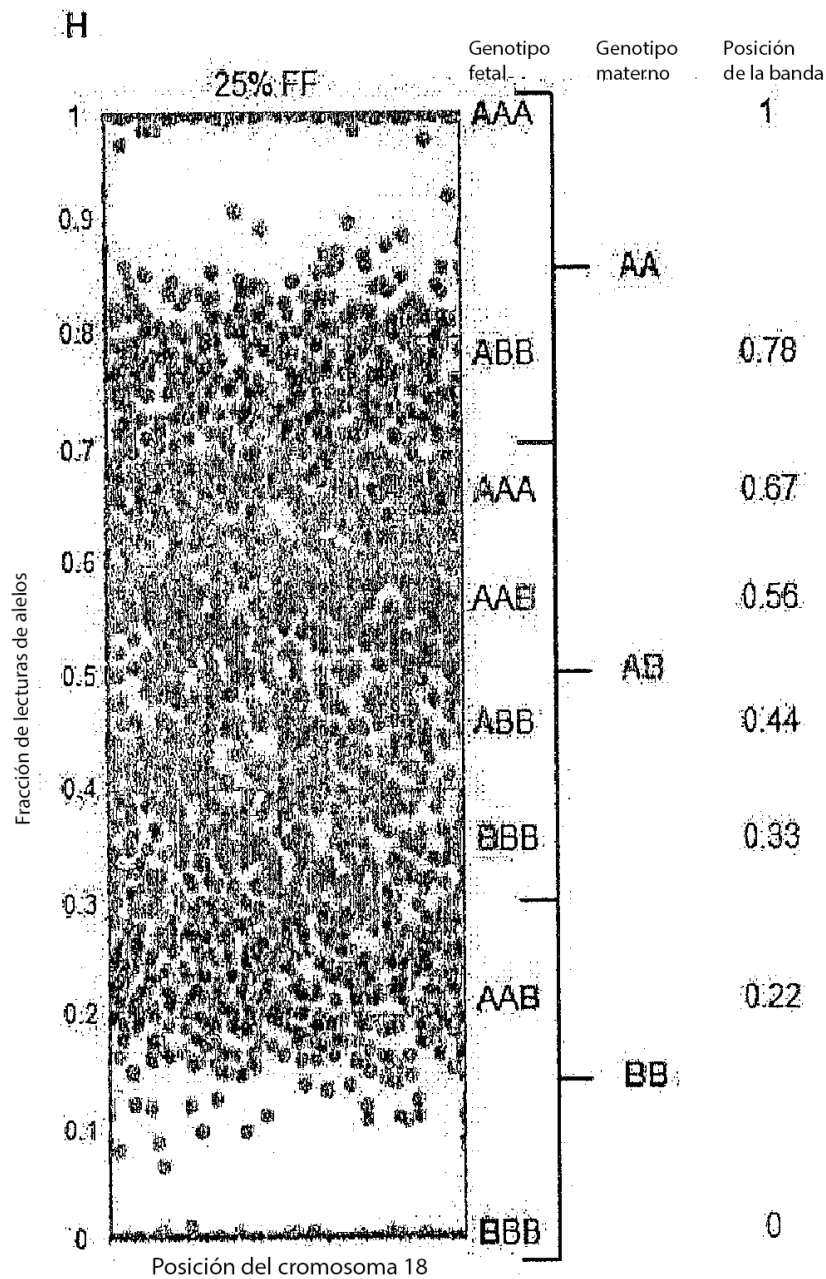


