(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 620 431**

(51) Int. CI.:

G06F 19/18 (2011.01)
C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **04.08.2009 PCT/US2009/052730**

(87) Fecha y número de publicación internacional: **11.02.2010 WO2010017214**

(96) Fecha de presentación y número de la solicitud europea: **04.08.2009 E 09805452 (1)**

(97) Fecha y número de publicación de la concesión europea: **11.01.2017 EP 2321642**

(54) Título: **Métodos para la determinación de alelos y de ploidía**

(30) Prioridad:

04.08.2008 US 137851 P
08.08.2008 US 188343 P
01.10.2008 US 194854 P
07.11.2008 US 198690 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
28.06.2017

(73) Titular/es:

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

(72) Inventor/es:

RABINOWITZ, MATTHEW;
GEMELOS, GEORGE;
BANJEVIC, MILENA;
RYAN, ALLISON y
SWEETKIND-SINGER, JOSHUA

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

DESCRIPCIÓN

Métodos para la determinación de alelos y de ploidía

Campo

La presente divulgación se refiere en general al campo de la adquisición y manipulación de datos genéticos de alta fidelidad a efectos de predicción médica.

Antecedentes

En 2006, en todo el mundo se llevaron a cabo aproximadamente 800 000 ciclos de fertilización *in vitro* (IVF). De los aproximadamente 150 000 ciclos realizados en EE.UU, aproximadamente 10 000 incluyeron diagnóstico genético preimplantacional (PGD). Las técnicas actuales de PGD no están reguladas, son caras y muy poco fiables: las tasas de error del cribado de loci vinculados con enfermedad o aneuplodía son del orden del 10%, cada test de cribado cuesta aproximadamente 5000 $, y típicamente una pareja se ve obligada a elegir entre la prueba de aneuplodía, que afecta a aproximadamente el 50% de los embriones de IVF, o el cribado de loci vinculados a enfermedad, para la célula individual. Hay una gran necesidad de una tecnología asequible, que pueda determinar de forma fiable los datos genéticos de una célula individual, para cribar en paralelo respecto a aneuplodía, enfermedades monogénicas, como la fibrosis quística, y la susceptibilidad a fenotipos de enfermedades complejas para los que se conocen los marcadores genéticos múltiples mediante estudios de asociación de genoma completo.

Hoy la mayoría de PGD van enfocados hacia anormalidades cromosómicas de alto nivel, como aneuploidía y translocaciones balanceadas, siendo las variables primarias la implantación con éxito y un bebé para llevarse a casa. El otro enfoque principal del PGD es el cribado de enfermedades genéticas, donde la variable principal es un bebé sano sin una enfermedad genéticamente hereditaria de la que sean portadores uno o los dos progenitores. En ambos casos, la probabilidad del resultado deseado está aumentada excluyendo embriones genéticamente subóptimos de la transferencia y la implantación en la madre.

El proceso del PGD durante la IVF implica actualmente la extracción de una sola célula de las aproximadamente ocho células de un embrión en fase temprana para su análisis. Aislar células individuales de embriones humanos, aunque es muy técnico, resulta ahora rutinario en las clínicas de IVF. Se han aislado con éxito tanto cuerpos polares como blastómeros. La técnica más común consiste en extraer blastómeros individuales de embriones de 3 días (estadio 6 u 8 células). Los embriones son transferidos a un medio de cultivo celular especial (medio de cultivo estándar sin calcio ni magnesio), y se introduce un orificio en la zona pelucida utilizando una solución ácida, láser o técnicas mecánicas. El técnico emplea entonces una pipeta de biopsia para extraer un solo blastómero con un núcleo visible. Las características del ADN del blastómero único (u ocasionalmente múltiple) se miden utilizando diversas técnicas. Como solo está disponible una sola copia del ADN de una célula, las mediciones del ADN tienen una alta tendencia al error, o son ruidosas. Resulta muy necesaria una técnica que pueda corregir, o hacer más precisas esas mediciones genéticas ruidosas.

Los seres humanos normales tienen dos juegos de 23 cromosomas en cada célula diploide, procediendo una copia de cada progenitor. La aneuploidía, el estado de una célula con cromosoma(s) extra(s) o ausente(s), y la disomía uniparental, el estado de una célula con dos de un cromosoma determinado, ambos procedentes de uno de los progenitores, son considerados responsables de un amplio porcentaje de implantaciones fallidas y abortos, y de algunas enfermedades genéticas. Cuando solo determinadas células de un individuo son aneuploides, se dice que el individuo presenta mosaicismo. La detección de anormalidades cromosómicas puede identificar a individuos o embriones con trastornos tales como el síndrome de Down, el síndrome de Klinefelter y el síndrome de Turner, entre otros, además de incrementar las probabilidades de una gestación con éxito. La identificación de anormalidades cromosómicas resulta especialmente importante al ir aumentando la edad de una madre potencial: entre los 35 y los 40 años, se calcula que entre el 40% y el 50% de los embriones son normales, y por encima de los 40 años, es probable que más de la mitad de los embriones sean anormales. La principal causa de aneuploidía es la no disyunción durante la meiosis. La no disyunción materna constituye aproximadamente el 88% de toda la no disyunción, de la cual aproximadamente el 65% se produce en meiosis I y un 23% en meiosis II. Los tipos comunes de aneuploidía humana incluyen la trisomía de no disyunción de meiosis I, la monosomía y la disomía uniparental. En un tipo particular de trisomía que se produce en no disyunción de meiosis II, o trisomía M2, un cromosoma extra es idéntico a uno de los dos cromosomas normales. Resulta especialmente difícil detectar la trisomía M2. Es muy necesario disponer de un método mejor que pueda detectar muchos o todos los tipos de aneuploidía en la mayoría de los cromosomas o todos ellos, de forma eficiente y con gran precisión, incluyendo un método que pueda diferenciar no solamente la euploidía de la aneuploidía, sino también que pueda diferenciar distintos tipos de aneuploidía entre sí.

El cariotipado, el método tradicional utilizado para la predicción de la aneuploidía y el mosaicismo, está cediendo terreno ante otros métodos de más alto rendimiento y más económicos, como la Citometría de Flujo (FC) y la hibridación fluorescente *in situ* (FISH). Actualmente, en la amplia mayoría de diagnósticos prenatales se utiliza la FISH, que puede identificar amplias aberraciones cromosómicas, y PCR/electroforesis, y que puede realizar numerosas determinaciones de SNPs u otros alelos. Una de las ventajas de la FISH es que resulta más barata que el cariotipado, pero la técnica es compleja y cara y se comprueba generalmente una pequeña selección de cromosomas (habitualmente los cromosomas 13, 18, 21, X, Y; a veces también 8, 9, 15, 16, 17, 22); además, la FISH tiene un bajo nivel de especificidad. Aproximadamente el setenta y cinco por ciento de PGD miden hoy

anormalidades cromosómicas de alto nivel, como la aneuploidía utilizando FISH con unas tasas de error del orden de 10-15%. Existe una gran demanda de un método de cribado de aneuploidía que tenga un mayor rendimiento, un menor coste y una mayor precisión.

El número de alelos genéticos asociados a enfermedad conocidos es de más de 380 según OMIM, y crece continuamente. En consecuencia, resulta cada vez más importante analizar múltiples posiciones en el ADN embriónico, o loci, que estén asociados a fenotipos determinados. Una evidente ventaja del diagnóstico genético preimplantación sobre el diagnóstico prenatal es que evita algunas de las cuestiones éticas sobre posibles vías de acción cuando se detectan fenotipos no deseados. Es necesario un método para un genotipado más extenso de los embriones en la fase preimplantación.

Existen diversas tecnologías avanzadas que permiten el diagnóstico de aberraciones genéticas en uno o unos pocos loci a nivel de célula individual. Entre ellos se incluyen la conversión cromosómica interfase, la hibridación genómica comparativa, la PCR fluorescente, la mini secuenciación y la amplificación de genoma completo. La fiabilidad de los datos generados por todas esas técnicas depende de la calidad de la preparación de ADN. Por consiguiente, se requieren mejores métodos para la preparación de ADN de una sola célula para amplificación y PGD y están siendo objeto de estudio. Todas las técnicas de genotipado, utilizadas en células individuales, números reducidos de células, o fragmentos de ADN, tienen problemas de integridad, principalmente pérdida de alelos (ADO). Esto se incrementa en el contexto de la fertilización in vitro, porque la eficiencia de la reacción de hibridación es baja, y el técnico debe actuar rápidamente para genotipar el embrión dentro del periodo de tiempo de máxima viabilidad del embrión. En US2007/184467 A1 se divulga un sistema y método para limpiar los datos genéticos ruidosos procedentes de individuos objetivo, utilizando datos genéticos de individuos relacionados modificados genéticamente. US 2003/077586 presenta un método y un aparato para combinar predicciones genéticas utilizando redes bayesianas. Es muy necesario disponer de un método que simplifique el problema de una elevada tasa de ADO al medir los datos genéticos procedentes de una célula o un reducido número de ellas, en especial cuando existen limitaciones de tiempo.

Resumen

La invención viene definida en las reivindicaciones del apéndice. En una realización de la presente divulgación, el método presentado permite la reconstrucción de datos genéticos incompletos o ruidosos, incluyendo la determinación de la identidad de alelos individuales, haplotipos, secuencias, inserciones, supresiones, repeticiones, y la determinación del número de copias de cromosomas en un individuo objetivo, todo ello con alta fidelidad, utilizando datos genéticos secundarios como fuente de información. Mientras la divulgación va enfocada hacia los datos genéticos procedentes de sujetos humanos, y más específicamente en embriones aún no implantados o fetos en desarrollo, así como en individuos relacionados, debe tenerse en cuenta que los métodos divulgados son aplicables a los datos genéticos de una gama de organismos, en una gama de contextos. Las técnicas descritas para limpiar datos genéticos resultan de la máxima importancia en el contexto del diagnóstico preimplantación durante la fertilización in vitro, el diagnóstico prenatal en conjunción con la amniocentesis, la biopsia de vellosidad coriónica, el muestreo de tejido fetal y el diagnóstico prenatal no invasivo, donde se aísla una pequeña cantidad de material genético fetal de la sangre materna. El empleo de este método puede facilitar los diagnósticos enfocados hacia las enfermedades hereditarias, las predicciones del número de copias de cromosomas, el incremento de probabilidad de defectos o anormalidades, y la predicción de la propensión a fenotipos de enfermedad y no enfermedad en individuos, para mejorar las decisiones clínicas y de estilo de vida.

En una realización de la presente invención, un método de determinación de un estado de ploidía de por lo menos un cromosoma en un individuo objetivo incluye obtener datos genéticos del individuo objetivo y de uno o más individuos relacionados; crear un conjunto de por lo menos una hipótesis de estado de ploidía para cada uno de los cromosomas del individuo objetivo; utilizar una o más técnicas especializadas para determinar una probabilidad estadística para cada hipótesis de estado de ploidía en el conjunto, para cada técnica especializada usada, dados los datos genéticos obtenidos; combinar, para cada hipótesis de estado de ploidía, las probabilidades estadísticas determinadas por la o las técnicas especializadas; y determinar el estado de ploidía de cada uno de los cromosomas en el individuo objetivo en base a las probabilidades estadísticas combinadas de cada una de las hipótesis de estado de ploidía.

En una realización de la presente divulgación, un método para determinar un estado alélico en un conjunto de alelos, en un individuo objetivo, y de uno o ambos progenitores del individuo objetivo, y opcionalmente de uno o más individuos relacionados, incluye la obtención de datos genéticos del individuo objetivo, y de uno o de ambos progenitores, y de cualquier individuo relacionado; crear un conjunto de por lo menos una hipótesis alélica para el individuo objetivo, y para uno o ambos progenitores, y opcionalmente para uno o más individuos relacionados, donde las hipótesis describen posibles estados alélicos en el conjunto de alelos; determinar una probabilidad estadística para cada hipótesis alélica en el conjunto de hipótesis con los datos genéticos obtenidos; y determinar el estado alélico de cada uno de los alelos en el conjunto de alelos del individuo objetivo, y de uno o ambos progenitores, y opcionalmente de uno o más individuos relacionados, en base a las probabilidades estadísticas de cada una de las hipótesis alélicas.

En una realización de la presente divulgación, un método para la determinación de un estado de ploidía de por lo menos un cromosoma en un individuo objetivo incluye obtener datos genéticos del individuo objetivo, y de ambos progenitores del individuo objetivo, y de uno o más hermanos del individuo objetivo, donde los datos genéticos incluyen datos relacionados con por lo menos un cromosoma; determinar un estado de ploidía de por lo menos un cromosoma en el individuo objetivo, y en uno o más hermanos del individuo objetivo utilizando una o más técnicas especializadas, donde ninguna de las técnicas especializadass requiere como entrada datos genéticos por fases; determinar datos genéticos por fases del individuo objetivo, y de los padres del individuo objetivo, y de uno o más hermanos del individuo objetivo, utilizando un método informático, y los datos genéticos obtenidos del individuo objetivo y de los padres del individuo objetivo, y de uno o más hermanos del individuo objetivo que se identificaron como euploides en ese cromosoma; y redeterminar el estado de ploidía de por lo menos un cromosoma del individuo objetivo, utilizando una o más técnicas especializadas, por lo menos una de las cuales requiere la entrada de datos genéticos por fases, y los datos genéticos por fases determinados del individuo objetivo, y de los padres del individuo objetivo, y de uno más hermanos del individuo objetivo.

En una realización de la presente divulgación, el método utiliza el conocimiento de los datos genéticos del embrión objetivo, los datos genéticos de la madre y del padre, como muestras de tejido diploide, y posiblemente datos genéticos de uno o más de lo siguiente: esperma del padre, muestras haploides de la madre o blastómeros del mismo u otros embriones derivados de los gametos de la madre y del padre, junto con el conocimiento del mecanismo de meiosis y la medición imperfecta de ADN embriónico objetivo, para reconstruir *in silico* el ADN embriónico en la localización de loci clave con un alto grado de confianza. En un aspecto de la presente divulgación, los datos genéticos derivados de otros individuos relacionados, tales como otros embriones, hermanos y hermanas, abuelos u otros parientes, pueden utilizarse también para aumentar la fidelidad del ADN embriónico reconstruido. En una realización de la presente divulgación, esos datos genéticos pueden ser utilizados para determinar el estado de ploidía en uno o más cromosomas del individuo. En un aspecto de la presente divulgación, cada uno del conjunto de datos genéticos medidos de un grupo de individuos relacionados se utiliza para aumentar la fidelidad de los otros datos genéticos. Es importante advertir que, en un aspecto de la presente divulgación, los datos genéticos parentales y otros datos genéticos secundarios permiten la reconstrucción no solamente de SNPs que fueron medidos deficientemente, sino también de inserciones, exclusiones, repeticiones y de SNPs o regiones completas de ADN que no fueron medidos en absoluto. En otro aspecto de la presente divulgación, los datos genéticos del individuo objetivo, junto con los datos genéticos secundarios de individuos relacionados se utilizan para determinar el estado de ploidía, o número de copias en uno, varios o todos los cromosomas del individuo.

En una realización de la presente divulgación, los datos genómicos fetales o embriónicos, con o sin el uso de datos genéticos procedentes de individuos relacionados, pueden ser utilizados para detectar si la célula es aneuploide; es decir, si en una célula está presente un número erróneo de un cromosoma, o si un número erróneo de cromosomas sexuales están presentes en la célula. También pueden utilizarse los datos genéticos para detectar la disomía uniparental, un trastorno en el que están presentes dos de un cromosoma determinado, ambos procedentes de un progenitor. Esto se hace creando un conjunto de hipótesis sobre los potenciales estados del ADN, y comprobando qué hipótesis tiene la mayor probabilidad de ser real teniendo en cuenta los datos medidos. Hay que advertir que el uso de datos de genotipado de alto rendimiento para el cribado de la aneuploidía permite utilizar un solo blastómero de cada embrión para medir múltiples loci vinculados a enfermedad, así como el cribado de la aneuploidía.

En una realización de la presente divulgación, las mediciones directas de la cantidad de material genético, amplificado o no, presente en diversos loci, pueden ser utilizadas para detectar la monosomía, la disomía uniparental, la trisomía coincidente, la trisomía no coincidente, la tetrasomía, y otros estados de aneuploidía. En una realización de la presente divulgación se aprovecha el hecho de que, en determinadas condiciones, el nivel medio de amplificación y salida de señal de medición es invariable en los cromosomas, y así la cantidad media de material genético medido en un conjunto de loci vecinos será proporcional al número de cromosomas homólogos presentes, y el estado de ploidía puede ser determinado de forma estadísticamente significativa. En otra realización, diferentes alelos tienen perfiles de amplificación característicos distintos estadísticamente, dado un contexto de padres determinado y un estado de ploidía determinado; estas diferencias características pueden utilizarse para determinar el estado de ploidía del cromosoma.

En una realización de la presente divulgación, el estado de ploidía, como se determina por un aspecto de la presente divulgación, puede utilizarse para seleccionar la entrada apropiada para una realización de determinación de alelo de la presente divulgación. En otro aspecto de la presente divulgación, los datos genéticos por fase reconstruidos del individuo objetivo y/o de uno o más individuos relacionados pueden ser utilizados como entrada para un aspecto de determinación de ploidía de la presente divulgación. En una realización de la presente divulgación, la salida de un aspecto de la presente divulgación puede ser utilizada como entrada, o para ayudar a seleccionar la entrada apropiada para otros aspectos de la presente divulgación en un proceso iterativo.

Considerando los beneficios de esta divulgación resultará evidente para un experto en la técnica, que los diversos aspectos y realizaciones de esta divulgación pueden ser implementados en combinación o por separado.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Las realizaciones que se divulgan aquí se explicarán en más detalle con referencia a los dibujos adjuntos, donde estructuras iguales serán designadas por números iguales en las distintas vistas. Los dibujos que se muestran no

son necesariamente a escala, poniéndose el énfasis en general en ilustrar los principios de las realizaciones divulgadas aquí.

La Figura 1 muestra las curvas de función de distribución acumulativa de un cromosoma disómico. Las curvas de función de distribución acumulativa se muestran para cada uno de los contextos parentales.

Las Figuras 2A-2D muestran las curvas de función de distribución acumulativa de cromosomas con diversos estados de ploidía. La Figura 2A muestra una curva de función de distribución acumulativa de un cromosoma disómico. La Figura 2B muestra una curva de función de distribución acumulativa de un cromosoma nulisómico. La Figura 2C presenta una curva de función de distribución acumulativa de un cromosoma monosómico. La Figura 2D presenta una curva de función de distribución acumulativa de un cromosoma trisómico materno. La relación entre curvas de función de distribución acumulativa de distintos contextos parentales varía con el estado de ploidía.

La Figura 3 muestra una distribución de hipótesis de diversos estados de ploidía utilizando la técnica de Media Cromosómica Completa que se divulga aquí. Se muestran estados de ploidía monosómicos, disómicos y trisómicos.

Las Figuras 4A y 4B muestran una distribución de los datos genéticos de cada uno de los progenitores utilizando la técnica de Presencia de Progenitores que se divulga aquí. La Figura 4A muestra una distribución donde están presentes datos genéticos de cada uno de los padres. La Figura 4B muestra una distribución donde están ausentes datos genéticos de cada uno de los padres.

La Figura 5 muestra qué distribuciones de las mediciones genéticas del padre varían cuando están presentes y no presentes datos genéticos utilizando la técnica de Presencia de Progenitores.

La Figura 6 muestra una representación gráfica de un conjunto de Polimorfismos de un Solo Nucleótido. La intensidad anormalizada de una salida de canal se representa contra la otra.