Fig 31A y 31B

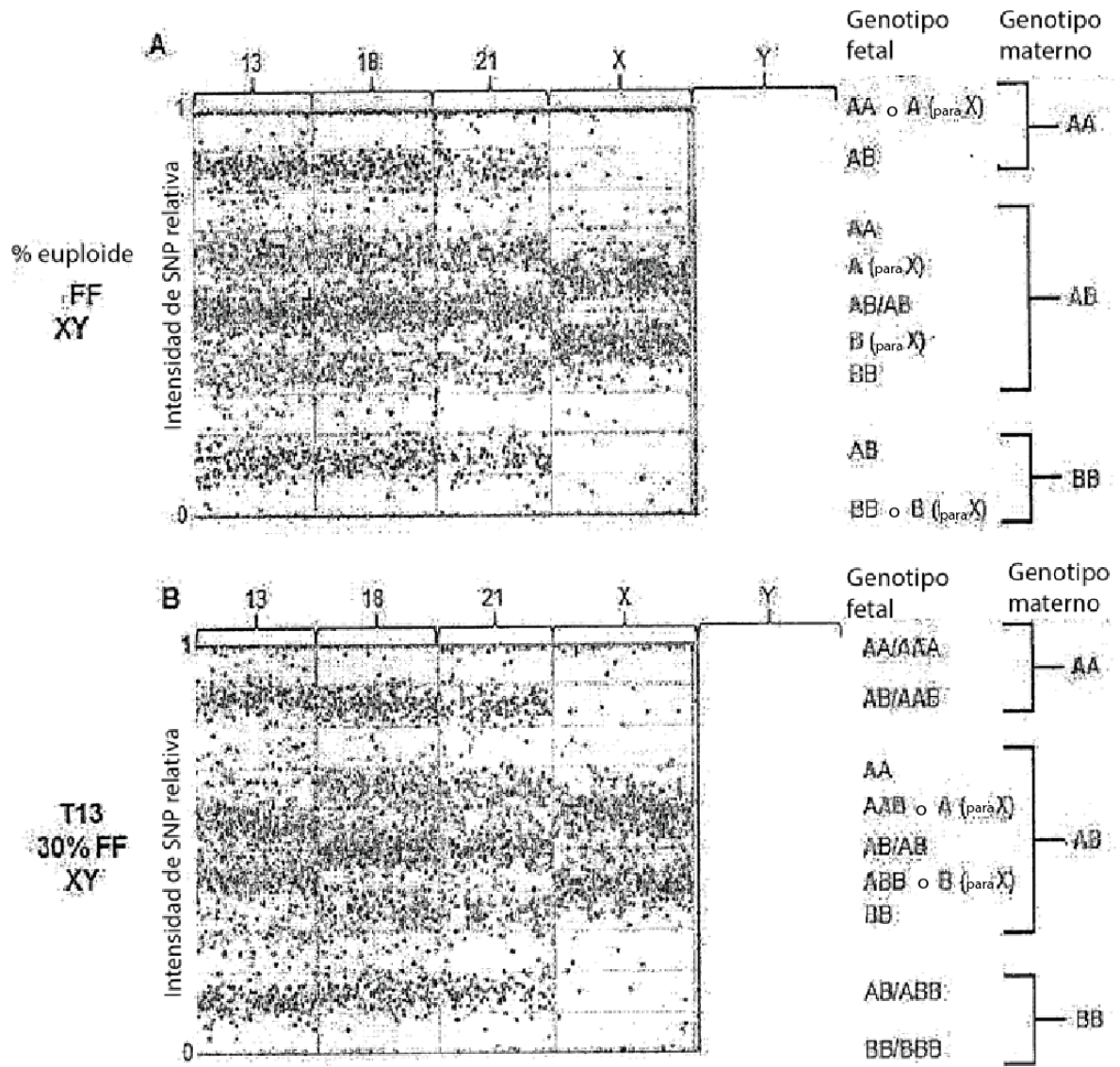