La Figura 7 muestra una representación gráfica de un conjunto de Polimorfismos de un Solo Nucleótido. La intensidad normalizada de la salida de un canal se representa contra la otra.

Las Figuras 8A-8C muestran ajustes de curva para datos alélicos para diferentes hipótesis de ploidía. La Figura 8A muestra ajustes de curva para datos alélicos para cinco hipótesis de ploidía distintas utilizando el método Kernel divulgado aquí. La Figura 8B muestra ajustes de curva para datos alélicos para cinco distintas hipótesis de ploidía utilizando un Ajuste Gaussiano divulgado aquí. La Figura 8C presenta un histograma de los datos alélicos medidos de un contexto, AA|BB - BB|AA.

La Figura 9 muestra una representación gráfica de meiosis.

Las Figuras 10A y 10B muestran la tasa de éxito real versus la confianza de determinación alélica para recipientes grandes. La Figura 10A muestra la media de tasa de éxito representada gráficamente contra una confianza prevista. La Figura 10B muestra la población relativa del recipiente.

Las Figuras 11A y 11B muestran la tasa de éxitos real versus la confianza de la determinación de alelos para recipientes pequeños. La Figura 11A muestra la tasa media real de éxitos representada gráficamente contra una confianza prevista. La Figura 11B muestra la población relativa del recipiente.

Las Figuras 12A y 12B muestran la confianza de alelos representada gráficamente junto a un cromosoma para determinar una localización de un cruce. La Figura 12A muestra las confianzas de la determinación de alelos para un conjunto de alelos localizados a lo largo de un cromosoma, como promedio sobre un conjunto de alelos vecinos. Los conjuntos de alelos utilizan distintos métodos. La Figura 12B muestra una localización de un cruce a lo largo del cromosoma.

Mientras los dibujos identificados más arriba exponen realizaciones divulgadas aquí, se contemplan también otras realizaciones, como se indica en la discusión.

Descripción detallada

La invención viene definida en las reivindicaciones adjuntas. En una realización de la presente divulgación, puede determinarse el estado genético de una célula o grupo de células. La determinación del número de copias es el concepto de establecer el número y la identidad de los cromosomas en una célula determinada, en un grupo de células o en un conjunto de ácido desoxirribonucleico (ADN). El concepto de la determinación de alelos consiste en identificar el estado alélico de una célula determinada, un grupo de células, o un conjunto de ADN, en un conjunto de alelos, incluyendo Polimorfismos de un Solo Nucleótido (SNPs), inserciones, supresiones, repeticiones, secuencias u otra información de pares de base. La presente divulgación permite la determinación de aneuploidía, así como la determinación alélica, de una célula sola, u otro pequeño conjunto de ADN, siempre que esté disponible el genoma de por lo menos uno o ambos progenitores. En algunos aspectos de la presente divulgación se utiliza el concepto de que dentro de un conjunto de individuos relacionados habrá conjuntos de ADN que serán casi idénticos, y que, utilizando las mediciones de los datos genéticos junto a un conocimiento del mecanismo de la meiosis, es posible determinar el estado genético de los individuos pertinentes, por deducción, con mayor precisión de lo que sería posible utilizando las mediciones individuales por sí solas. Esto se hace determinando qué segmentos de cromosomas de individuos relacionados estaban involucrados en la formación de gametos y, cuando sea necesario, dónde pueden haberse producido cruces durante la meiosis, y por consiguiente qué segmentos de los genomas de

individuos relacionados puede esperarse que sean casi idénticos a secciones del genoma objetivo. Esto puede resultar especialmente útil en el caso del diagnóstico genético preimplantación, o el diagnóstico prenatal, donde está disponible una cantidad limitada de ADN, y donde la determinación del estado de ploidía de un objetivo, en esos casos un embrión o feto, tiene un elevado impacto clínico.

Existen muchas técnicas matemáticas posibles para determinar el estado de aneuploidía de un conjunto de datos genéticos objetivo. En esta divulgación se comentan algunas de tales técnicas, pero otras podrían utilizarse igualmente bien. En una realización de la presente divulgación, pueden utilizarse datos cualitativos y/o cuantitativos. En una realización de la presente divulgación pueden utilizarse datos parentales para deducir datos del genoma objetivo que puedan haber sido medidos deficientemente, de forma incorrecta o de ningún modo. En una realización, se pueden usar datos genéticos deducidos de uno o más individuos para aumentar la probabilidad de que el estado de ploidía sea determinado correctamente. En una realización de la presente divulgación, pueden utilizarse diversas técnicas, cada una de las cuales puede descartar determinados estados de ploidía, o determinar la probabilidad relativa de determinados estados de ploidía, y las probabilidades de esas predicciones pueden ser combinadas para obtener una predicción del estado de ploidía con mayor confianza de la que es posible utilizando solamente una técnica. Se puede calcular una confianza para cada determinación cromosómica realizada.

Las mediciones de DNA, obtenidas por técnicas de secuenciación, matrices de genotipado o cualquier otra técnica, contienen cierto grado de error. La confianza relativa en una medición determinada de ADN se ve afectada por diversos factores, incluyendo el método de amplificación, la tecnología utilizada para medir el ADN, el protocolo seguido, la cantidad de ADN usada, la integridad del ADN utilizado, el operador, y la frescura de los reactivos, por mencionar solo algunos. Una forma de incrementar la precisión de las mediciones es aplicar técnicas basadas en informática para inferir el estado genético correcto del ADN en el objetivo, en base al conocimiento del estado genético de individuos relacionados. Dado que se espera que los individuos relacionados compartan determinado aspecto de su estado genético, cuando se consideran en su conjunto los datos genéticos de diversos individuos relacionados, es posible identificar probables errores en las mediciones, e incrementar la precisión del conocimiento de los estados genéticos de todos los individuos relacionados. Además, se puede calcular una confianza para cada determinación efectuada.

En algunos aspectos de la presente divulgación, el individuo objetivo es un embrión, y el motivo de aplicar el método divulgado a los datos genéticos del embrión es permitir a un médico u otro agente proceder a una elección informada respecto a qué embrión o embriones deben ser implantados durante la IVF. En otro aspecto de la presente divulgación, el individuo objetivo es un feto, y el motivo de aplicar el método divulgado a los datos genéticos del feto es permitir a un médico u otro agente proceder a una elección informada sobre posibles decisiones clínicas u otras acciones a tomar respecto al feto.

Definiciones

SNP (Single Nucleotide Polymorphism, Polimorfismo de un Solo Nucleótido) puede referirse a un solo nucleótido que puede ser distinto 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.

Determinar un SNP puede referirse al acto de tomar una decisión sobre el estado real de un par de bases determinado, teniendo en cuenta la evidencia directa e indirecta.

Secuencia puede referirse a una secuencia de ADN o una secuencia genética. Puede referirse a la estructura primaria, física de la molécula o cadena de ADN en un individuo.

Locus puede referirse 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 supresión, o el punto de otra variación genética relevante. SNPs vinculados a enfermedad pueden referirse también a loci vinculados a enfermedad.

Alelo puede referirse a los genes que ocupan un locus determinado.

Determinar un alelo puede referirse al acto de determinar el estado genético de un locus concreto de ADN. Esto puede implicar determinar un SNP, diversos SNPs, o determinar si hay o no una inserción o supresión en ese locus, o determinar el número de inserciones que pueden estar presentes en ese locus, o determinar si está presente en ese locus alguna otra variante genética.

Correcta determinación de alelo puede referirse a una determinación de alelo que refleja correctamente el estado real del verdadero material genético de un individuo.

Limpiar datos genéticos puede referirse a la acción de tomar datos genéticos imperfectos y corregir algún error o todos ellos, o completar datos que faltaban en uno o más loci. En el contexto de esta divulgación, esto puede implicar utilizar los datos genéticos de individuos relacionados y el método que se describe aquí.

Aumentar la fidelidad de las determinaciones de alelos puede referirse a la acción de limpiar datos genéticos respecto a un conjunto de alelos.

Datos genéticos imperfectos puede referirse a datos genéticos con algo de lo siguiente: pérdida de alelos, mediciones imprecisas de pares de base, mediciones incorrectas de pares de base, falta de mediciones de pares de base, mediciones dudosas de inserciones o supresiones, determinaciones dudosas del número de copias de segmentos cromosómicos, señales falsas, ausencia de mediciones, otros errores o combinaciones de lo anterior.

Datos genéticos ruidosos puede referirse a datos genéticos imperfectos, llamados también datos genéticos incompletos.

Datos genéticos sin limpiar puede referirse a datos genéticos tal como se han medido, es decir, sin utilizar ningún método para corregir la presencia de ruido o errores en los datos genéticos en bruto; llamado también datos genéticos crudos.

Confianza puede referirse 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 copias de cromosoma”, o “determinación del número de copias (CNC), puede ser la acción de determinar la cantidad e identidad cromosómica de uno o más cromosomas presentes en una célula.

Aneuploidía puede referirse al estado en que están presentes en una célula un número erróneo de cromosomas. 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. Cuando se refiere a un solo cromosoma, puede referirse al caso en el que más o menos de dos cromosomas homólogos estén presentes.

Estado de ploidía puede ser la cantidad e identidad cromosómica de uno o más cromosomas en una célula.

Identidad cromosómica puede referirse al número de cromosomas referente. 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 de los padres. 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 SNPs en el ADN, puede referirse a los haplotipos de fase del material genético, y puede referirse a la secuencia del ADN, incluyendo inserciones, supresiones, repeticiones y mutaciones. También puede referirse al estado de ploidía de uno o más cromosomas, segmentos cromosómicos o conjunto de segmentos cromosómicos.

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

Estado alélico puede referirse 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.

Error de copia emparejada, también “aneuploidía de cromosoma emparejado”, o “MCA”, puede ser 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 mitosis, y puede ser denominada error de no disyunción mitótica.

Error de copia no emparejada, también “Aneuploidía de Cromosoma Único” o “UCA”, puede ser 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 referirse como error meiótico.

Mosaicismo puede referirse a un conjunto de células en un embrión u otro individuo que son heterogéneas respecto a su estado de ploidía.

Cromosomas homólogos pueden ser cromosomas que contienen el mismo conjunto de genes que pueden normalmente emparejarse durante la meiosis.

Cromosomas idénticos pueden ser 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) puede referirse a la situación en la que no se detecta uno de los pares de base en un conjunto de pares de base de cromosomas homólogos en un alelo determinado.

Pérdida de locus (Locus Drop Out, LDO) puede referirse a la situación en la que no se detectan ambos pares de base en un conjunto de pares de base 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.

Región cromosómica puede referirse a un segmento de cromosoma o un cromosoma completo.

Segmento de un cromosoma puede referirse a una sección de un cromosoma que en cuanto a tamaño puede ir de un par de base al cromosoma completo.

Cromosoma puede referirse a un cromosoma completo, o también a un segmento o sección de un cromosoma.

Copias puede referirse al número de copias de un segmento de cromosoma, 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 es una combinación de alelos en múltiples loci que son transmitidos 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 (SNPs) en una sola cromátida que están asociados estadísticamente.

Datos haplotípicos llamados también “datos por fase” 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 puede referirse 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 puede referirse a los datos genéticos donde se ha determinado el haplotipo.

Datos de determinación de alelos por fases puede referirse a los datos alélicos, donde ha sido determinado el estado alélico, incluyendo datos del haplotipo. En una realización, los datos de determinación de alelos parentales por fase, determinados por un método informático, pueden ser utilizados como datos genéticos obtenidos en un aspecto de determinación de ploidía de la presente divulgación.

Datos genéticos no ordenados pueden referirse a datos combinados derivados de mediciones de dos o más cromosomas en un genoma diploide o poliploide; ej. las copias materna y paterna de un cromosoma concreto en un genoma diploide.

Datos genéticos “de” “en” “desde” o "sobre " un individuo puede referirse a datos describiendo aspectos del genoma de un individuo. Puede referirse a un locus o un conjunto de loci, secuencias parciales o enteras, cromosomas parciales o enteros o el genoma completo.

Hipótesis puede referirse a un conjunto de posibles estados 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”, puede referirse a una hipótesis sobre cuántas copias de un cromosoma determinado hay 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.

Hipótesis alélica puede referirse a un posible estado alélico de un conjunto determinado de alelos. Un conjunto de hipótesis alélicas puede referirse a un conjunto de hipótesis que describan, juntas, todos los posibles estados alélicos en el conjunto de alelos. 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 objetivo puede referirse al individuo cuyos datos genéticos están siendo determinados. En un contexto, hay disponible solamente una cantidad limitada de ADN del individuo objetivo. En un contexto, el individuo objetivo es un embrión o un feto. En algunas realizaciones, puede haber más de un individuo objetivo. En algunas realizaciones, cada niño, embrión, feto o esperma derivados de un par de progenitores puede ser considerado individuo objetivo.

Individuo relacionado puede referirse a cualquier individuo que esté relacionado genéticamente, y comparta por tanto bloques de haplotipos con el individuo objetivo. En un contexto, el individuo objetivo puede ser un progenitor genético del individuo objetivo, 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 o abuelo.

Hermano puede referirse a cualquier individuo cuyos padres sean los mismos que los del individuo en cuestión. En algunas realizaciones, puede referirse a un niño ya nacido, a un embrión, o a un feto, o una o más células procedentes de un niño ya nacido, de un embrión o de 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.

Progenitor puede referirse a la madre o el padre genéticos de un individuo. Un individuo tiene típicamente dos progenitores, una madre y un padre. Un progenitor puede ser considerado un individuo.

Contexto parental puede referirse al estado genético de un SNP determinado, en cada uno de los dos cromosomas relevantes para cada uno de los dos progenitores del objetivo.

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

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

Decisión clínica puede referirse a toda decisión de emprender o no una acción, que tenga un resultado que afecte a la salud o la supervivencia de un individuo. En el contexto de la IVF, una decisión clínica puede referirse a la decisión de implantar o no uno o más embriones. 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 a la decisión de realizar más pruebas.

Respuesta de plataforma puede referirse a la caracterización matemática de las características de entrada/ salida de una plataforma de medición genética, y puede utilizarse como medida de las diferencias de medición previsibles estadísticamente.

Método basado en informática puede referirse 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. 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™.

Técnica especializada puede referirse a un método usado para determinar un estado genético. En una realización puede referirse a un método utilizado para determinar o colaborar en la determinación del estado de ploidía de un individuo. Puede referirse a un algoritmo, un método cuantitativo, un método cualitativo, y/o una técnica basada en ordenador.

Intensidad de canal puede referirse a la potencia del fluorescente u otra señal asociada a un alelo determinado, un par de base u otro marcador genético que sea resultado de un método utilizado para medir datos genéticos. Puede referirse a un conjunto de resultados. En una realización, puede referirse al conjunto de resultados de una matriz de genotipado.

Curva de función de distribución acumulativa (CDF) puede referirse a un incremento monótono, la correcta distribución de probabilidad continua de una variable, donde la coordenada "y" de un punto en la curva se refiere a la probabilidad de que la variable tome un valor inferior o igual a la coordenada "x" del punto.

Contexto parental

El contexto parental puede referirse al estado genético de un SNP determinado, en cada uno de los dos cromosomas relevantes para cada uno de los dos progenitores del objetivo. Hay que advertir que en una realización, el contexto parental no se refiere al estado alélico del objetivo, sino al estado alélico de los padres. El contexto parental de un SNP determinado puede consistir en cuatro pares de base, dos paternos y dos maternos; pueden ser iguales o distintos entre sí. Esto viene expresado típicamente como "mim2|fif2", donde mi y m2 son el estado genético del SNP determinado en los dos cromosomas maternos, y fi y f2 son el estado genético de dicho SNP en los dos cromosomas paternos. En algunas realizaciones, el contexto parental puede venir expresado como "f^lminV'. 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 base; A o B podrían representar igualmente bien a C (citosina), G (guanina), A (adenina) o T (timina). Por ejemplo, si en un alelo determinado, el genotipo materno era T en un cromosoma, y G en el cromosoma homólogo, y el genotipo paterno en ese alelo es G en ambos cromosomas homólogos, se podría decir que el alelo

del individuo objetivo tiene el contexto parental de AB|BB. Véase que, en teoría, cualquiera de los cuatro alelos 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 base se observan en un alelo determinado. En esta divulgación, en el debate se supone que se observarán solamente dos posibles pares de base en un alelo determinado, aunque resultará obvio para un experto en la técnica cómo pueden modificarse las realizaciones divulgadas aquí para tener en cuenta los casos en los que este supuesto no es válido.

Un “contexto parental” puede referirse a un conjunto o subconjunto de SNPs objetivo que tienen el mismo contexto parental. Por ejemplo, si hubiera que medir 1000 alelos en un cromosoma determinado en un individuo objetivo, el contexto AA|BB podría referirse al conjunto de todos los alelos en el grupo de 1000 alelos donde el genotipo de la madre del objetivo era homocigoto, y el genotipo del padre del objetivo 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 por fases, y por tanto AB ≠ BA, hay dieciséis posibles contextos parentales distintos: 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 SNPs en los cromosomas sexuales, tiene uno de esos contextos parentales. El conjunto de SNPs donde el contexto parental en un progenitor es heterocigoto puede ser denominado el contexto heterocigoto.

Hipótesis

Una hipótesis puede referirse 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 refiere a un conjunto de posibles estados genéticos. En algunas realizaciones, un conjunto de hipótesis puede ser diseñado de forma que una de las hipótesis del conjunto corresponda al estado genético real de un individuo determinado.

En algunas realizaciones, un conjunto de hipótesis puede estar diseñado 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, o sección de un cromosoma, en el individuo objetivo. Puede referirse también al estado de ploidía en más de uno de los 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 individuo normal contiene uno de cada 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 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, las realizaciones solo consideran las hipótesis posibles en las que 0, 1, o 2 copias de un cromosoma determinado proceden 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). 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. Es posible utilizar otros conjuntos de hipótesis, y debería ser evidente para un experto en la técnica cómo modificar el método divulgado para tener en cuenta un número de hipótesis distinto.

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 objetivo. 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 objetivo y el individuo relacionado, es

posible inferir los datos genéticos correctos de un individuo objetivo con mayor confianza que utilizando solamente las mediciones genéticas del individuo objetivo. 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 objetivo.

Una hipótesis alélica, o una "hipótesis de estado alélico" puede referirse a una hipótesis referente a un posible estado alélico de un conjunto de alelos. En algunas realizaciones, es una clave de este método que, como se ha descrito más arriba, individuos relacionados pueden compartir bloques haplotípicos, lo que puede ayudar en la reconstrucción de datos genéticos que no hayan sido medidos perfectamente. Una hipótesis alélica puede referirse también 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. La teoría de la meiosis nos dice que cada cromosoma en un individuo viene heredado de uno de los dos progenitores, y esto es una copia casi idéntica de un cromosoma parental. Por consiguiente, si se conocen los haplotipos de los padres, es decir, el genotipo ajustado por fases de los padres, se puede inferir también el genotipo del hijo. (Aquí el término hijo incluye todo individuo formado a partir de dos gametos, uno de la madre y uno del padre). En una realización de la presente divulgación, la hipótesis alélica describe un posible estado alélico en un conjunto de alelos, incluyendo los haplotipos, así como qué cromosomas de individuos relacionados pueden coincidir con el o los cromosomas que contienen el conjunto de alelos.

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 a los expertos 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, determinadas por los dos métodos de experto, 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 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 entre el 99% y aproximadamente el 99,9%. En algunas realizaciones el umbral puede situarse por encima de aproximadamente el 99,9%.

Algunas realizaciones

En una realización de la presente invención, un método para determinar un estado de ploidía de por lo menos un cromosoma en un individuo objetivo incluye la obtención de datos genéticos del individuo objetivo, y de uno o más individuos relacionados; crear un conjunto de por lo menos una hipótesis de estado de ploidía para cada uno de los cromosomas del individuo objetivo; utilizar una o más técnicas especializadas para determinar una probabilidad estadística para cada hipótesis de estado de ploidía en el conjunto, para cada técnica especializada utilizada, considerando los datos genéticos obtenidos; combinar, para cada hipótesis de estado de ploidía, las probabilidades

estadísticas determinadas por una o más técnicas especializadas; y determinar el estado de ploidía de cada uno de los cromosomas en el individuo objetivo en base a las probabilidades estadísticas combinadas de cada una de las hipótesis de estado de ploidía.

En una realización, se puede determinar el estado de ploidía de cada uno de los cromosomas del individuo objetivo en el contexto de la fertilización in vitro, y donde el individuo objetivo es un embrión. En una realización, se puede determinar el estado de ploidía de cada uno de los cromosomas en el individuo objetivo en el contexto del diagnóstico prenatal no invasivo, y donde el individuo objetivo es un feto. Se puede determinar el estado de ploidía de cada uno de los cromosomas en el individuo objetivo en el contexto del cribado de una situación cromosómica seleccionada del grupo que incluye, entre otros, euploidía, nulisomía, monosomía, disomía uniparental, trisomía, trisomía emparejada, trisomía no emparejada, tetrasomía, otra aneuploidía, translocación no balanceada, supresiones, inserciones, mosaicismo y combinaciones de lo anterior. En una realización, se puede determinar el estado de ploidía de cada uno de los cromosomas en el individuo objetivo para diversos embriones, y se utiliza para seleccionar por lo menos un embrión para su inserción en un útero. Se toma una decisión clínica tras determinar el estado de ploidía de cada cromosoma en el individuo objetivo.

En algunas realizaciones de la presente invención, un método para la determinación del estado de ploidía de uno o más cromosomas en un individuo objetivo puede incluir los siguientes pasos:

Primero, se pueden obtener datos genéticos del individuo objetivo y de uno o más individuos relacionados. En una realización, los individuos relacionados incluyen a los dos progenitores del individuo objetivo. En una realización, los individuos relacionados incluyen los hermanos del individuo objetivo. Estos datos genéticos de individuos pueden obtenerse de distintas maneras, incluyendo, entre otras, mediciones de salida de una plataforma de genotipado; datos secuenciados medidos en el material genético del individuo; datos genéticos in silico; datos de salida de un método informático diseñado para limpiar datos genéticos o pueden ser de otras fuentes.

El material genético utilizado para mediciones puede ser amplificado por diversas técnicas conocidas por los expertos.

Los datos genéticos del individuo objetivo pueden ser medidos utilizando herramientas o técnicas tomadas de un grupo incluyendo, entre otros, Sondas de Inversión Molecular (MIP), Micromatrices de Genotipado, el TaqMan SNP Genotyping Assay, el Illumina Genotyping System, otros ensayos de genotipado, hibridación in situ fluorescente, (FISH), secuenciación, otras plataformas de genotipado de alto rendimiento y combinaciones de lo anterior. Los datos genéticos de individuo objetivo pueden ser medidos analizando sustancias tomadas de un grupo incluyendo, entre otras, una o más células diploides del individuo objetivo, una o más células haploides del individuo objetivo, uno o más blastómeros del individuo objetivo, material genético extracelular hallado en el individuo objetivo, material genético extracelular del individuo objetivo hallado en sangre materna, células del individuo objetivo halladas en sangre materna, material genético del que se sabe que procede del individuo objetivo, y combinaciones de lo anterior. Los datos genéticos del individuo relacionado pueden ser medidos analizando sustancias tomadas de un grupo incluyendo, entre otras, masa de tejido diploide del individuo relacionado, una o más células diploides del individuo relacionado, una o más células haploides tomadas del individuo relacionado, uno o más embriones creados a partir de un o unos gametos del individuo relacionado, uno o más blastómeros tomados de tal embrión, material genético extracelular hallado en el individuo relacionado, material genético del que se sabe que procede del individuo relacionado, y combinaciones de lo anterior.

Segundo, se puede crear un conjunto de por lo menos una hipótesis de estado de ploidía para cada uno de los cromosomas del individuo objetivo. Cada una de las hipótesis de estado de ploidía puede referirse a un posible estado de ploidía del cromosoma del individuo objetivo. El conjunto de hipótesis puede incluir todos los posibles estados de ploidía que se puede esperar que tenga el cromosoma del individuo objetivo.

Tercero, utilizando una o más de las técnicas especializadas discutidas en esta divulgación, se puede determinar una probabilidad estadística para cada hipótesis de estado de ploidía en el conjunto. En algunas realizaciones, la técnica especializada puede incluir un algoritmo que opere sobre los datos genéticos obtenidos, y el resultado puede ser una probabilidad estadística determinada para cada una de las hipótesis consideradas. En una realización, por lo menos en una de las técnicas especializadass se utilizan datos de determinación de alelos parentales por fases, es decir, que utiliza como entrada datos alélicos de los padres del individuo objetivo, donde han sido determinados los haplotipos de los datos alélicos. En una realización, por lo menos una de las técnicas especializadas es específica para un cromosoma sexual. El conjunto de probabilidades determinadas puede corresponder al conjunto de hipótesis. En una realización, la probabilidad estadística incluye trazar una curva de función de distribución acumulativa para uno o más contextos parentales. En una realización, la determinación de la probabilidad estadística para cada una de las hipótesis de estado de ploidía puede incluir comparar las intensidades de los datos de salida de genotipado, promediadas sobre un conjunto de alelos, a intensidades esperadas. Las matemáticas de base de las diversas técnicas especializadas se describen en otro punto de esta divulgación.

Cuarto, el conjunto de probabilidades determinadas puede ser entonces combinado. Esto puede implicar, para cada hipótesis, multiplicar las probabilidades determinadas por cada técnica, y puede implicar también normalizar las hipótesis. En algunas realizaciones, se pueden combinar las probabilidades bajo la suposición de que son independientes. El conjunto de los productos de las probabilidades para cada hipótesis es entonces el resultado como probabilidades combinadas de las hipótesis.

Por último, el estado de ploidía del individuo objetivo se determina como el estado de ploidía asociado a la hipótesis de mayor probabilidad. En algunos casos, una hipótesis tendrá una probabilidad combinada normalizada superior a 90%. Cada hipótesis es asociada a un estado de ploidía, y el estado de ploidía asociado a la hipótesis cuya probabilidad combinada normalizada es superior a 90%, o algún otro valor umbral, puede ser elegido como el estado de ploidía determinado.

En una realización de la presente divulgación, un método para la determinación de un estado alélico en un conjunto de alelos de un individuo objetivo, procedente de uno o ambos progenitores del individuo objetivo, y posiblemente de uno o más individuos relacionados, incluye la obtención de datos genéticos del individuo objetivo, y de uno o ambos progenitores, y de individuos relacionados; crear un conjunto de por lo menos una hipótesis alélica para el individuo objetivo, y para uno o ambos progenitores, y opcionalmente para el o los individuos relacionados, donde las hipótesis describen posibles estados alélicos en el conjunto de alelos; determinar una probabilidad estadística para cada hipótesis alélica en el conjunto de hipótesis considerando los datos genéticos obtenidos; y determinar el estado alélico para cada uno de los alelos en el conjunto de alelos del individuo objetivo, y para uno o ambos progenitores, y opcionalmente para el individuo o individuos relacionados, basado en las probabilidades estadísticas de cada una de las hipótesis alélicas. En una realización, el método tiene en cuenta una posibilidad de cruces de ADN que puede producirse durante la meiosis. En una realización, el método puede llevarse a cabo paralelamente o conjuntamente con un método que determine el número de copias de un segmento de cromosoma determinado presente en el o los individuos objetivo, y donde ambos métodos utilizan una misma célula, o grupo de células, procedentes de uno o más individuos objetivo como fuente de datos genéticos.

En una realización, la determinación del estado alélico puede llevarse a cabo en el contexto de la fertilización in vitro, y donde por lo menos uno de los individuos objetivo es un embrión. En una realización, la determinación del estado alélico puede llevarse a cabo donde por lo menos uno de los individuos objetivo es un embrión, y donde la determinación del estado alélico en el conjunto de alelos del o de los individuos objetivo se lleva a cabo para seleccionar por lo menos un embrión para su transferencia en el contexto de la IVF, y donde los individuos objetivo son seleccionados del grupo incluyendo, entre otros, uno o más embriones que son de los mismos progenitores, uno o más espermas del padre, y combinaciones de lo anterior. En una realización, la determinación del estado alélico puede ser realizada en el contexto del diagnóstico prenatal no invasivo, y donde por lo menos uno de los individuos objetivo es un feto. En una realización, la determinación del estado alélico del conjunto de alelos del o de los individuales objetivo puede incluir un genotipo por fases en un conjunto de alelos para esos individuos. Se puede tomar una decisión clínica tras determinar el estado alélico en el conjunto de alelos del individuo o individuos objetivo.

En algunas realizaciones de la presente divulgación, un método para determinar los datos alélicos de uno o más individuos objetivo, y uno o ambos progenitores de los padres de los individuos objetivo, en un conjunto de alelos, puede incluir los siguientes pasos:

Primero, se pueden obtener datos genéticos del individuo o individuos objetivo, de uno o ambos progenitores, y de cero o más individuos relacionados. Estos datos genéticos de individuos pueden obtenerse de distintos modos incluyendo, entre otros, mediciones de salida de una plataforma de genotipado; pueden ser datos secuenciados medidos en el material genético del individuo; pueden ser datos genéticos in silico; pueden ser datos de salida de un método informático diseñado para limpiar datos genéticos, o pueden ser de otras fuentes. En una realización, los datos genéticos obtenidos pueden incluir polimorfismos de un solo nucleótido medidos de una matriz de genotipado. En una realización, los datos genéticos obtenidos pueden incluir datos de secuencia de ADN, es decir, la secuencia genética medida representando la estructura primaria del ADN del individuo. El material genético utilizado para las mediciones puede ser amplificado mediante diversas técnicas conocidas por los expertos. En una realización, los individuos objetivo son todos hermanos. En una realización, una o más de las mediciones genéticas de los individuos objetivo se efectuaron sobre células individuales. En una realización, se pueden utilizar modelos de respuesta de plataforma para determinar una probabilidad de un genotipo real considerando las mediciones genéticas observadas, y un sesgo de medición característico de la técnica de genotipado.

Los datos genéticos del individuo objetivo pueden medirse utilizando herramientas y o técnicas tomadas de un grupo incluyendo, entre otras, Molecular Inversion Probes (MIP), Genotyping Microarrays, el TaqMan SNP Genotyping Assay, el Illumina GenotypingSystem, otras matrices de genotipado, hibridación in situ fluorescente (FISH), secuenciación, otras plataformas de genotipado de alto rendimiento y combinaciones de ellos. Los datos genéticos del individuo objetivo pueden ser medidos analizando sustancias tomadas de un grupo incluyendo, entre otras, una o más células diploides del individuo objetivo, una o más células haploides del individuo objetivo, uno o más blastómeros del individuo objetivo, material genético extracelular hallado en el individuo objetivo, material genético extracelular del individuo objetivo hallado en sangre materna, células del individuo objetivo halladas en la sangre materna, material genético del que se sabe procede del individuo objetivo y combinaciones de lo anterior. Los datos genéticos del individuo relacionado pueden ser medidos analizando sustancias tomadas de un grupo incluyendo, entre otras, masa de tejido diploide del individuo relacionado, una o más células diploides del individuo relacionado, una o más células haploides tomadas del individuo relacionado, uno o más embriones creados de un gameto o gametos del individuo relacionado, uno o más blastómeros tomados de tal embrión, material genético extracelular

hallado en el individuo relacionado, material genético del que se sabe que procede del individuo relacionado, y combinaciones de lo anterior.

Segundo, se puede crear un conjunto de diversas hipótesis alélicas para el conjunto de alelos, para cada uno de los individuos. Cada una de las hipótesis alélicas puede referirse a una posible identidad para cada uno de los alelos en el conjunto de alelos para ese individuo. En una realización, la identidad de los alelos de un individuo objetivo puede incluir el origen del alelo, es decir, el progenitor del que procede genéticamente el alelo, y el cromosoma específico del que deriva genéticamente el alelo. El conjunto de las hipótesis puede incluir todos los posibles estados alélicos que cabe esperar que tenga el individuo objetivo dentro de ese conjunto de alelos.

Por último, se puede determinar una probabilidad estadística para cada una de las hipótesis alélicas considerando los datos genéticos obtenidos. La determinación de la probabilidad de una hipótesis determinada puede establecerse utilizando cualquiera de los algoritmos descritos en esta divulgación, específicamente los de la sección de la determinación de alelos. El conjunto de hipótesis alélicas de un individuo puede incluir todos los posibles estados alélicos de dicho individuo, en el conjunto de alelos. Las hipótesis que coinciden más exactamente con los datos genéticos medidos ruidosos del individuo objetivo son las que probablemente sean las correctas. La hipótesis que se corresponde exactamente con los datos genéticos reales del individuo objetivo es la de mayor posibilidad de ser identificada como con una probabilidad muy alta. Se puede determinar que el estado alélico es el que corresponde a la hipótesis determinada como con la mayor probabilidad. En algunas realizaciones el estado alélico puede ser determinado para varios subconjuntos del conjunto de alelos.

Parental Support

Algunas realizaciones de la presente divulgación pueden utilizar el método PARENTAL SUPPORT™ (PS) basado en informática. En algunas realizaciones, el método PARENTAL SUPPORT ™ es un conjunto de métodos que pueden ser utilizados para determinar los datos genéticos, con alta precisión, de una célula o un número reducido de células, específicamente para determinar alelos vinculados a enfermedad, otros alelos de interés, y/o el estado de ploidía de la célula o células.

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 objetivo, y posiblemente de uno o más individuos relacionados, para reconstruir, *insilico,* el genotipo en diversos alelos, y/o el estado de ploidía de un embrión o de cualquier célula o células objetivo, y el ADN objetivo en la localización de loci clave con un alto grado de confianza. El método PARENTAL SUPPORT™ puede reconstruir no solamente polimorfismos de un solo nucleótido que fueran medidos deficientemente, sino también inserciones y supresiones, y SNPs o regiones completas de ADN que no fueron medidas en absoluto. Además, el método PARENTAL SUPPORT™ puede medir múltiples loci vinculados a 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 biopsiados 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 objetivo (embrión) utilizando el genotipo de individuos relacionados (padres) como referencia. El PARENTAL SUPPORT™ puede ser especialmente pertinente cuando se dispone solamente de una pequeña cantidad de material genético (ej. PGD) y donde las mediciones directas de los genotipos son inherentemente ruidosas, debido a las limitadas cantidades de material genético. 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 SNPs. 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 SNPs en base a datos obtenidos de la base de datos HapMap, desarrollada por el International HapMapProject. Con el modelo genético subyacente y el modelo de error de medición, puede utilizarse una estimación *maximum a posteriori* (MAP), con modificaciones para eficiencia informática, para calcular los valores de alelos ordenados correctos en cada SNP en el embrión.

Un aspecto de la tecnología PARENTAL SUPPORT™ es un algoritmo de determinación del número de copias cromosómicas que en algunas realizaciones utilizan contextos de genotipo parental. Para determinar el número de copias, el algoritmo puede utilizar el fenómeno de pérdida de locus (LDO) combinado con distribuciones de genotipos embriónicos esperados. Durante la amplificación de genoma completo se produce necesariamente LDO. La tasa de LDO concuerda con el número de copias del material genético del que deriva, es decir, menos copias cromosómicas producen más LDO, y viceversa. Como tal, se deduce que loci con determinados contextos de genotipos parentales se comportan de un modo característico en el embrión, relacionado con la probabilidad de contribuciones alélicas al embrión. Por ejemplo, si ambos progenitores tienen estados BB homocigotos, el embrión no debería tener nunca estados AB o AA. En este caso, se espera que las mediciones en el canal de detección A tengan una distribución determinada por el ruido de fondo y varias señales de interferencia, pero no genotipos

válidos. A la inversa, si ambos progenitores tienen estados AA homocigotos, el embrión no debería tener nunca estados AB o BB, y se espera que las mediciones del canal A tengan la máxima intensidad posible considerando la tasa de LDO en una amplificación de genoma completo determinada. Cuando el estado de número de copias subyacente del embrión es distinto de disomía, los loci correspondientes a los contextos parentales específicos se comportan de forma predecible, basado en el contenido alélico adicional aportado por uno de los progenitores o carente en él. Esto permite la determinación del estado de ploidía en cada cromosoma, o segmento de cromosoma. Los detalles de una realización de este método se describen en otra parte de la presente divulgación.

Determinación del número de copias utilizando contextos parentales

El concepto de los contextos parentales puede resultar útil en el contexto de la determinación del número de copias (denominado también “determinación de ploidía”). Cuando están genotipados, cabe esperar que todos los SNPs en un primer contexto parental se comporten estadísticamente del mismo modo al ser medidos respecto a un estado de ploidía determinado. Por el contrario, se puede esperar que algunos conjuntos de SNPs de un segundo contexto parental se comporten de forma distinta a los del primer contexto parental en determinadas circunstancias, como en ciertos estados de ploidía, y la diferencia en comportamiento puede ser característica para un estado o conjunto de estados de ploidía determinados. Existen diversas técnicas estadísticas que podrían ser utilizadas para analizar las respuestas medidas en los diversos loci dentro de los distintos contextos parentales. En algunas realizaciones de la presente divulgación, se pueden utilizar técnicas estadísticas para obtener probabilidades para cada una de las hipótesis. En algunas realizaciones de la presente divulgación, se pueden utilizar técnicas estadísticas para obtener probabilidades para cada una de las hipótesis junto con confianzas en las probabilidades calculadas. Algunas técnicas, si se utilizan individualmente, pueden no ser adecuadas para determinar el estado de ploidía de un cromosoma concreto con un nivel de confianza determinado.

La clave de un aspecto de la presente divulgación es el hecho de que algunas técnicas especializadas especializadas resultan especialmente buenas para confirmar o eliminar de la discusión determinados estados de ploidía o conjuntos de estados de ploidía, pero pueden no ser adecuadas para determinar correctamente el estado de ploidía si se utilizan solas. Esto contrasta con algunas técnicas especializadas que pueden ser relativamente buenas diferenciando entre sí la mayoría o todos los estados de ploidía, pero no con tanta confianza como pueden tener algunas técnicas especializadas especializadas diferenciando un subconjunto concreto de estados de ploidía. Algunos métodos utilizan una técnica generalizada para determinar el estado de ploidía. No obstante, la combinación del conjunto apropiado de técnicas especializadas especializadas puede resultar más precisa en las determinaciones de ploidía, que la utilización de una técnica especializada generalizada. Por ejemplo, una técnica especializada puede ser capaz de determinar si un objetivo es monosómico con un muy alto nivel de confianza, una segunda técnica especializada puede ser capaz de determinar si un objetivo es o no trisómico o tetrasómico con un muy elevado nivel de confianza, y una tercera técnica puede ser capaz de detectar disomía uniparental con un muy alto grado de confianza.