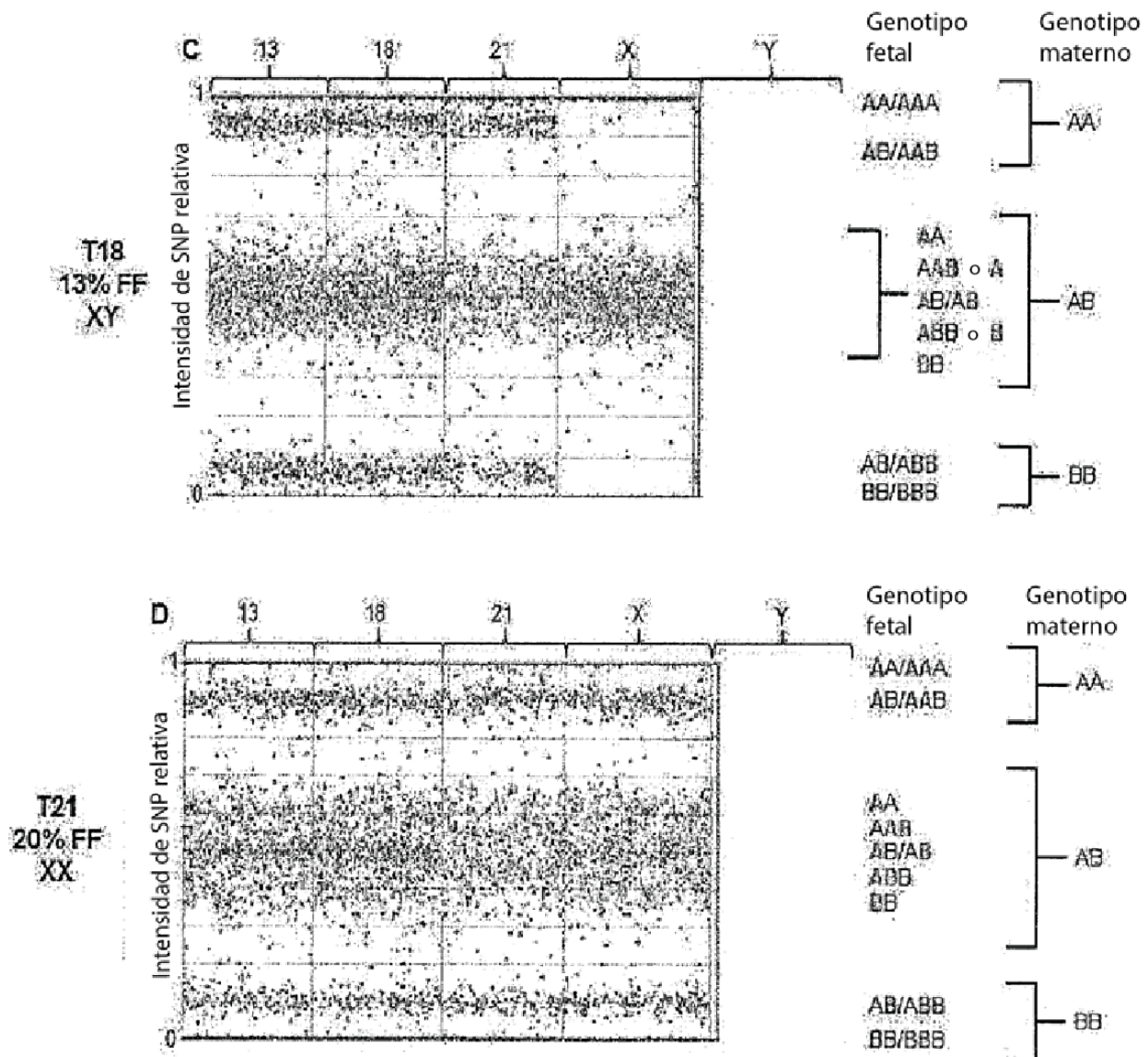