Ninguna de estas técnicas puede ser capaz de efectuar una determinación precisa de ploidía por sí sola, Pero cuando se emplean combinadas estas tres técnicas especializadas especializadas, pueden ser capaces de proceder a la determinación de ploidía con mayor precisión que cuando se usa una técnica especializada que puede diferenciar razonablemente bien todos los estados de ploidía. En algunas realizaciones de la presente divulgación, se pueden combinar los resultados de probabilidades de múltiples técnicas para llegar a una determinación del estado de ploidía con alto nivel de confianza. En algunas realizaciones de la presente divulgación, las probabilidades que predice cada una de las técnicas para una hipótesis determinada pueden ser multiplicadas juntas, y ese producto considerarse la probabilidad combinada para esa hipótesis. El estado o estados de ploidía asociados a la hipótesis que tiene la mayor probabilidad combinada puede ser considerado el estado de ploidía correcto. Si se elige adecuadamente el conjunto de técnicas especializadas, el producto combinado de las probabilidades puede permitir una determinación más exacta del estado de ploidía que con una sola técnica. En algunas realizaciones de la invención, las probabilidades de las hipótesis con más de una técnica pueden multiplicarse, utilizando por ejemplo álgebra lineal, y renormalizando, para obtener las probabilidades combinadas. En una realización, las confianzas de las probabilidades pueden combinarse de forma similar a las probabilidades. En una realización de la presente divulgación, las probabilidades de las hipótesis pueden combinarse bajo el supuesto de que son independientes. En algunas realizaciones de la presente divulgación, el resultado de una o más técnicas puede utilizarse como entrada para otras técnicas. En una realización de la presente divulgación, la determinación de ploidía, realizada utilizando una o un conjunto de técnicas especializadas, puede utilizarse para determinar la entrada apropiada para la técnica de determinación de alelos. En una realización de la presente divulgación, el resultado de datos genéticos limpios por fase de la técnica de determinación de alelos puede ser utilizado como entrada para una o un conjunto de técnicas especializadas de determinación de ploidía. En algunas realizaciones de la presente divulgación, puede reiterarse el uso de varias técnicas.

En algunas realizaciones de la presente divulgación, el estado de ploidía puede ser determinado con un nivel de confianza mayor de aproximadamente 80%. En algunas realizaciones de la presente divulgación, el estado de ploidía puede ser determinado con un nivel de confianza superior a aproximadamente 90%.

En algunas realizaciones de la presente divulgación, el estado de ploidía puede determinarse con un nivel de confianza superior a aproximadamente 95%. En algunas realizaciones de la presente divulgación, el estado de ploidía puede ser determinado con un nivel de confianza superior a aproximadamente 99%. En algunas

realizaciones de la presente divulgación, el estado de ploidía puede ser determinado con un nivel de confianza superior a aproximadamente 99,9%. En algunas realizaciones de la presente divulgación, un alelo o un conjunto de alelos pueden determinarse con un nivel de confianza superior a aproximadamente 80%. En algunas realizaciones de la presente divulgación, el o los alelos pueden ser determinados con un nivel de confianza superior a aproximadamente 90%. En algunas realizaciones de la presente divulgación, el o los alelos pueden ser determinados con un nivel de confianza superior a aproximadamente 95%. En algunas realizaciones de la presente divulgación, el o los alelos pueden ser determinados con un nivel de confianza superior a aproximadamente 99%. En algunas realizaciones de la presente divulgación, el o los alelos pueden ser determinados con un nivel de confianza superior a aproximadamente 99,9%. En algunas realizaciones de la presente divulgación, los datos resultado de la determinación de alelos son ajustados por fases, diferenciando los datos genéticos de los dos cromosomas homólogos. En algunas realizaciones de la presente divulgación, los datos de determinación de alelos por fases son resultados para todos los individuos.

Más abajo se da una descripción de varias técnicas estadísticas que se pueden utilizar en la determinación del estado de ploidía. No se pretende que esta lista sea exhaustiva respecto a las posibles técnicas especializadas. Es posible utilizar cualquier técnica estadística que sea capaz de atribuir probabilidades y/o confianzas en el conjunto de las hipótesis del estado de ploidía de un objetivo. Se pueden combinar cualquiera de las siguientes técnicas, o pueden combinarse éstas con otras técnicas no comentadas en esta divulgación.

Técnica de permutación

La tasa de LDO concuerda con el número de copias del material genético del que deriva, es decir, menos copias de cromosomas dan como resultado mayor LDO y viceversa. La consecuencia es que loci con determinados contextos de genotipos parentales se comportan de forma característica en el embrión, relacionado con la probabilidad de contribuciones alélicas al embrión. En una realización de la presente invención, llamada la “técnica de permutación”, es posible utilizar el comportamiento característico de loci en los diversos contextos parentales, para deducir el estado de ploidía de esos loci. Específicamente, esta técnica implica comparar la relación existente entre distribuciones observadas de datos de medición de alelos para distintos contextos parentales, y determinar qué estado de ploidía coincidía con el conjunto observado de relaciones entre las distribuciones. Esta técnica resulta especialmente útil para determinar el número de cromosomas homólogos presentes en la muestra. Proyectando gráficamente una curva de función de distribución acumulativa (CDF) para cada uno de los contextos parentales, se puede observar que varios contextos se agrupan juntos. Obsérvese que una curva CDF es solamente una forma de visualizar y comparar las distribuciones observadas de los datos de medición de alelos. Por ejemplo, la Figura 1 muestra una curva CDF de un cromosoma disómico. En particular, la Figura 1 muestra cómo los datos de medición de alelos de determinados contextos de genotipos parentales (Madre(Padre) se comportan de una forma característica en el embrión, relacionada con la probabilidad de contribuciones alélicas al embrión. Los nueve contextos parentales se agrupan en cinco bloques cuando el cromosoma en cuestión es disómico. En el trazado de la curva CDF, la variable independiente, a lo largo del eje x, es la respuesta del canal, y la variable dependiente, a lo largo del eje y, es el porcentaje de alelos en ese contexto cuya respuesta de canal es inferior al valor umbral.

Por ejemplo, si ambos progenitores tienen estados BB homocigotos, el embrión no debería tener nunca estados AB o AA. En este caso, las mediciones en el canal de detección A tendrán probablemente una distribución determinada por ruido de fondo y varias señales de interferencia, pero no genotipos válidos. Inversamente, si ambos progenitores tienen estados AA homocigotos, el embrión no debería tener nunca estados AB o BB, y las mediciones en el canal A tendrían probablemente la mayor intensidad posible dada la tasa de LDO en una amplificación de genoma completo determinada. Cuando el estado del número de copias subyacente del embrión es distinto de disomía, los loci correspondientes a los contextos parentales específicos se comportan de una forma predecible, basada en el contenido alélico adicional aportado o ausente de uno de los progenitores. Los trazados de función de densidad acumulativa de intensidad de sonda de microarray en un canal de detección, separados por contexto de genotipo parental, ilustran el concepto (ver Figura 2). Las Figuras 2A-2D muestran específicamente cómo la relación entre las curvas de contexto en un gráfico CDF cambian de forma previsible con un cambio en el número de copias cromosómicas. La Figura 2A muestra una curva de función de distribución acumulativa de un cromosoma disómico, la Figura 2B muestra una curva de función de distribución acumulativa de un cromosoma nulisómico, la Figura 2C muestra una curva de función de distribución acumulativa de un cromosoma monosómico, y la Figura 2D muestra una curva de función de distribución acumulativa de un cromosoma trisómico materno.

Cada contexto está representado como M1M2IF1F2, donde Mi y M2 son los alelos maternos, y F_1 y F_2 son los alelos paternos. Hay nueve posibles contextos parentales (ver la leyenda de las Figuras 2A-2D), donde en un cromosoma disómico, forman cinco bloques en el gráfico CDF. En las nulisomías, todas las curvas de contexto parental se agrupan con fondo en el gráfico CDF. En el caso de la monosomía, se puede esperar ver solamente tres grupos de curvas de contexto, porque la eliminación de un contexto parental da solo tres posibles resultados embriónicos: homocigoto AA, heterocigoto AB, y homocigoto BB. Cabría esperar que la trisomía tuviera también una distinta topología de curva CDF, de forma que haya siete grupos, causados por alelos extra en un solo canal de detección y de un solo progenitor.

Un conjunto de topologías canónicas esperadas viene ilustrado en las Figuras 2A-2D, para las cuales el estado de ploidía puede ser determinado por inspección visual de los gráficos. En algunos casos, puede no resultar tan fácil interpretar los datos de una muestra como los datos mostrados en las Figuras 2A-2D.

Muchos factores pueden influir en la claridad de los datos, incluyendo: ADN degradado de blastómeros que provoca señales con un cociente señal-a-ruido muy bajo; errores de ploidía parciales que se encuentran frecuentemente durante la IVF, como translocaciones; y sesgos de amplificación específicos de cromosoma y de segmento de cromosoma, causados posiblemente por las posiciones físicas de los cromosomas en el núcleo o fenómeno epigenético, como distintos niveles de metilación y estructuras proteicas en torno a los cromosomas. Esos y otros diversos fenómenos pueden afectar diferencialmente a cada cromosoma de un par homólogo, en cuyo caso resulta difícil distinguirlos de estados de ploidía. En una realización de la presente divulgación, para acomodar esos varios efectos, puede utilizarse un algoritmo estadístico para analizar datos como los que se ilustran en las Figuras 2A-2D y generar una determinación de ploidía junto con una confianza en la exactitud de esa determinación.

En una realización de la presente divulgación, para reforzar las diferencias que puedan existir entre una muestra y otra, o entre muestras de líneas celulares y blastómeros, el algoritmo puede ser *no paramétrico* y no depende de los valores esperados de estadísticas o umbrales probados en determinadas muestras y aplicados a otras. En una realización de la presente divulgación, el algoritmo utiliza estadísticas de rango cuantil (un método de permutación no paramétrico), que calcula primero el rango de la curva CDF de cada contexto a una intensidad en la que el contexto de fondo está dentro de aproximadamente el 80% de una densidad de aproximadamente 1. En otra realización el algoritmo puede calcular el rango de la curva CF de cada contexto a una intensidad en la que el contexto de fondo está dentro de aproximadamente el 90% de una densidad de aproximadamente 1. En otra realización, El algoritmo puede calcular el rango de la curva CDF de cada contexto a una intensidad a la que el contexto de fondo se halla dentro de aproximadamente el 95% de una densidad de aproximadamente 1. Entonces el algoritmo compara el rango de los datos con el rango esperado para varios estados de ploidía. Por ejemplo, si el contexto AB|BB tiene el mismo rango que el contexto BB|AA, esto difiere de lo esperado en disomía, pero coincide con la trisomía materna. Entonces se puede examinar la distribución de los datos de cada muestra para determinar la probabilidad de que dos curvas CDF puedan haber intercambiado rangos por azar, y utilizar entonces información, combinada con las estadísticas de rango, para proceder a determinaciones del número de copias y calcular confianzas explícitas. El resultado de esta técnica estadística es un diagnóstico de alta precisión de número de copias cromosómicas, combinado con una confianza explícita en cada determinación.

Dado que la determinación del número de copias de técnica de permutación para un cromosoma específico es independiente de todos los otros cromosomas, sin perder generalidad es posible el enfoque en un solo cromosoma específico. Para un genotipo materno gM y un genotipo paterno gT determinados se puede utilizar gM|gF para denotar el contexto parental; ej. AB|BB se refiere a los SNPs donde el genotipo de la madre es AB y el genotipo del padre es BB.

Para un contexto gM|gF determinado, $X_{gM|gF}$ indica el conjunto de respuestas de canal x para todos los SNPs en el contexto gM|gF. De forma similar, se puede utilizar $Y_{gM|gF}$ para indicar el conjunto de respuestas del canal y. Además, para un número C positivo determinado se puede definir $I_{\{x\leq c\}}$

$$n^{x}_{gM|gF}(c) = \sum_{x \in X_{gM|gF}} I_{\{x\leq c\}} \quad \text{y} \quad n^{y}_{gM|gF}(c) = \sum_{y \in Y_{gM|gF}} I_{\{y<c\}}$$

Se puede utilizar también $N_{gM|gF}$ para indicar el número de SNPs en el contexto gM|gF. Es posible definir

$$\hat{p}^{x}_{gM|gF}(c) = (n^{x}_{gM|gF}(c)) / (N_{gM|gF}) \quad \text{y} \quad \hat{p}^{y}_{gM|gF}(c) = (n^{y}_{gM|gF}(c)) / (N_{gM|gF})$$

Se puede pensar en $\hat{p}^{x}_{gM|gF}(c), \hat{p}^{y}_{gM|gF}(c)$ como el valor de CF empírico del canal x- , canal y, respuesta de contexto gM|gF en el punto c. Se puede indicar las verdaderas CDFs como $p^{x}_{gM|gF}(c)$, y $p^{y}_{gM|gF}(c)$

El algoritmo

La idea básica tras el algoritmo es que para un entero positivo determinado c, el orden

$$p^{x}_{AA|AA}(c), p^{x}_{AB|AA}(c), p^{x}_{BB|AA}(c), p^{x}_{AA|AB}(c), p^{x}_{AB|AB}(c), p^{x}_{BB|AB}(c), p^{x}_{AA|BB}(c), p^{x}_{AB|BB}(c), \text{ y } p^{x}_{BB|BB}(c),$$

puede variar en base al número de copias de cromosoma. Lo mismo es aplicable al canal y. En una realización de la presente divulgación, puede utilizarse ese orden para determinar el número de copias cromosómicas. Dado que el canal x y el canal y se tratan independientemente, en adelante está discusión se centrará solo en el canal x.

Cálculos

El primer paso consiste en atribuir un valor a c que maximice la posibilidad de distinguir entre los contextos; es decir, el valor de *c* que maximiza la diferencia entre los dos contextos extremos, AA|AA y BB|BB. Más exactamente, se puede definir:

$c_x = \underset{c\in\{0,100,\ldots,66000\}}{\operatorname{argmax}} \hat{p}^{x}_{BB|BB}(c) - \hat{p}^{x}_{AA|AA}(c)$ y $e_x = \hat{p}^{x}_{BB|BB}(c_x) - \hat{p}^{x}_{AA|AA}(c_x)$, y también

$c_y = \underset{c\in\{0,100,\ldots,66000\}}{\operatorname{argmax}} \hat{p}^{y}_{BB|BB}(c) - \hat{p}^{y}_{AA|AA}(c)$ y $e_x = \hat{p}^{y}_{BB|BB}(c_y) - \hat{p}^{y}_{AA|AA}(c_y)$.

Por consiguiente, en este debate se tomará c_x como el punto muestra y todas las otras comparaciones se harán respecto a $\hat{p}^{x}_{AA|AA}(c_x)$, $\hat{p}^{x}_{AB|AA}(c_x)$, $\hat{p}^{x}_{BB|AA}(c_x)$, $\hat{p}^{x}_{AA|AB}(c_x)$, $\hat{p}^{x}_{AB|AB}(c_x)$, $\hat{p}^{x}_{BB|AB}(c_x)$, $\hat{p}^{x}_{AA|BB}(c_x)$, $\hat{p}^{x}_{AB|BB}(c_x)$, $\hat{p}^{x}_{BB|BB}(c_x)$. En adelante la discusión abandonará la dependencia de c_x. Para asignar una confianza a la determinación del número de copias cromosómicas, es importante determinar una variante para cada $p^{x}_{gM|gF}$. Esto puede hacerse utilizando un modelo binomial. En particular, se puede observar que cada $N^{x}_{gM|gF}$ es la suma de variables aleatorias I.I.D. Bernoulli, y por tanto la suma normalizada tiene desviación estándar.

$$\sigma^{x}_{gM|gF} = \sqrt{\frac{p^{x}_{gM|gF}\,(1-p^{x}_{gM|gF})}{N_{gM|gF}}}$$

Cálculo de la confianza

Se describe aquí un método para calcular la confianza de una hipótesis determinada de número de copias. Cada hipótesis tiene un conjunto de permutaciones válidas de

$\hat{p}^{x}_{AA|AA} \approx p^{x}_{AA|AA}$ $\hat{p}^{x}_{AA|AB} \approx p^{x}_{AA|AB}$ $\hat{p}^{x}_{BB|AB} \approx p^{x}_{BB|AB}$

$\hat{p}^{x}_{AB|AA} \approx p^{x}_{AB|AA}$ $\hat{p}^{x}_{AB|AB} \approx p^{x}_{AB|AB}$ $\hat{p}^{x}_{AA|BB} \approx p^{x}_{AA|BB}$

$\hat{p}^{x}_{BB|AA} \approx p^{x}_{AB|AA}$ $\hat{p}^{x}_{BB|AB} \approx p^{x}_{BB|AB}$ $\hat{p}^{x}_{AB|BB} \approx p^{x}_{AB|BB}$

Por ejemplo, una hipótesis de disomía puede tener el siguiente conjunto de permutaciones válidas:

$\hat{p}^{x}_{AA|AA} \approx p^{x}_{AA|AA} : 1$ $\hat{p}^{x}_{AA|AB} \approx p^{x}_{AA|AB} : 2$ $\hat{p}^{x}_{AA|BB} \approx p^{x}_{AA|BB} : 3$

$\hat{p}^{x}_{AB|AA} \approx p^{x}_{AB|AA} : 2$ $\hat{p}^{x}_{AB|AB} \approx p^{x}_{AB|AB} : 3$ $\hat{p}^{x}_{AB|BB} \approx p^{x}_{AB|BB} : 4$

$\hat{p}^{x}_{BB|AA} \approx p^{x}_{AB|AA} : 3$ $\hat{p}^{x}_{BB|AB} \approx p^{x}_{BB|AB} : 4$ $\hat{p}^{x}_{BB|BB} \approx p^{x}_{BB|BB} : 5$

donde se da el mismo valor a dos entradas si su orden relativo no se especifica en la hipótesis. Por consiguiente, hay 12 permutaciones válidas para disomía. La confianza para una hipótesis determinada se calcula hallando la permutación válida que coincide con los datos observados. Esto se hace ordenando los elementos de los grupos invariables, grupos con los mismos números de orden, respecto a su estadística observada.

Por ejemplo, dado que se observa el siguiente orden:

$$\begin{bmatrix} \hat{p}^{x}_{AA|AA} \\ \hat{p}^{x}_{AB|AA} \\ \hat{p}^{x}_{BB|AA} \\ \hat{p}^{x}_{AA|AB} \\ \hat{p}^{x}_{AB|AB} \\ \hat{p}^{x}_{BB|AB} \\ \hat{n}^{x}_{AA|BB} \\ \hat{p}^{x}_{AB|BB} \\ \hat{p}^{x}_{BB|BB} \end{bmatrix}$$

La permutación que es consistente con disomía y coincide con los datos es

$$\begin{bmatrix} p^{x}_{AA|AA} \\ p^{x}_{AB|AA} \\ p^{x}_{BB|AA} \\ p^{x}_{AA|AB} \\ p^{x}_{AB|AB} \\ p^{x}_{BB|AB} \\ p^{x}_{AA|BB} \\ p^{x}_{AB|BB} \\ p^{x}_{BB|BB} \end{bmatrix}$$

Se puede entonces calcular la probabilidad de los datos del canal x observados considerando una hipótesis de disomía como Pr{datos x | $H_{1,1}$} = Pr{datos x | mejor orden de concordancia}

$$\overset{(a)}{\approx} \Pr\{\hat{p}^{x}_{AA|AA}, \hat{p}^{x}_{AB|AA} \mid p^{x}_{AA|AA} \leq p^{x}_{AB|AA}\}$$

$$\cdot \Pr\{\hat{p}^{x}_{AB|AA}, \hat{p}^{x}_{AA|AB} \mid p^{x}_{AB|AA} \leq p^{x}_{AA|AB}\}$$

$$\cdot \Pr\{\hat{p}^{x}_{AA|AB}, \hat{p}^{x}_{BB|AA} \mid p^{x}_{AA|AB} \leq p^{x}_{BB|AA}\}$$

$$\cdot \Pr\{\hat{p}^{x}_{BB|AA}, \hat{p}^{x}_{AA|BB} \mid p^{x}_{BB|AA} \leq p^{x}_{AA|BB}\}$$

$$\cdot \Pr\{\hat{p}^{x}_{AA|BB}, \hat{p}^{x}_{AB|AB} \mid p^{x}_{AA|BB} \leq p^{x}_{AB|AB}\}$$

$$\cdot \Pr\{\hat{p}^{x}_{AB|AB}, \hat{p}^{x}_{BB\ AB} \mid p^{x}_{AB|AB} \leq p^{x}_{BB|AB}\}$$

$$\cdot \Pr\{\hat{p}^{x}_{BB|AB}, \hat{p}^{x}_{AB\ BB} \mid p^{x}_{BB|AB} \leq p^{x}_{AB|BB}\}$$

$$\cdot \Pr\{\hat{p}^{x}_{AB|BB}, \hat{p}^{x}_{BB\ BB} \mid p^{x}_{AB|BB} \leq p^{x}_{BB|BB}\}$$

En este caso, se hace la aproximación (a) para hacer computable la probabilidad. Finalmente, para dos contextos cualquiera gM1|gF1 y gM2|gF se puede calcular:

$$\Pr\{\hat{p}^x_{gM1|gF1}, \hat{p}^x_{gM2|gF2} \mid p^x_{gM1|gF1} \le p^x_{gM2|gF2}\}$$

$$= \frac{1}{\Pr\{p^x_{gM1|gF1} \le p^x_{gM2|gF2}\}} \Pr\{\hat{p}^x_{gM1|gF1}, \hat{p}^x_{gM2|gF2}, p^x_{gM1|gF1} \le p^x_{gM2|gF2}\}$$

$$\overset{(a)}{=} \frac{1}{\Pr\{p^x_{gM1|gF1} \le p^x_{gM2|gF2}\}} \cdot \int_{p^x_{gM1|gF1} \le p^x_{gM2|gF2}} Pr\{\hat{p}^x_{gM1|gF1}, \quad \hat{p}^x_{gM2|gF2}, \quad p^x_{gM1|gF1}, p^x_{gM2|gF2}\}\, dp^x_{gM1|gF1}\, dp^x_{gM2|gF2}$$

$$\overset{(b)}{=} \alpha \int_{p^x_{gM1|gF1} \le p^x_{gM2|gF2}} Pr\{\hat{p}^x_{gM1|gF1}, \hat{p}^x_{gM2|gF2} \mid p^x_{gM1|gF1}, p^x_{gM2|gF2}\}\, dp^x_{gM1|gF1}\, dp^x_{gM2|gF2}$$

$$\overset{(c)}{=} \alpha \int_{p^x_{gM1|gF1} \le p^x_{gM2|gF2}} f_{p^x_{gM1|gF1}, \sigma^x_{gM1|gF1}}(\hat{p}^x_{gM1|gF1})\, f_{p^x_{gM2|gF2}, \sigma^x_{gM2|gF2}}(\hat{p}^x_{gM2|gF2})\, dp^x_{gM1|gF1}\, dp^x_{gM2|gF2}$$

$$= \alpha \int_{p^x_{gM1|gF1} \le p^x_{gM2|gF2}} f_{\hat{p}^x_{gM1|gF1}, \sigma^x_{gM1|gF1}}(p^x_{gM1|gF1})\, f_{\hat{p}^x_{gM2|gF2}, \sigma^x_{gM2|gF2}}(p^x_{gM2|gF2})\, dp^x_{gM1|gF1}\, dp^x_{gM2|gF2}$$

donde

(a) y (b) derivan de la independencia, y una suposición de una distribución uniforme en el $p^x_{gM|gF}$ y (c) siguen del uso de $f_{\mu,\sigma}$ para indicar el PDF normal con media μ y desviación estándar a y una aplicación del CLT. Finalmente, de (1) es posible derivar:

$\Pr\{\hat{p}^x_{gM1|gF1}, \hat{p}^x_{gM2|gF2} \mid p^x_{gM1|gF1} \le p^x_{gM2|gF2}\} = \Pr\{W_1 \le W_2\}$, donde

$W_1 \sim N(p^x_{gM1|gF1}, \sigma^x_{gM1|gF1})$ y

$W_2 \sim N(p^x_{gF2|gF2}, \sigma^x_{gM2|gF2})$

Las confianzas del canal x y del canal y se combinan bajo el supuesto de independencia; es decir

Pr{datos | H1,1} = Pr{datos x | H1,1} Pr{datos y | H1,1}.

De esta forma es posible calcular la probabilidad de los datos considerando cada hipótesis. En una realización, puede aplicarse la regla de Bayes para hallar la probabilidad de cada hipótesis considerando los datos.

Nulisomía

En una realización de la presente divulgación, cuando se usa la técnica de permutación, las nulisomías se tratan de una forma especial. Además de asignar una confianza atribuida a la determinación del número de copias, también es posible realizar un test de envoltura. Si la envoltura e_x o e_y es inferior a un umbral, la probabilidad de nulisomía se sitúa en aproximadamente 1 y la probabilidad de las otras hipótesis es de aproximadamente 0. En una realización de la presente divulgación, este umbral puede fijarse en aproximadamente 0,05. En una realización de la presente divulgación, este umbral puede establecerse en aproximadamente 0,1. En una realización de la presente divulgación, este umbral puede establecerse en aproximadamente 0,2. El conjunto de permutaciones de nulisomía para el canal x es como sigue:

- $p^x_{AA|AA} \geq p^x_{BB|BB}$
- $p^x_{AB|AA} \geq p^x_{BB|BB}$
- $p^x_{AA|AB} \geq p^x_{BB|BB}$
- $p^x_{AA|AA} \geq p^x_{BB|AB}$
- $p^x_{AB|AA} \geq p^x_{BB|AB}$
- $p^x_{AA|AB} \geq p^x_{BB|AB}$
- $p^x_{AA|AA} \geq p^x_{AB|BB}$
- $p^x_{AB|AA} \geq p^x_{BB|AB}$
- $p^x_{AA|AB} \geq p^x_{AB|BB}$

donde el orden de todos los contextos no listados se elige para maximizar la probabilidad. De forma similar, el conjunto de permutaciones de nulisomía para el canal y es como sigue:

- $p^y_{BB|BB} \geq p^y_{AA|AA}$
- $p^y_{AB|BB} \geq p^y_{AA|AA}$
- $p^y_{BB|AB} \geq p^y_{AA|AA}$
- $p^y_{BB|BB} \geq p^y_{AA|AB}$
- $p^y_{AB|BB} \geq p^y_{AA|AB}$
- $p^y_{BB|AB} \geq p^y_{AA|AB}$
- $p^y_{BB|BB} \geq p^y_{AB|AA}$
- $p^y_{AB|BB} \geq p^y_{AB|AA}$
- $p^y_{BB|AB} \geq p^y_{AB|AA}$

Segmentación

El algoritmo de permutación estándar descrito más arriba funciona bien en la mayoría de los casos, y proporciona confianzas teóricas que se corresponden a índices de error empíricos. El problema que ha surgido es la conducta específica regional en un pequeño subconjunto de los datos cromosómicos. Esta conducta puede ser debida a que las proteínas bloqueen algunas secciones de los cromosomas, o a una translocación. Para solventar tales problemas regionales, es posible utilizar una interfaz de protocolo segmentada al método de permutación.

Si a un cromosoma se le atribuye una confianza inferior a un umbral, el cromosoma se descompone en varias regiones, y el algoritmo de segmentación se ejecuta sobre cada segmento. En una realización de la presente divulgación, se pueden utilizar aproximadamente cinco segmentos iguales. En una realización de la presente divulgación, se pueden utilizar aproximadamente de dos a cinco segmentos. En una realización se pueden utilizar aproximadamente de seis a diez segmentos. En una realización de la presente divulgación se pueden utilizar más de aproximadamente diez segmentos. En una realización de la presente divulgación este umbral puede ser fijado en aproximadamente 0,6. En una realización de la presente divulgación, este umbral puede ser fijado en aproximadamente 0,8. En una realización de la presente divulgación, este umbral puede ser fijado en aproximadamente 0,9.

Luego nos podemos centrar en los segmentos con confianzas asignadas superiores a un umbral, e intentar hallar un voto mayoritario entre esos segmentos de alto nivel de confianza. En una realización de la presente divulgación, este umbral puede ser fijado en aproximadamente 0,5. En una realización de la presente divulgación, este umbral puede ser fijado en aproximadamente 0,7. En una realización de la presente divulgación, este umbral puede fijarse en aproximadamente 0,8. Por ejemplo, en el caso de que se empleen cinco segmentos iguales, si no hay una mayoría de tres o más, la técnica puede dar las confianzas de los algoritmos de permutación estándar, mientras que si existe una mayoría de tres o más segmentos de alto nivel de confianza, estos segmentos pueden ser combinados juntos, y el algoritmo de permutación estándar se aplica a los datos combinados. La técnica puede entonces proporcionar las confianzas de los datos combinados como la confianza del cromosoma completo.

En una realización de la presente divulgación, si uno de los segmentos minoritarios tiene una confianza superior a un umbral, ese cromosoma puede ser marcado como segmentado. En una realización de la presente divulgación, este umbral puede ser fijado en aproximadamente 0,8. En una realización de la presente divulgación, este umbral puede fijarse en aproximadamente 0,9. En una realización de la presente divulgación, este umbral puede establecerse en aproximadamente 0,95

Media del cromosoma completo

En algunos casos, distintos cromosomas pueden tener diferentes perfiles de amplificación. En una realización de la presente divulgación, es posible utilizar la siguiente técnica, denominada técnica de la “media del cromosoma completo”, para aumentar la precisión de los datos corrigiendo, o corrigiendo parcialmente, este sesgo de amplificación. Esta técnica sirve también para corregir, o corregir parcialmente cualquier medición u otros sesgos que puedan existir en los datos. Esta técnica no se basa en la identidad de cualquiera de los alelos medida por diversas técnicas de genotipado, sino que se basa solamente en la intensidad global de las mediciones de genotipado. Típicamente, los datos de salida en bruto de una técnica de genotipado, como una matriz de genotipado, es un conjunto de intensidades medidas de los canales que corresponden a cada uno de los cuatro pares de base, A, C, G y T. Estas intensidades medidas, tomadas de los resultados de los canales, están diseñadas para correlacionarse con la cantidad de material genético presente, así el par de base cuya intensidad medida es la mayor es considerado con frecuencia como el alelo correcto. En algunas realizaciones, las intensidades medidas para determinados conjuntos de SNPs se promedian, y la conducta característica de esas medias se utiliza para determinar el estado de ploidía del cromosoma.

El primer paso consiste en normalizar cada objetivo en cuanto a variación en la amplificación. Esto se hace utilizando un método alternado para realizar una determinación inicial de estado de ploidía. Luego se seleccionan todos los cromosomas con una determinación de ploidía con una confianza superior a un umbral establecido. En una realización de la presente divulgación, este umbral se fija en aproximadamente 99%. En una realización de la presente divulgación, este umbral se fija en aproximadamente 95%. En una realización de la presente divulgación, este umbral se fija en aproximadamente 90%. Después, las medias ajustadas de los cromosomas seleccionados se utilizan como una medida de la amplificación general del objetivo. En una realización de la presente divulgación, solo se utiliza la intensidad de la sonda fluorescente, promediada sobre el cromosoma completo. En una realización, se utilizan las intensidades de los datos resultado del genotipado sobre un conjunto de alelos.

A continuación, se ajustan las medias respecto a la determinación del número de copias del cromosoma, normalizando respecto a disomía; es decir, las monosomías se escalan por disomías por 1 y trisomías por 2/3. Las medias de cada cromosoma del objetivo se dividen entonces por la media de esas medias ajustadas de alta confianza. Esas medias normalizadas pueden ser denominadas medias ajustadas de amplificación. En una realización, solo se utilizan los alelos resultado del canal de determinados contextos. En una realización, solo se usan los alelos de AA|AA o BB|BB.

Una vez normalizados los objetivos respecto a variaciones de amplificación, se puede normalizar cada cromosoma respecto a varianza de amplificación específica de cromosoma. Para el cromosoma k^{th} se hallan todos los objetivos que tienen disomía determinada de cromosoma k con una confianza superior a la confianza umbral. Se toma la media de sus medias ajustadas de amplificación. Esto servirá como la amplificación media del cromosoma k, que puede ser denominada b{k}. Sin pérdida de generalidad, se ajusta b {1} a 1 dividiendo todas las otras b{k} por b {1}.

Las medias normalizadas de amplificación pueden ser normalizadas respecto a variación cromosómica dividiendo por el vector [b{1},...,b{24}]. Estas medias son denominadas medias estandarizadas. A partir de un conjunto de prueba compuesto de datos históricos, puede ser posible hallar medias y desviaciones estándar para esas medias estandarizadas, bajo el supuesto de monosomía, disomía y trisomía. Estas medias estandarizadas, bajo los diversos supuestos de estado de ploidía, puede ser consideradas intensidades esperadas a efectos comparativos. En una realización, se puede calcular una probabilidad utilizando métodos estadísticos conocidos por los expertos en la técnica, y usando las intensidades medias medidas de los datos resultado de genotipado, y las intensidades medias esperadas de los datos resultado de genotipado. Se puede calcular una probabilidad para cada una de las hipótesis de estado de ploidía, según una hipótesis gaussiana, o mediante un método no paramétrico, como un método kernel para la estimación de densidad. Después se combinan todos los datos con una determinación de ploidía dada y una confianza superior a un umbral establecido. En una realización, el umbral es aproximadamente 80%. En una realización, el umbral es aproximadamente 90%. En una realización, el umbral es aproximadamente 95%. Suponiendo distribuciones gaussianas, el resultado debería ser un conjunto de hipótesis de distribuciones. La Figura 3 muestra una hipótesis de distribución de monosomía (izquierda), disomía (centro), y trisomía (derecha) utilizando la técnica de Media de Cromosoma Completo, y usando datos históricos internos como datos de prueba. En el primer paso del método de medias de cromosoma completo, cada objetivo puede ser normalizado respecto a variación de amplificación. Esto puede hacerse sin normalizar primero respecto a variación cromosómica. En una realización de la presente divulgación, tras calcular el vector [b{1},...,b{24}] de las medias normalizadas de amplificación, el vector puede ser usado para ajustar las medias empleadas para determinar la amplificación del objetivo. Esto dará como resultado nuevas medias normalizadas de amplificación y por tanto un nuevo vector [b { 1 b { 2 4 }] . Esto se puede repetir hasta alcanzar un punto determinado.

Técnica de presencia de progenitores

En una realización de la presente invención, se puede utilizar una técnica especializada estadística denominada “Presencia de Progenitores” (POP), descrita en esta sección, que es especialmente buena para diferenciar las hipótesis que impliquen no contribución de uno o más progenitores (es decir, nulisomía, monosomía, y disomía uniparental) de las que sí lo implican. La técnica estadística descrita en esta sección puede detectar, independientemente para cada progenitor, y para un cromosoma determinado, si hay o no contribución del genoma del progenitor. La determinación se efectúa basada en distancias entre conjuntos de contextos en el punto más

amplio de las curvas CDF. La técnica asigna probabilidades a cuatro hipótesis: {ambos progenitores presentes, ningún progenitor presente, solo la madre, solo el padre}. Las probabilidades se asignan calculando una estadística resumida para cada progenitor, y comparándolo con modelos de datos de prueba para los dos casos de “presente” y “no presente”.

Cálculo de estadística resumida

El algoritmo POP se basa en la idea de que si un progenitor determinado no contribuye, ciertos pares de contextos deben comportarse de forma idéntica. La estadística resumida X^P para progenitor *p* en un cromosoma individual es una medida de la distancia entre esos pares de contextos. En una realización de la presente divulgación, en un cromosoma arbitrario, se pueden definir cinco distancias de contexto d_c^{p1} a d_c^{p5} para cada canal *c* ϵ *X, Y* y cada progenitor p ϵ *{padre, madre}.AABBx* se define como el valor de la curva CDF del contexto *AABB* en el canal X medido en la mayor anchura de envoltura, etc.

$$d_c^{m1} = AABB_c - BBBB_c$$

$$d_c^{m2} = AABB_c - BBBB_c$$

$$d_c^{m3} = AAAB_c - BBAB_c$$

$$d_c^{m4} = AAAA_c - BBAA_c$$

$$d_c^{m5} = AAAA_c - ABAA_c$$

Cuando no hay contribución de la madre, todos los diez de $\{d_c^{mi}\}$ deben ser cero. Cuando hay contribución de la madre, el conjunto de cinco $\{d_X^{mi}\}$ debe ser negativo, y el conjunto de cinco $\{d_Y^{mi}\}$ debe ser positivo. De forma similar, se pueden definir diez distancias $d_c^{f1} \cdots d_c^{f5}$ para el padre, y deben ser cero cuando no existe contribución del padre.

$$d_c^{f1} = BBAB_c - BBBB_c$$

$$d_c^{f2} = BBAA_c - BBBB_c$$

$$d_c^{f3} = ABAA_c - ABBB_c$$

$$d_c^{f4} = AAAA_c - AAAB_c$$

$$d_c^{f5} = AAAA_c - AABB_c$$

Cada distancia puede ser normalizada por la anchura de la envoltura del canal para formar la i[th] distancia normalizada s_c^{pi} para el progenitor *p* en el canal *c.* La anchura de envoltura se mide también en su punto más amplio.

$$s_c^{pi} = d_c^{pi}/abs(AAAA_c - BBBB_c)$$

Se forma una estadística única para el progenitor *p* en el cromosoma actual sumando las distancias normalizadas en los cinco pares de contextos *i* y ambos canales.

$$X^p = \sum_{i=1}^{5} s_Y^{pi} - \sum_{i=1}^{5} s_X^{pi}$$

Distribuciones de prueba

Habiendo calculado una estadística Xp para cada progenitor en un cromosoma determinado, se puede comparar para las distribuciones en los casos de “progenitor presente” y “progenitor no presente” para calcular la probabilidad de cada uno.

En una realización de la presente divulgación, las distribuciones de datos de prueba pueden basarse en un conjunto de blastómeros que hayan sido filtrados usando una o una combinación de otras técnicas de determinación de número de copias. En una realización de la presente divulgación, se consideran determinaciones de hipótesis de la técnica de permutación y la WCM, detectándose la nulisomía utilizando el mínimo criterio de anchura de envoltura requerido. En una realización, para ser incluido en los datos de prueba, un cromosoma debe ser determinado con un alto nivel de confianza. En una realización de la presente divulgación, esta confianza puede ser fijada a aproximadamente 0,6. En una realización de la presente divulgación, esta confianza puede ser fijada a aproximadamente 0,8. En una realización de la presente divulgación, esta confianza puede fijarse a aproximadamente 0,9. En una realización de la presente divulgación, esta confianza puede fijarse a aproximadamente 0,95. Los cromosomas con determinaciones de alta confianza de monosomía paternal o disomía paternal uniparental, se incluyen en el conjunto de datos de “madre no presente”. Los cromosomas no nulisómicos con determinaciones de alta confianza en todas las otras hipótesis se incluyen en el conjunto de datos de “madre presente”, y los conjuntos de datos del padre se construyen de forma similar.