Fig 31C y 31D



Figures 31E and 31F

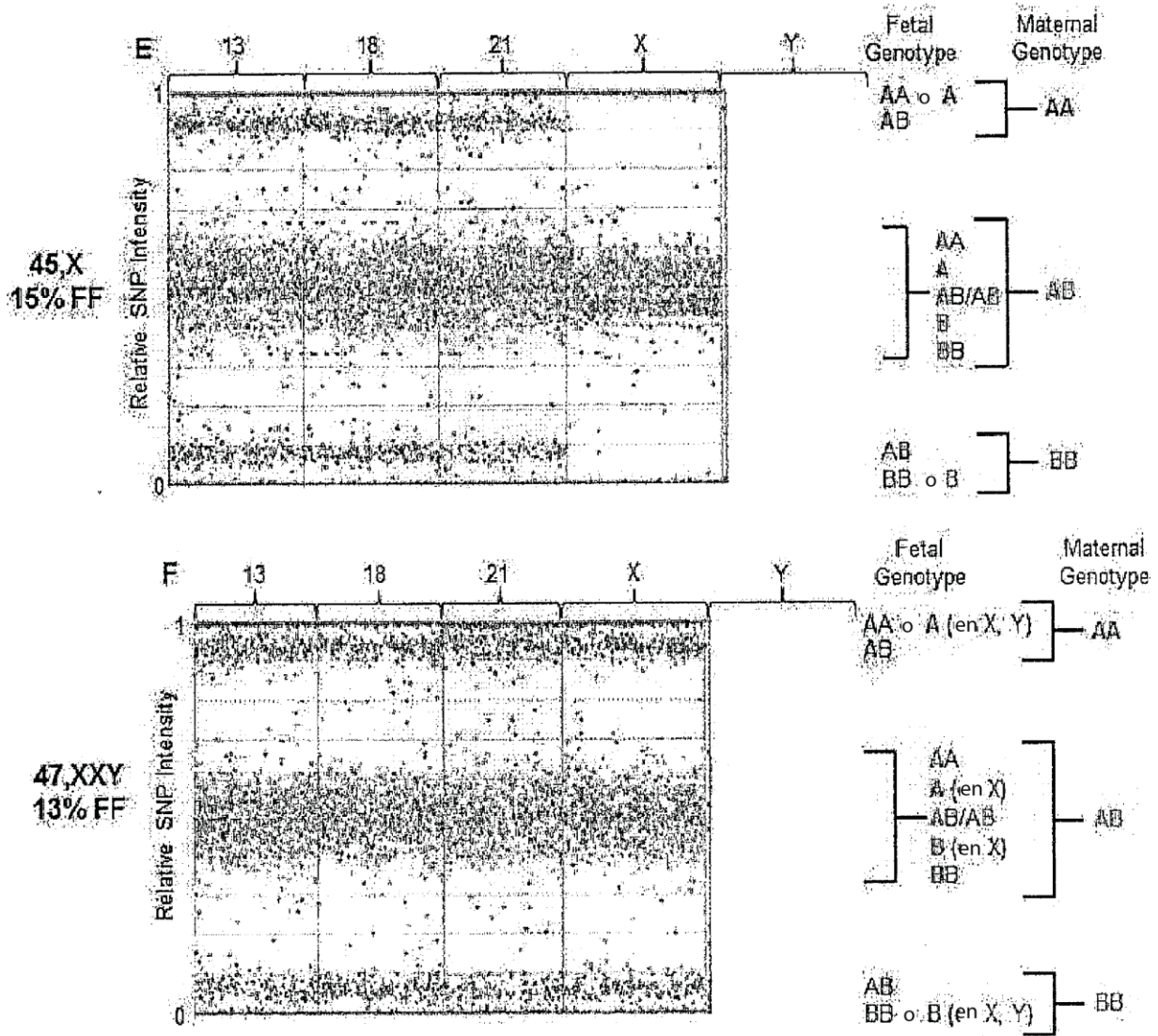


Fig 31G

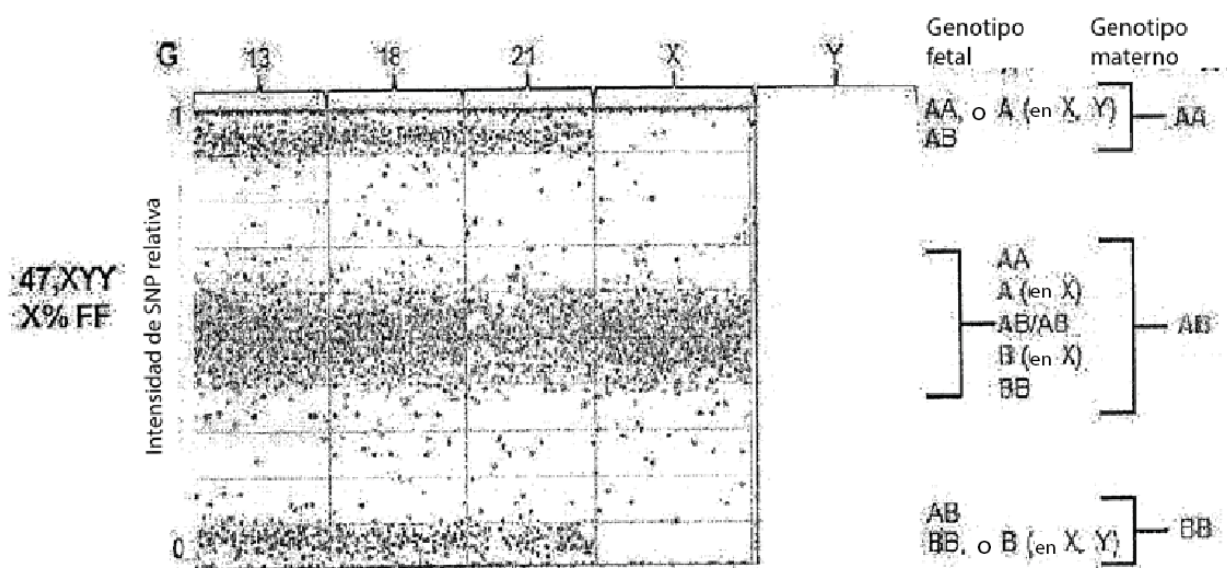


Fig 32