En una realización de la presente divulgación, se puede formar una densidad kernel de cada conjunto de datos, dando como resultado cuatro distribuciones en *X.* Se utiliza una amplia anchura kernel cuando el progenitor está presente, y una estrecha anchura kernel cuando el progenitor no está presente. En una realización de la presente divulgación, la amplia anchura kernel puede ser aproximadamente 0,9, 0,8 o 0,6. En una realización de la presente divulgación, la anchura de kernel estrecha puede ser de aproximadamente 0,1, 0,2, o 0,4. Varios ejemplos de las distribuciones estadísticas resultantes de las técnicas de Presencia de Progenitores se muestran en la Figura 4A-4B. La Figura 4A muestra una distribución de datos genéticos de cada uno de los progenitores cuando hay presentes datos genéticos de los progenitores; la Figura 4B muestra una distribución en ausencia de datos genéticos de cada progenitor. Obsérvese que las distribuciones “presentes” (izquierda) son multimodales, representando los escenarios de “una copia presente” y “dos copias presentes”. Las distribuciones presentes y no presentes en la estadística del padre se muestran en la misma representación gráfica en la Figura 5, poniendo de relieve que X^f puede ser utilizado para distinguir de modo fiable entre los dos casos.

Probabilidades de hipótesis

Las probabilidades de hipótesis de un cromosoma se calculan comparando las estadísticas representativas X^m y X^f con las distribuciones de datos de prueba. La estadística de madre-presente *m* proporciona las funciones de probabilidad *m* = *p*(X^m\ madre presente) y $\overline{m}$ =*p*(X^m\ madre no presente)*f* = *p*(X^f\madre no presente), y la estadística de padre-presente proporciona las funciones de probabilidad *f* = p(X^f| padre presente) y *f*= p(X*f*| padre no presente). Considerando que la presencia de la madre y del padre es independiente, la probabilidad conjunta de una hipótesis sobre ambos progenitores puede ser calculada multiplicando las probabilidades de los progenitores individuales. Por consiguiente, la estructura de probabilidades de hipótesis usual conteniendo nueve probabilidades p(datos| hipótesis) para números de copias parentales, que vayan de cero a dos pueden ser construida como se muestra en la Tabla 1.

	0 padre	1 padre	2 padre
0 madre	$\overline{m}\overline{f}$	$\overline{m}f$	$\overline{m}f$
1 madre	$\overline{m}\overline{f}$	$\overline{m}f$	
2 madre	$\overline{m}\overline{f}$	$\overline{m}f$	

Tabla 1: Probabilidad de hipótesis de datos dados combinando madre y padre

Técnica de presencia de homólogos

Este algoritmo, denominado la técnica de “Presencia de Homólogos” (POH), utiliza información genética parental ajustada por fases, y puede distinguir entre genotipos heterogéneos. Es difícil detectar genotipos donde haya dos cromosomas idénticos utilizando una técnica especializada enfocada en las determinaciones de alelos. La detección de individuos homólogos del progenitor es solo posible utilizando información parental por fases. Sin información parental ajustada por fases, solo pueden identificarse genotipos parentales AA, BB, o AB/BA (heterocigoto). La información parental ajustada por fases distingue entre los genotipos heterocigotos AB y BA. El algoritmo POH se basa en el examen de SNPs donde el progenitor de interés es heterocigoto, y el otro progenitor es homocigoto, como AA|AB, BB|AB, AB|AA o AB|BB. Por ejemplo, la presencia de un B en el blastómero en un SNP donde la madre es AB y el padre AA indica la presencia de M2. Como los datos de una única célula están sujetos a elevadas

tasas de ruido y pérdidas, el cromosoma se segmenta en regiones no superpuestas, y se evalúan las hipótesis en base a estadísticas de los SNPs en una región, más que individualmente.

Es frecuente que resulte difícil diferenciar la trisomía mitótica de la disomía, y algunos tipos de disomía uniparental, donde están presentes dos cromosomas idénticos de un progenitor, se diferencian a veces con dificultad de la monosomía. La trisomía meiótica se distingue por la presencia de ambos homólogos de un único progenitor, en la totalidad del cromosoma en el caso de trisomía de meiosis uno (M1), o en pequeñas secciones del cromosoma en el caso de trisomía de meiosis dos (M2). Esta técnica resulta especialmente útil para la detección de la trisomía M2. La capacidad de diferenciar la trisomía mitótica de la trisomía meiótica es útil, por ejemplo, la detección de trisomía mitótica en blastómero biopsiado de un embrión indica una razonable probabilidad de que el embrión sea mosaico, y se desarrolle normalmente, mientras que una trisomía meiótica indica una muy baja probabilidad de que el embrión sea mosaico, y la probabilidad de que se desarrolle normalmente es menor. Esta técnica resulta particularmente útil en la diferenciación de la trisomía mitótica, la trisomía meiótica y la disomía uniparental. Esta técnica es efectiva para proceder a determinaciones correctas de número de copias con alta precisión.

La presencia de un único progenitor homólogo en el ADN del embrión puede ser detectada examinando contextos indicadores del homólogo. Los contextos indicadores de un homólogo (uno en cada canal) pueden ser definidos como los contextos en los que una señal en ese contexto solo puede proceder de ese homólogo concreto. Por ejemplo, el homólogo de la madre 1 (M_1) se indica en el canal X en el contexto AB|BB, y en el canal Y en el contexto BA|AA.

En una realización de la presente invención, la estructura del algoritmo es como sigue:

1) Ajustar progenitores por fase y calcular umbrales de ruido por cromosoma

2) Segmentar cromosomas

3) Calcular tasas de pérdida de SNP por segmento para cada contexto de interés

4) Calcular la tasa de pérdida de alelos (ADO) para cada progenitor en cada cromosoma objetivo, y las probabilidades de hipótesis en cada segmento

5) Combinar entre segmentos para producir probabilidad de hipótesis de cadena parental de los datos dados para el cromosoma completo

6) Comprobar determinaciones no válidas y luego calcular resultados

(1) Ajuste por fases y cálculo de umbral de ruido

El ajuste por fases del progenitor puede realizarse mediante diversas técnicas. En una realización de la presente divulgación, los datos genéticos parentales son ajustados por fases utilizando un método divulgado en este documento. En una realización de la presente divulgación, pueden ser necesarios aproximadamente 3, 4, 5 o más embriones. En algunas realizaciones de la presente divulgación, el cromosoma puede ser ajustado por fases en segmentos de forma que el ajuste por fases entre un segmento y otro puede no ser consistente. El método de ajuste por fases puede distinguir fenotipos AB y BA con una confianza reportada. En una realización de la presente divulgación, los SNPs no ajustados por fases con la confianza mínima requerida no son asignados a ningún contexto. En una realización de la presente divulgación, la mínima confianza de fase permitida es de aproximadamente 0,8. En una realización de la presente divulgación, la mínima confianza de fase permitida es de aproximadamente 0,9. En una realización de la presente divulgación, la mínima confianza de fase permitida es de aproximadamente 0,95.

El cálculo del umbral de ruido puede basarse en una especificación de percentil. En una realización de la presente divulgación, la especificación de percentil es de aproximadamente 0,90, 0,95 o 0,98. En una realización de la presente divulgación, el umbral de ruido del canal X es el valor del percentil 98 en el contexto BBBB, y de forma similar en el canal Y. Se puede considerar que se ha perdido un SNP si cae por debajo del umbral de ruido de su canal. Se puede calcular un umbral de ruido diferente para cada objetivo, cromosoma y canal.

(2) Segmentación de cromosomas

La segmentación de cromosomas, es decir, aplicar el algoritmo en segmentos de un cromosoma en lugar de un cromosoma completo, es una parte de esta técnica, porque los cálculos se basan en tasas de pérdidas, que se calculan sobre segmentos. Los segmentos que son demasiado pequeños pueden no contener SNPs en todos los contextos requeridos, en especial según disminuye la confianza del ajuste por fases. Los segmentos que son demasiado grandes es más probable que contengan cruces homólogos (es decir, cambio de M1 a M2) que pueden ser confundidos con trisomía. Como las tasas de pérdidas de alelos pueden ser de hasta aproximadamente el 80 por ciento, se pueden requerir muchos SNPs en un segmento para distinguir de forma fiable la pérdida de alelos de la falta de una señal, es decir, donde la tasa de pérdidas esperada sea de aproximadamente un 95 por ciento o superior).

Otra razón de que la segmentación de cromosomas puede ser beneficiosa para la técnica es que permite una ejecución de la técnica más rápidamente, con un nivel determinado de velocidad y potencia de cálculo. Dado que el número de hipótesis y con ello las necesidades del cálculo de la técnica, se escalan más o menos, según el número de alelos considerados aumenta a la n potencia, donde *n* es el número de individuos relacionados, reduciendo el número de alelos considerados se puede mejorar significativamente la velocidad del algoritmo. Segmentos pertinentes pueden ser empalmados juntos después de haber sido ajustados por fases.

En una realización de la presente divulgación, el método de ajuste por fases segmenta cada cromosoma en regiones de 1000 SNPs antes del ajuste por fases. Los segmentos resultantes pueden tener distintos números de SNPs ajustados por fases por encima de un nivel de confianza determinado. En una realización de la presente divulgación, los segmentos del algoritmo utilizado para calcular las tasas de pérdidas pueden no cruzar límites de segmentos de ajuste por fases, porque las definiciones de cadena pueden no ser consistentes. Por consiguiente, la segmentación se realiza por subdivisión de los segmentos de ajuste por fases. En una realización se emplean aproximadamente de 2 a 4 segmentos para un cromosoma. En una realización se utilizan de aproximadamente de 5 a 10 segmentos para un cromosoma. En una realización se utilizan aproximadamente de 10 a 20 segmentos para un cromosoma. En una realización se utilizan aproximadamente de 20 a 30 segmentos para un cromosoma. En una realización se utilizan aproximadamente de 30 a 50 segmentos para un cromosoma. En una realización se utilizan más de aproximadamente 50 segmentos para un cromosoma

En una realización de la presente divulgación, se utilizan aproximadamente 20 segmentos en grandes cromosomas, y aproximadamente 6 segmentos en cromosomas muy pequeños. En una realización de la presente divulgación el número de segmentos usados se calcula para cada cromosoma, oscilando entre aproximadamente 6 y 20, y varía linealmente con el número total de SNPs en el cromosoma. En una realización de la presente divulgación, si el número de segmentos de ajuste por fases es superior o igual al número de segmentos deseado, se utilizan los segmentos de ajuste por fases como tal, y si no, los segmentos de ajuste por fases se subdividen uniformemente en *n* segmentos cada uno, donde *n* es el mínimo requerido para alcanzar el número de segmentos deseado.

(3) Cálculo de tasas de pérdidas

Los datos sobre un segmento particular de cromosoma se resumen por las tasas de pérdidas en un conjunto de contextos. La tasa de pérdidas puede definirse, para esta sección, como la fracción de SNPs en el contexto en cuestión (con su canal especificado) que mide por debajo del umbral de ruido. Se pueden medir seis contextos para cada progenitor. Las tasas de pérdidas $\hat{a}_x$ y $\hat{a}_y$ pueden reflejar la tasa de pérdidas de alelos, y las tasas de pérdidas y pueden indicar la presencia de homólogo. La siguiente tabla muestra un ejemplo de los contextos asociados a cada tasa de pérdidas para cada progenitor. La tasa de pérdidas medida y el número de SNPs para cada contexto deben ser almacenadas. Obsérvese que cada una de las tres tasas de pérdidas en la Tabla 2 es medida en dos contextos distintos para cada progenitor.

	mom, X	mom, Y	dad, X	dad, Y
$\hat{a}$	AABB	BBAA	BBAA	AABB
$\hat{a}^1$	ABBB	BAAA	BBAB	AABA
$\hat{a}^2$	BABB	ABAA	BBBA	AAAB

Tabla 2: Contextos para tasas de pérdidas requeridas

(4) Estimación de probabilidad máxima de ADO

Esta sección contiene un debate sobre un método para calcular la tasa de pérdida de alelos a* para cada progenitor en cada objetivo, en base a probabilidades de la forma p (D_s|Mi, a) y p(D_s|Fi,a). ADO puede definirse como la probabilidad de pérdida de señal en un SNP AB. D_s puede definirse como el conjunto de tasas de pérdida del contexto medido en un segmento de un cromosoma, y *Mi, Fi,* son las hipótesis de cadena parental. En una realización de la presente divulgación, los cálculos se realizan utilizando probabilidades log debido a las probabilidades relativamente pequeñas generadas por multiplicación entre contextos y segmentos.

La tasa de pérdidas de alelos puede ser calculada utilizando una estimación de probabilidad máxima calculada por búsqueda de cuadrícula de fuerza bruta en el rango permisible. En una realización de la presente divulgación, el rango de búsqueda [a_{min}, a_{max}] puede ser establecido en aproximadamente [0,4; 0,7]. A niveles altos de ADO, resulta difícil distinguir entre presencia y ausencia de una señal, porque el ADO se aproxima a la tasa de pérdidas de umbral de ruido de aproximadamente 0,95.

En una realización de la presente divulgación, la tasa de pérdidas de alelos se calcula para un objetivo particular, para cada progenitor, utilizando el siguiente algoritmo. En una realización de la presente divulgación, el cálculo puede realizarse utilizando operaciones matriz, en lugar de para cada objetivo y cromosoma individualmente.

Para un $\epsilon\ [a_{min}, a_{max}]$

para $\in [1, 22]$ (22 cromosomas)

Calcular $P(D_S|Mi, a)$ $\forall I, \forall s$ en cromosoma

$\mathcal{M}_s^* =$ arg max $P(D_S|Mi, a)$ (maximizar sobre hipótesis en cada segmento)

$P(D_{ch}|M_s^*, a) = \prod_s \quad P(D_s|M_s^*, a)$ (combinar entre segmentos en el cromosoma)

$\Lambda(a) = \prod_{ch} \quad P(D_{ch}|M_s^*, a)$ (combinar entre cromosomas)

$a^* = \arg\max \Lambda(a)$ (optimizar sobre a)

Modelado de probabilidades de datos

En una realización de la presente divulgación, la optimización de ADO puede utilizar un modelo de tasa de pérdidas en varios contextos como una función de hipótesis de cadena parental y ADO. Las pérdidas de SNP en un solo segmento de cromosoma pueden ser consideradas variables I.I.D. Bernoulli, y se puede esperar que la tasa de pérdidas esté distribuida normalmente con una media μ y una desviación estándar $\sigma = \sqrt{\mu(1-\mu)/N}$ donde N es el número de SNPs medidos. El modelo de tasa de pérdida de alelos puede calcular μ como una función de la hipótesis, ADO, y contexto.

La hipótesis y el contexto juntos determinan un genotipo para un SNP, como AB. El genotipo y la tasa de ADO determinan entonces μ. En una realización de la presente divulgación, las hipótesis para la madre son {M_0, Mi, M_2, M12, M11, M22}. Otros conjuntos de hipótesis pueden ser utilizados igualmente bien. Mo significa que no hay presente ningún homólogo de la madre. M11 y M22 son casos donde están presentes dos copias idénticas de la madre. No indican trisomía meiótica. Las hipótesis consistentes con disomía son M1 y M2.

La Tabla 3 recoge μ por hipótesis de madre y las diversas mediciones de tasas de pérdidas en esta realización de la presente divulgación. La tabla idéntica puede ser utilizada para las correspondientes hipótesis de cadena de padre. Recordemos que p es la tasa de pérdidas que define el umbral de ruido, y por consiguiente la tasa de pérdidas esperada de un canal sin presencia de alelo.

	M_0	M_1	M_2	M_{12}	M_{11}	M_{22}
$\hat{a}$	p	a	a	a^2	a^2	a^2
$\hat{s}^1$	p	a	p	a	a^2	p
$\hat{s}^2$	p	p	a	a	p	a^2

Tabla 3: Modelo de tasa de pérdidas de segmento esperada por hipótesis de cadena

En cada segmento, las tres tasas de pérdidas $\hat{a},\ \hat{s}^1$ and $\hat{s}^2$ se miden en ambos canales. Así, los datos totales D_s de un segmento consisten en 6 mediciones de tasa de pérdidas, y la probabilidad $P(D_S|Mi, a)$ es el producto de las 6 probabilidades correspondientes bajo las distribuciones normales determinadas por μ de la Tabla 3.

Como la identificación de los SNPs para las tasas de pérdidas $\hat{s}^1$ y $\hat{s}^2$ depende del ajuste por fases parental, en algunos contextos puede no haber SNPs identificados. Cada una de las tres tasas de pérdidas medidas $\hat{a},\ \hat{s}^1$ y $\hat{s}^2$ puede ser medida en dos contextos distintos correspondiendo a los dos canales. Si alguna de las tres no tiene datos en ninguno de sus contextos, no se pueden calcular las probabilidades para ese segmento. Los cromosomas con determinación de nulisomía por el test de anchura de envoltura estándar pueden no estar incluidos.

(5) Calcular las probabilidades cromosómicas combinando segmentos

Los cálculos de probabilidades descritos más arriba proporcionan una probabilidad de datos $P(D_S|Mi)$ en cada segmento s para cada hipótesis de cadena parental M. Los dos progenitores pueden aún ser considerados independientemente. Las probabilidades de cadena pueden normalizarse entonces, de forma que la suma de todas las probabilidades en un solo segmento sea uno. Las probabilidades normalizadas del segmento s serán denominadas $\{\Lambda_S(Mi)\}$. Este proceso dependerá también de las longitudes de los segmentos normalizadas $\{x_s\}$, definidas como la fracción de los SNPs de un cromosoma contenida en el segmento s.

En una realización de la presente divulgación, las probabilidades de todos los segmentos pueden combinarse ahora para formar un conjunto de probabilidades cromosómicas para el número de cadenas distintas presentes. Todos los datos de un cromosoma se combinan en D_{ch}. Las hipótesis cromosómicas son S_0^m, S_1^m, S_2^m para la madre. S_1^m es la hipótesis de que solo un homólogo distinto está presente cada vez, lo que permite las hipótesis de cadena M1; M11; M2; M22 S_2^m es la hipótesis de trisomía meiótica, donde la madre ha aportado dos cadenas distintas. Se comentarán las hipótesis sobre el número de cadenas de la madre; las hipótesis sobre la cadena del padre pueden ser calculadas de forma análoga.

S_0^m se corresponde uno-a-uno con la hipótesis de no cadena M_0. Por consiguiente, la probabilidad de no copias es simplemente la suma (ponderada por longitud de segmento) de las probabilidades de no cadena en cada segmento.

$$P(D_{ch}|S_0^m) = \sum_s \Lambda_s(M_0)x_s$$

S_1^m (una copia cada vez) corresponde a las hipótesis de cadena M1; M11; M_2; M22. Sin hacer ninguna suposición sobre recombinación, cabe esperar que una sola copia parental sea la cadena M1 o M2 en todos los segmentos. En esta realización de la presente divulgación, en lugar de intentar detectar cuántas copias de una sola cadena están presentes, se incluyen también las hipótesis de doble cadena M11 y M22. En otra realización de la presente divulgación, se pueden agrupar M1 y M2 en una hipótesis, y M11 y M22 pueden agruparse en otra hipótesis. En otras realizaciones, otras hipótesis pueden referirse a otros agrupamientos del estado real del material genético. Nuevamente, la probabilidad cromosómica es simplemente una suma ponderada.

$$P(D_{ch}|S_1^m) = \sum_s (\Lambda_s(M_1) + \Lambda_s(M_{11}) + \Lambda_s(M_2) + \Lambda_s(M_{22}))x_s$$

La trisomía meiótica se caracteriza por la presencia de dos cromosomas no idénticos de un solo progenitor. Dependiendo del tipo de error meiótico, puede ser una copia completa de cada uno de los homólogos del progenitor (meiosis-1), o puede ser dos recombinaciones distintas de los homólogos del progenitor (meiosis-2). En el primer caso, el resultado es la hipótesis de cadena M_{I2} en todos los segmentos, pero en el segundo caso el resultado es M_{I2} solo donde las dos combinaciones distintas no coinciden. Por consiguiente, el enfoque de la suma ponderada aplicado a las otras hipótesis puede no ser apropiado

El cálculo de la probabilidad de trisomía meiótica se basa en el supuesto de que recombinaciones únicas serán distintas en por lo menos una región contínua cubriendo por lo menos una cuarta parte del cromosoma. En otras realizaciones, se pueden utilizar otros tamaños de región continua en la que recombinaciones únicas sean distintas. Un umbral de detección que sea demasiado bajo puede dar como resultado la determinación incorrecta de trisomías, debido a recombinaciones de segmento medio y ruido. Como la trisomía de meiosis-2 no se corresponde con ninguna hipótesis de cadena de cromosoma completo, la probabilidad puede no ser proporcional a la suma de las probabilidades de segmentos, como lo es para los otros dos números de copias. En vez de eso, la confianza en la hipótesis meiótica depende de si se ha cumplido o no el umbral meiótico, y la confianza global del cromosoma.

En una realización de la presente divulgación, los cromosomas pueden ser reconstruidos recombinando los segmentos junto con sus probabilidades relativas utilizando los siguientes pasos:

1. Hallar la longitud x de la región continua más larga con A(M12) > 0:8 combinando 5 segmentos adyacentes
2. Si x > 0:25 establecer la marca meiótica como cierta. De lo contrario establecerla como falsa.
3. Calcular la confianza general en el cromosoma promediando la confianza en la hipótesis más probable de cada segmento $C = \Sigma_s x_s \max\Lambda_s(Mi)4$. Si la marca meiótica es cierta, dejar $P(D_{ch}|S_2^m) = C.$ normalizado. De lo contrario dejar $P(D_{ch}|S_2^m) = 1\text{-}C.$

El resultado es que si se activa la marca meiótica en un cromosoma de alta confianza, la hipótesis meiótica tendrá en correspondencia una alta confianza. Si no se activa la marca meiótica, la hipótesis meiótica tendrá una baja confianza.

(6) Comprobación de determinaciones no válidas y cálculo de resultados CNC

El paso final es calcular probabilidades en números de copias parentales verdaderos sin distinción entre error meiótico y mitótico. La anotación estándar HNmNf se adaptará para progenitores individuales, donde N_m es el número de cadenas de la madre presentes, y Nf es el número de cadenas del padre presentes.

$$P(D_{ch}|H0x) = P(D_{ch}|S_0^m)$$

$$P(D_{ch}|H1x) = P(D_{ch}|S_1^m)$$

$$P(D_{ch}|H2x) = P(D_{ch}|S_2^m)\,P(\text{meiotic}) + P(D_{ch}|S_1^m)\,P(\text{mitotic})$$

La fórmula final viene explicada por el hecho de que la trisomía puede derivar de dos eventos dispares meiótico y error mitótico. El error meiótico corresponde a la hipótesis S_2^m (2 copias diferentes) y el error mitótico corresponde a la hipótesis S_2^m (duplicado del mismo homólogo). Las probabilidades previas de esos dos eventos se supone que son iguales. Como resultado, una confianza muy alta en la hipótesis pone aproximadamente una confianza igual en H1x y H2x, pero una confianza muy alta en la hipótesis S_2^m favorece solamente a H2x.

Este algoritmo es muy adecuado para detectar segmentación en cromosomas. Una disomía segmentada se caracteriza por la presencia de una copia de cada progenitor, donde por lo menos una copia parental es incompleta. Si un progenitor tiene una confianza superior a aproximadamente el 80 por ciento en la hipótesis de 0 cadenas (M0 o F0) para por lo menos una cuarta parte del cromosoma, este cromosoma puede ser marcado como “monosomía segmentada” incluso si los cálculos de confianza utilizando otras técnicas especializadas dan como resultado una determinación de disomía. Esta marca de segmentación puede combinarse con la marca de segmentación de la técnica de Permutación, de forma que cada una de ellas puede detectar un error independientemente. Si la determinación de algoritmo global es monosomía, la marca segmentada puede no ser activada porque resultaría redundante.

En este punto de la ejecución de la técnica, se han asignado confianzas de hipótesis de copias para cada progenitor para cada cromosoma donde había disponibles tasas de pérdidas para por lo menos un segmento. No obstante, algunos cromosomas pueden no haber sido ajustados por fases con una alta confianza, y sus probabilidades pueden reflejar tasas de pérdidas que estaban solo disponibles para una fracción muy pequeña del cromosoma. En una realización de la presente divulgación, para evitar realizar determinaciones basadas en datos insuficientes o no claros, pueden realizarse comprobaciones para eliminar determinaciones en cromosomas con ajuste por fases incompleto o resultados con mucho ruido.

Después de realizar las comprobaciones, las hipótesis de copias parentales pueden ser convertidas a hipótesis CNC estándar. Para las copias maternas N_m y las copias paternas Nf, la probabilidad de las hipótesis CNC HN_mNf es simplemente una multiplicación de las probabilidades de copias parentales independientes. Si uno de los progenitores no se había determinado debido a un ajuste por fases incompleto o datos con ruido, el algoritmo puede dar probabilidades uniformes en este progenitor, pero seguir determinando el otro progenitor.

$$P(D|HN_mN_f) = P(D|HN_mx)\,P(D|HxN_f)$$

Comprobación de ajuste por fases incompleto

La cobertura del ajuste por fases en un cromosoma es la suma de las longitudes de segmento para las que se han calculado las probabilidades. En algunas realizaciones de la presente divulgación, no se calculan probabilidades cuando alguna de las tres mediciones de tasas de pérdidas carece de datos. Si la cobertura de ajuste por fases es inferior a la mitad, no se produce determinación. En el caso donde la trisomía meiótica sea marcada por una secuencia de segmentos M12 o F12 de longitud combinada de aproximadamente 0,25, toda cobertura de ajuste por fases de menos de 0,75 no es suficiente para descartar tal segmento. No obstante, si se detecta un segmento meiótico de longitud 0,25 puede aún determinarse. En una realización de la presente divulgación, con una cobertura de ajuste por fases de entre aproximadamente 0,5 y 75 se procede como sigue.

*si está marcado como trisomía, la determinación de ploidía está como completamente ajustada por fases

* si la determinación es monosomía parcial o completa, la determinación de ploidía es como si estuviera completamente ajustada por fases

* de lo contrario, no realizar determinación (establecer probabilidad uniforme para las copias de este progenitor)

Comprobar cromosomas ruidosos

Algunos cromosomas pueden resistirse a la clasificación utilizando este algoritmo. A pesar de un ajuste por fases de alta confianza y de las probabilidades de segmentos, los resultados de cromosoma completo no son claros. En algunos casos, esos cromosomas se caracterizan por cambios frecuentes entre hipótesis de máxima probabilidad. Aunque se esperan solamente unos pocos eventos de recombinación por cromosoma, estos cromosomas pueden presentar cambios casi aleatorios entre hipótesis. Como la hipótesis meiótica es activada por una secuencia meiótica de aproximadamente 0,25, se pueden provocar con frecuencia falsas trisomías en cromosomas ruidosos.

En algunas realizaciones de la presente divulgación, el algoritmo declara un “cromosoma ruidoso” combinando segmentos adyacentes con la misma hipótesis de probabilidad máxima. La longitud media de esos nuevos segmentos se compara con la longitud media del conjunto de segmentos originales. Si este cociente es inferior a dos, pocos segmentos adyacentes pueden tener hipótesis coincidentes, y el cromosoma puede ser considerado ruidoso. Este test se basa en el supuesto de que se espera que la segmentación original sea un tanto uniforme y densa. Un cambio a un algoritmo de segmentación óptimo requeriría un nuevo criterio.

Si un cromosoma es declarado ruidoso en un progenitor específico, las hipótesis de copias de ese progenitor pueden ser consideradas como uniformes, y las marcas de monosomía meiótica y segmentada como falsas.

Técnica del cromosoma sexual

Las técnicas descritas más arriba han sido diseñadas para cromosomas autosómicos. Dado que los estados genéticos probables de los cromosomas sexuales (X e Y) son distintos, técnicas diferentes pueden ser más apropiadas. En esta sección se describen otras técnicas diseñadas específicamente para la determinación del estado de ploidía de los cromosomas sexuales.

Además de que sean distintos los números de los cromosomas sexuales esperados, la determinación del estado de ploidía de los cromosomas sexuales se complica por el hecho de que hay regiones en el cromosoma X e Y que son homólogas, y otras que son similares pero no polimórficas. El cromosoma Y puede ser considerado un mosaico de distintas regiones, y el comportamiento de las sondas Y depende en gran medida de la región a la que se unen en el cromosoma Y. Muchas de las sondas Y no miden SNPs per se; en lugar de eso se unen a localizaciones que son no polimorfas en ambos cromosomas X e Y. En algunos casos, una sonda se unirá a una localización que es siempre AA en el cromosoma X, pero siempre BB en el cromosoma Y, o viceversa. Estas sondas son denominadas de “dos grupos” (clusters) porque cuando una de esas sondas se aplica a un conjunto de muestras masculinas y femeninas, el gráfico de dispersión resultante se divide siempre en dos grupos, segregados por sexo. Los machos son siempre heterocigotos y las hembras siempre homocigotas.

Técnica del cromosoma XYZ

En una realización de la presente divulgación, la determinación de ploidía de los cromosomas sexuales se enfoca considerando un cromosoma abstracto denominado *“cromosoma 23”,* compuesto por cuatro subcromosomas distintos, denominados X, Y, XY, y Z. El cromosoma XY corresponde a las sondas que hibridan con los cromosomas X y los cromosomas Y en lo que es conocido como las regiones pseudoautosómicas. Por el contrario, las sondas asociadas al cromosoma X se espera que hibriden solo con el cromosoma X, y las sondas asociadas al cromosoma Y se espera que hibriden solamente con el cromosoma Y. El cromosoma Z corresponde a esas sondas de “dos grupos” que hibridan con el cromosoma Y, en lo que se conoce como la región X transpuesta – la región que concuerda en aproximadamente el 99,9% con una región similar en el cromosoma X, y cuyos valores de alelos son polares a sus cognados en X. Así, una sonda Z medirá AB (sin tener en cuenta el ruido) en una muestra masculina, y AA o BB en una muestra femenina, dependiendo del locus.

El siguiente debate describe las matemáticas base de esta técnica. En términos de los cromosomas sexuales componentes, el objetivo de esta técnica es distinguir los siguientes casos: $\{\varnothing, X, Y, XX, XY, YY, XXX, XXY, XYY, XXYY\}$. Hay que advertir que si el cromosoma 23 es euploide, debe ser uno de $\{XX, XY\}$ y por tanto debe tener un número de copias de 2. En los casos de disomía uniparental: XX de la madre y nada del padre, o YY del padre, se puede asignar arbitrariamente un número de copias de 5, o fusionarlas con las hipótesis de monosomía.

La conexión entre los subcromosomas X e Y se expresa solamente en la distribución previa conjunta $P(n_X^F, n_Y^F)$ en el número de subcromosomas de X e Y aportados por el padre.

Anotación

1. *n* es el número de copias del cromosoma para el cromosoma 23.

2. n_X^M es el número de copias del subcromosoma X aportado al embrión de la madre: 0, 1, o 2. A efectos de anotación, es conveniente también definir $n_Y^M = 0$ como el número de copias del subcromosoma Y aportadas al embrión por la madre.

3. (n_X^F, n_Y^F) es el número de copias de subcromosomas X e Y aportados conjuntamente al embrión por el padre. Esos pares de números de copias deben pertenecer al conjunto $\{(0,0),(0,1),(1,0),(2,0),(1,1),(0,2)\}$.

Obsérvese que las precedentes tres variables definidas satisfacen la limitación $n_X^M + n_X^F + n_Y^F = n.$

4. Definir $n_{XY}^M = n_X^M,\ n_{XY}^F = n_X^F + n_Y^F$

5. Definir $n_Z^M = n_X^M,\ n_Z^F = n_X^F + n_Y^F$

6. Pd (¿) es la tasa de pérdidas, y $f(p_d)$ es una previa de esta tasa.

7. Pa es la tasa de entradas, y $f(p_a)$ es una previa de esta tasa.

8. c es el umbral de corte para no determinaciones.

9. $D_X = \{(x_{Xk}, y_{Xk})\}$ es el conjunto de respuestas de plataforma en bruto en canales x e y en todos los SNPs k en el subcromosoma X. De forma similar $D_Y = \{(x_{Yk}, y_{Yk})\}$ es el conjunto de respuestas de plataforma en bruto en los canales x e y en todos los SNPs k en el subcromosoma Y, $D_{XY} = \{(x_{XYk}, y_{XYk})\}$ es el conjunto de respuestas de plataforma en bruto en canales x e y en todos los SNPs k en el subcromosoma XY, y $D_Z = \{(x_{Zk}, y_{Zk})\}$ es el conjunto de respuestas de plataforma en bruto en los canales x e y sobre todos los SNPs k en el subcromosoma Z.

10 $D_X(c) = \{G(x_{Xk}, y_{Xk}); c\} = \{\hat{g}_{Xk}^{(c)}\}$ es el conjunto de determinaciones de genotipo en todos los SNPs k en el subcromosoma X, y de forma similar para los subcromosomas Y, XY, y Z. Adviértase que las determinaciones de genotipo dependen del umbral de corte de no determinación e.

11 Definir un índice j de subcromosoma, donde $j \in \{X, Y, XY, Z\}.$ En este caso, podemos hacer referencia a $D_j(c)$ en relación con los datos asociados con el subcromosoma j.

12 $\hat{g}_{jk}^{(c)}$ es la determinación de genotipo en el SNP kth (como opuesto al valor real) en el subcromosoma j: uno de AA, AB, BB, o NC (sin determinación).

13 Considerando una determinación de genotipo $\hat{g}$ en SNP k, las variables $(\hat{g}^A, \hat{g}^B)$ son variables indicadoras (1o 0). Formalmente, $\hat{g}^A = (A \in \hat{g}),$ y $\hat{g}^B = (B \in \hat{g})$.

14 $M - \{g_{jk}^M\}$ es la secuencia real conocida de determinaciones de genotipo en la madre, en el subcromosoma j. g^M se refiere al valor del genotipo en algún locus particular. Adviértase que, para j $= \mathrm{Y}, \{g_{Yk}^M\}$ se considera como una secuencia de no determinaciones: NC.

15 $F = (g^F_{jk})$ es la secuencia real conocida de determinaciones de genotipo en el padre, en el subcromosoma j. g^F se refiere al valor del genotipo en algún locus particular.

16 $C_{MF}(j)$ es la clase de genotipos parentales conjuntos concebible que puede ocurrir en el subcromosoma j. Cada elemento de $C_{MF}(j)$ es una tupla de la forma, ej., (AA, AB), y describe uno de los posibles genotipos conjuntos del padre y la madre. Los conjuntos $C_{MF}(j)$ se enumeran en su totalidad aquí:

a. $C_{MF}(X) = \{AA, AB, BB\} \times \{AA, BB\}$

b. $C_{MF}(Y) = \{NC\} \times \{AA, BB\}$

c. $C_{MF}(XY) = \{AA, AB, BB\} \times \{AA, AB, BB\}$

d. $C_{MF}(Z) = \{AA, BB\} \times \{AB\}$

17 n^A_j, n^B_j son el número de copias verdadero de A y B en el embrión (implícitamente en el locus k), respectivamente en el subcromosoma j. Los valores deben estar en 0,1,2,3,4 para $j \in \{X, XY, Z\}$ y en 0,1,2 para $j \in \{Y\}$.

18 c^{AM}_j, c^{BM}_j son el número de alelos A y alelos B aportados respectivamente por la madre al embrión (implícitamente en el locus k) en el subcromosoma j. Para j = X o XY o Z, los valores deben estar en 0, 1,2, y no deben sumar más de 2. Para j = Y, los valores deben ser (0,0). De forma similar, c^{AF}_j, c^{BF}_j son el número de alelos A y alelos B aportados respectivamente por el padre al embrión (implícitamente en el locus k) en el subcromosoma j. El padre tiene la limitación adicional para j = X o j = Y que uno de c^{AF}_j, c^{BF}_j debe ser cero, reflejando el hecho de que el padre no puede aportar material heterocigoto de cualquier cromosoma sexual. Para j=XY, no hay tal limitación.

Para j=Z, las limitaciones son como sigue:

1. Cuando el locus es homo AA en la madre, tenemos $c^{AF}_Z = n^F_X$ y $c^{BF}_Z = n^F_Y$.
2. Cuando el locus es homo BB en la madre, tenemos $c^{BF}_Z = n^F_X$ y $c^{AF}_Z = n^F_Y$.

En conjunto los cuatro valores $\{c^{AM}_j, c^{BM}_j, c^{AF}_j, c^{BM}_j\}$ determinan exactamente el verdadero genotipo del embrión en el subcromosoma j. Por ejemplo, si los valores fueran (1,1) y (1,0), el embrión tendría tipo AAB.

Adviértase también que las siguientes limitaciones se aplican a todos los j:

1. $c^{AM}_j + c^{BM}_j = n^M_j$
2. $c^{AF}_j + c^{BF}_j = n^F_j$

La siguiente solución es aplicable solamente al cromosoma 23 y tiene en cuenta la interrelación entre subcromosomas X,Y, y XY.

$$P(n|D_X(c),D_Y(c),D_{XY}(c),M,F) = \sum_{(n_X^M,n_X^F,n_Y^F)\in n} P(n_X^M,n_X^F,n_Y^F|D_X(c),D_Y(c),D_{XY}(c),M,F)$$

$$P(n_X^M,n_X^F,n_Y^F|D_X(c),D_Y(c),D_{XY}(c),M,F) = \frac{P(n_X^M)P(n_X^F,n_Y^F)P(D_X(c),D_Y(c),D_{XY}(c)|n_X^M,n_X^F,n_Y^F,M,F)}{\sum_{(n_X^M,n_X^F,n_Y^F)} P(n_X^M)P(n_X^F,n_Y^F)P(D_X(c),D_Y(c),D_{XY}(c)|n_X^M,n_X^F,n_Y^F,M,F)}$$

$P(n_X^F,n_Y^F)$ es una distribución previa que puede ser establecida razonablemente. Las probabilidades de (1,0) y (0,1) pueden ser establecidas razonablemente altas, dado que son los estados de euploidía.

$$P(D_X(c),D_Y(c),D_{XY}(c)|n_X^M,n_X^F,n_Y^F,M,F) = P(D_X(c)|n_X^M,n_X^F,M,F) \times P(D_Y(c)|n_Y^F,M,F) \times P(D_{XY}(c)|n_{XY}^M,n_{XY}^F,M,F)$$

En lo anterior hay que tener en cuenta que $n_{XY}^M = n_X^M$ y $n_{XY}^F = n_X^F + n_Y^F$.

$$P(D_j(c)|n_j^M,n_j^F,M,F) = \int\int f(p_d)f(p_a)P(D_j(c)|n_j^M,n_j^F,M,F,p_d,p_a)dp_d dp_a$$

$$(*)\ P(D_j(c)|n_j^M,n_j^F,M,F,p_d,p_a) = \prod_k P(G(x_{jk},y_{jk};c)|n_j^M,n_j^F,g_{jk}^M,g_{jk}^F,p_d,p_a)$$

Procedimiento (*) en el cromosoma XY

El caso del cromosoma XY se comporta de forma similar a cualquier autosoma. Las matemáticas se comentan aquí.

$$P(D_X(c)|n_X^M,n_X^F,M,F,p_d,p_a) = \prod_k P(G(x_{Xk},y_{Xk};c)|n_X^M,n_X^F,g_{Xk}^M,g_{Xk}^F,p_d,p_a)$$

$$= \prod_{\substack{g^M\in\{AA,AB,BB\}\\ g^F\in\{AA,AB,BB\}\\ \hat{g}\in\{AA,AB,BB,NC\}}} \prod_{\{x:\, g_{Xk}^M=g^M,\, g_{Xk}^F=g^F,\, \hat{g}_{Xk}^{(c)}=\hat{g}\}} P(\hat{g}|n_X^M,n_X^F,g^M,g^F,p_d,p_a)$$

$$= \prod_{\substack{g^M \in \{AA,AB,BB\} \\ g^F \in \{AA,AB,BB\} \\ \hat{g} \in \{AA,AB,BB,NC\}}} P(\hat{g}|n_X^M, n_X^F, g^M, g^F, p_d, p_a)^{\left|\left\{k:\, g_{Xk}^M = g^M, g_{Xk}^F = g^F, \hat{g}_{Xk}^{(c)} = \hat{g}\right\}\right|}$$

$$= exp\left(\sum_{\substack{g^M \in \{AA,AB,BB\} \\ g^F \in \{AA,AB,BB\} \\ \hat{g} \in \{AA,AB,BB,NC\}}} \left|\left\{k:\, g_{Xk}^M = g^M, g_{Xk}^F = g^F, \hat{g}_{Xk}^{(c)} = \hat{g}\right\}\right| \times \log P(\hat{g}|n_X^M, n_X^F, g^M, g^F, p_d, p_a) \right)$$

$$P(\hat{g}|n_X^M, n_X^F, g^M, g^F, p_d, p_a)$$

$$= \sum_{n^A, n^B} \underbrace{P(n^A, n^B|n_X^M, n_X^F, g^M, g^F,)}_{genetic\ modeling} \overbrace{P(\hat{g}^A|n^A, p_d, p_a) P(\hat{g}^B|n^B, p_d, p_a)}^{platform\ modeling}$$

Procedimiento (*) en el cromosoma X

Aquí las limitaciones adicionales son que el padre no es nunca heterocigoto en X.

$$P(D_X(c)|n_X^M, n_X^F, M, F, p_d, p_a) - \prod_k P(G(x_{Xk}, y_{Xk}; c)|n_X^M, n_X^F, g_{Xk}^M, g_{Xk}^F, p_d, p_a)$$

$$= \prod_{\substack{g^M \in \{AA,AB,BB\} \\ g^F \in \{AA,BB\} \\ \hat{g} \in \{AA,AB,BB,NC\}}} \prod_{\left\{k:\, g_{Xk}^M = g^M, g_{Xk}^F = g^F, \hat{g}_{Xk}^{(c)} = \hat{g}\right\}} P(\hat{g}|n_X^M, n_X^F, g^M, g^F, p_d, p_a)$$

$$= \prod_{\substack{g^M \in \{AA,AB,BB\} \\ g^F \in \{AA,BB\} \\ \hat{g} \in \{AA,AB,BB,NC\}}} P(\hat{g}|n_X^M, n_X^F, g^M, g^F, p_d, p_a)^{\left|\left\{k:\, g_{Xk}^M = g^M, g_{Xk}^F = g^F, \hat{g}_{Xk}^{(c)} = \hat{g}\right\}\right|}$$

$$= exp\left(\sum_{\substack{g^M \in \{AA,AB,BB\} \\ g^F \in \{AA,BB\} \\ \hat{g} \in \{AA,AB,BB,NC\}}} \left|\left\{k:\, g_{Xk}^M = g^M, g_{Xk}^F = g^F, \hat{g}_{Xk}^{(c)} = \hat{g}\right\}\right| \times \log P(\hat{g}|n_X^M, n_X^F, g^M, g^F, p_d, p_a) \right)$$

$$P(\hat{g}|n_X^M, n_X^F, g^M, g^F, p_d, p_a)$$

$$= \sum_{n^A, n^B} \underbrace{P(n^A, n^B|n_X^M, n_X^F, g^M, g^F,)}_{genetic\ modeling} \overbrace{P(\hat{g}^A|n^A, p_d, p_a) P(\hat{g}^B|n^B, p_d, p_a)}^{platform\ modeling}$$

Procedimiento (*) en el cromosoma Y

Aquí las limitaciones son que el número de copias de la madre es 0, y el padre no es nunca heterocigoto en Y.

$$P(D_Y(c)|n_Y^F, M, F, p_d, p_a) = \prod_k P(G(x_{Yk}, y_{Yk}; c)|n_Y^F, g_{Yk}^F, p_d, p_a)$$

$$= \prod_{\substack{g^F \in \{AA,BB\} \\ \hat{g} \in \{AA,AB,BB,NC\}}} \prod_{\{k: g_{Yk}^F = g^F, \hat{g}_{Yk}^{(o)} = \hat{g}\}} P(\hat{g}|n_Y^F, g^F, p_d, p_a)$$

$$= \prod_{\substack{g^F \in \{AA,BB\} \\ \hat{g} \in \{AA,AB,BB,NC\}}} P(\hat{g}|n_Y^F, g^F, p_d, p_a)^{\left|\left\{k: g_{Yk}^F = g^F, \hat{g}_{Yk}^{(o)} = \hat{g}\right\}\right|}$$

$$= exp\left(\sum_{\substack{g^F \in \{AA,BB\} \\ \hat{g} \in \{AA,AB,BB,NC\}}} \left|\left\{k: g_{Yk}^F = g^F, \hat{g}_{Yk}^{(o)} = \hat{g}\right\}\right| \times \log P(\hat{g}|n_Y^F, g^F, p_d, p_a)\right)$$

$$P(\hat{g}|n_Y^F, g^F, p_d, p_a) = \sum_{n^A, n^B} \underbrace{P(n^A, n^B \mid n_Y^F, g^F)}_{genetic\ modeling} \overbrace{P(\hat{g}^A|n^A, p_d, p_a) P(\hat{g}^B|n^B, p_d, p_a)}^{platform\ modeling}$$

$$P(n^A, n^B \mid n_Y^F, g^F,) = P(n^A, n^B \mid n_Y^F, g^F, n_Y^M = 0, g^M = NC)$$

Aquí la solución se continúa para todos los subcromosomas. Téngase en cuenta que cuando j = Y, $n_j^M = 0$ y $g_{jk}^M = NC$ para todos los k.

$$P(\hat{g}|n_j^M, n_j^F, g^M, g^F, p_d, p_a)$$

$$= \sum_{n^A, n^B} \underbrace{P(n^A, n^B|n_j^M, n_j^F, g^M, g^F,)}_{genetic\ modeling} \overbrace{P(\hat{g}^A|n^A, p_d, p_a) P(\hat{g}^B|n^B, p_d, p_a)}^{platform\ modeling}$$

$$P(\hat{g}^A|n^A, p_d, p_a)$$

$$= \hat{g}^A \left((1 - p_d{}^{n^A}) + (n^A = 0) p_a \right)$$

$$+ (1 - \hat{g}^A) \left((n^A > 0) p_d{}^{n^A} + (n^A = 0)(1 - p_a) \right)$$

$$P(\hat{g}^B|n^B, p_d, p_a)$$

$$= \hat{g}^B \left((1 - p_a{}^{n^B}) + (n^B = 0) p_a \right)$$

$$+ (1 - \hat{g}^B) \left((n^B > 0) p_d{}^{n^B} + (n^B = 0)(1 - p_a) \right)$$

$$P(n^A, n^B|n_j^M, n_j^F, g^M, g^F,) = \sum_{\substack{c_j^{AM} + c_j^{AF} = n^A \\ c_j^{BM} + c_j^{BF} = n^B}} P(c_j^{AM}, c_j^{BM}|n_j^M, g^M) P(c_j^{AF}, c_j^{BF}|n_j^F, g^F)$$

Subcasos maternos: para j en {X, XY}, tenemos

$$P(c_j^{AM}, c_j^{BM}|n_j^M, g^M) = (c_j^{AM} + c_j^{BM} = n_j^M) \begin{cases} (c_j^{BM} = 0), & g^M = AA \\ (c_j^{AM} = 0), & g^M = BB \\ \dfrac{1}{n_j^M + 1}, & g^M = AB \end{cases}$$

Para j=Y, que está degenerado para la madre, tenemos:

$$P(c_Y^{AM}, c_Y^{BM}|n_Y^M, g^M) = (c_Y^{AM} + c_Y^B = 0)(n_Y^M = 0)(g^M = NC)$$

Subcasos paternos: para j en {X,Y}, tenemos:

$$P(c_j^{AF}, c_j^{BF}|n_j^F, g^M) =$$

$$(c_j^{AF} + c_j^{BF} = n_j^F)((c_j^{AF} = 0) \cup (c_j^{BF} = 0)) \begin{cases} (c_j^{BF} = 0), & g^F = AA \\ (c_j^{AF} = 0), & g^F = BB \end{cases}$$

Para j = XY, las matemáticas son las mismas que para la madre, a saber:

$$P(c_{XY}^{AF}, c_{XY}^{BF} | n_{XY}^{F}, g^{F}) = (c_{XY}^{AF} + c_{XY}^{BF} = n_{XY}^{F}) \begin{cases} (c_{XY}^{BF} = 0), & g^{F} = AA \\ (c_{XY}^{AF} = 0), & g^{F} = BB \\ \frac{1}{n_{XY}^{F} + 1}, & g^{F} = AB \end{cases}$$

Técnica del cromosoma X

En una realización de la presente divulgación, la técnica del cromosoma X, descrita aquí, puede determinar el estado de ploidía del cromosoma X con un alto nivel de confianza. En la práctica, esta técnica tiene similitudes con la técnica de permutación, dado que la determinación se lleva a cabo examinando las curvas CDF características de los distintos contextos. Esta técnica utiliza específicamente la distancia entre ciertas curvas CDF de contexto para determinar el número de copias del cromosoma sexual.

En una realización de la presente divulgación, el algoritmo puede ser modificado del siguiente modo para optimizarlo para el cromosoma X. En esta realización, se pueden introducir ligeras modificaciones en la distribución de alelos, el modelo de respuesta y posibles hipótesis. La fórmula es:

$$P(g_{ij}|D,F) = \frac{P(D_{ij}^{s}|g_{ij},F_j^{s})}{P(D|F)} \sum_{g^M,g^F} P(g^M)P(g^F)P(D_i^m|g^M)P(D_i^f|g^F) \sum_h P(g_{ij}|g^M,g^F,h,F_j^s) * Q(h,g^M,g^F,F,D,i,j)$$

donde

$$Q(h,g^M,g^f,F,D,i,j) = \sum_{\substack{H_i \\ H_{ij}^s = h}} \prod_{\substack{a=1,\ldots,k, \\ a \neq j}} P(D_{ia}^s|g^M,g^F,H_{ia}^s,F_a^s) \prod_{a=1,\ldots,l} P(D_{ia}^s|g^F,H_{ia}^s,F_a^s)$$

$$* W_1(H_i,D,i,F) * W_2(H_i,D,i,F)$$

Además, pueden hacerse algunos de los siguientes cambios:

- El modelo de respuesta $P(D_{ij}^s|g_{ij},F_j^s)$ depende de F_j^s. Si $F_j^s = 0$, 2 copias, esto puede modelarse como antes, si $F_j^s = 1$, se utiliza una copia y puede ser modelado del mismo modo que para el esperma.
- $P(g^F)$ es p, (1-p), para AA, BB respectivamente, omitiendo AB.
- $P(D_i^f|g^F)$ es igual que antes, dado que se supone progenitores 100% correctos, hay que asegurarse de omitir todo recorte con $D_i^f = AB$
- *h,* la hipótesis del embrión en (madre, padre), tenía previamente 4 posibilidades, ahora solo considera 2 posibilidades para M1, M2, dado que la contribución del padre no existe (para $F_j^s = 0$), o solo tiene una hipótesis (para $F_j^s = 1$). Esto es válido para cada embrión. De forma similar sobre el esperma hay solamente una hipótesis.
- $P(g_{ij}|g^M,g^F,h,F_j^s)$ puede calcularse de forma ligeramente distinta dependiendo de F_j^s, es decir, dependiendo de si consideramos la contribución del padre.

- $Q(h, g^M, g^F, F, D, i, j)$ puede calcularse del mismo modo que antes, teniendo en cuenta la reducción en el espacio de hipótesis, y los cambios mencionados más arriba dependiendo de F_j^e.

Distancia de contexto: cromosoma X

En otra realización de la presente divulgación, se puede determinar el estado de ploidía del cromosoma X como sigue. El primer paso es determinar la distancia entre los cuatro contextos siguientes: AA|BB y BB|AA en el canal X, AA|BB y BB|AA en el canal Y, AB|BB y BB|AA en el canal X, y AB|AA y AA|BB en el canal Y. Estas distancias pueden ser tomadas en el punto donde AA|AA y BB|BB están más alejados, y luego se normalizan por la distancia entre AA|AA y BB|BB. Esta normalización es una forma de eliminar cualquier variación en el proceso de amplificación. Entonces pueden crearse distribuciones para cada una de las distancias normalizadas bajo las hipótesis H10, H01, H11, H21 y H12 utilizando determinaciones de ploidía de alta confianza en los cromosomas autosómicos. En una realización de la presente divulgación, el conjunto de prueba está restringido a cromosomas 1-15

La Figura 6 y la Figura 7 presentan dos gráficos mostrando el agrupamiento de los diversos contextos tomados de datos reales. La Figura 6 muestra una representación gráfica de un primer conjunto de SNPs, con la intensidad normalizada de una salida de canal representada contra la otra. La Figura 7 muestra una representación gráfica de un segundo conjunto de SNPs, con la intensidad normalizada de la salida de un canal representada contra la otra. Los datos presentados en esas dos figuras muestran que los datos de los diversos contextos se agrupan bien, y las hipótesis son claramente separables. Adviértase que se utilizaron solamente cromosomas con una confianza superior a aproximadamente 0,9 para el conjunto de prácticas. Un ejemplo de la distribución de las distancias puede verse en las Figuras 8A-8C, que muestran ajustes de curva para datos alélicos de distintas hipótesis de ploidía. La Figura 8A muestra ajustes de curva para datos alélicos para cinco hipótesis de ploidía distintas utilizando el método Kernel comentado aquí, la Figura 8B muestra ajustes de curva para datos alélicos para cinco hipótesis de ploidía distintas utilizando un Gaussian Fit comentado aquí, y la Figura 8C muestra un histograma de los datos alélicos reales medidos de un contexto parental, AA|BB - BB|AA, en el canal X, comparado con los ajustes de curva de todos los datos. El estado de ploidía cuya hipótesis coincide mejor con los datos alélicos reales medidos se determina como el estado de ploidía real. Esta técnica determina el estado de ploidía de la célula cuyos datos se muestran en las Figuras 6 - 8 como XX con una confianza de 0,999 o mejor. Este método efectúa también determinaciones correctas en células individuales aisladas de líneas celulares con estados de ploidía conocidos.

Cromosoma Y

En una realización de la presente divulgación, el estado de ploidía del cromosoma Y puede ser determinado como se describe en otra parte de esta divulgación, con las siguientes modificaciones. En una realización es posible utilizar la técnica de la presencia parental, con modificaciones apropiadas para el cromosoma Y.

Tomemos $F^e j = 0$, $g_{ij} = NaN$. Para $F^e_i = 1$, $g_{ij} = g^F$, es decir, lo mismo que el padre. En otra realización, es posible tener en cuenta posibles errores en la medición paterna:

$$P(g_{ij}|D,F) = P(g_{ij})P(D_i^f|g_{ij}) \prod_{a=1,\ldots,k_c} P(D_{ia}^c|g_{ij},F_a^c) \prod_{a=1,\ldots,l} P(D_{ia}^z|g_{ij},F_a^z)$$

donde $P(g_{ij})$ es la frecuencia de población en este recorte, $P(D_i^f|g_{ij})$ va a ser 0/1. En una realización de la presente divulgación, se puede suponer que no hay error parental, en cuyo caso el algoritmo del cromosoma Y es simple. En otra realización, se puede utilizar un modelo de error para los padres en el cromosoma Y, en cuyo caso $P(D_{ia}^z|g_{ij},F_a^z)$, que es simple si F_a=0, o se puede usar un modelo de error en el objetivo, y en el cromosoma Y.

Cromosoma XY

Para el cromosoma “XY”, es posible utilizar el mismo algoritmo que para otros cromosomas autosómicos.

Cromosoma Z

En una realización, el cromosoma “Z” se ha definido de forma que los alelos deben ser AB para machos y AA/BB para hembras, determinado por frecuencia de población. En esta realización se pueden realizar las siguientes modificaciones:

$$g_{ij} = \begin{cases} AB & F_j^s = 1 \\ AA & F_j^s = 0, p(A) = 1 \\ BB & F_j^s = 0, p(A) = 0 \end{cases}$$

En otros aspectos la determinación del estado de ploidía del cromosoma Z puede realizarse como se describe en otra parte en esta divulgación.

Técnica no paramétrica

En otra realización de la presente divulgación, puede utilizarse un enfoque denominado “técnica no paramétrica”. Esta técnica no establece supuestos sobre la distribución de los datos. Para un conjunto determinado de SNPs, definido típicamente por un contexto parental, crea la distribución esperada sobre bases hipotéticas o empíricas. La determinación de las probabilidades de las hipótesis se hace comparando la relación entre las distribuciones observadas de los contextos parentales con las relaciones esperadas entre las distribuciones de los contextos parentales. En una realización, las medias, cuartiles o quintiles de las distribuciones observadas pueden utilizarse para representar matemáticamente las distribuciones. En una realización, las relaciones esperadas pueden predecirse utilizando simulaciones teóricas, o pueden predecirse viendo los datos empíricos de conjuntos conocidos de relaciones en cromosomas con estados de ploidía conocidos. En una realización, las distribuciones teóricas para un contexto parental dado pueden ser construidas mezclando las distribuciones observadas de otros contextos parentales. Las distribuciones esperadas de contextos parentales bajo hipótesis distintas pueden ser comparadas con las distribuciones observadas de contextos parentales, y se espera que solo la distribución bajo las hipótesis correctas coincida con la distribución observada.

Se esboza en esta sección un método para calcular posteriores probabilidades como $P(H_i|\text{“data”})$ donde H_i es una hipótesis que es una cierta combinación de los conjuntos de distribuciones esperados para los casos donde un progenitor aporta 0, 1, o 2 cromosomas. En los casos en los que el progenitor aporta dos cromosomas, hay dos posibles subcasos: error de copia Ml (error de copia no emparejada) (2a), o error de copia M2 (error de copia emparejada) (2b). Esto origina 16 hipótesis totales: cuatro hipótesis para el padre, multiplicadas por cuatro para la madre. El caso en que la madre o el padre aportan por lo menos un cromosoma se discutirá primero, y el caso en el que un progenitor no aporta cromosomas se discutirá después. Consideremos los siguientes puntos:

(A) Bajo los contextos parentales AB|AA y AA|AB, bajo las 8 hipótesis de contribución cromosómica parental, donde cada progenitor aporta por lo menos un cromosoma, pero sin incluir el caso en que ambos progenitores aportaron dos cromosomas debido a un error de copia M2, la distribución de los genotipos objetivo puede separarse en una distribución que puede calcularse empíricamente partiendo de los datos. Además, la distribución del estado euploide puede separarse de las otras hipótesis.

(B) Si las distribuciones de los objetivos son distintas, existe una estadística *T* (formalmente aquí una variable aleatoria) que las distingue. La distribución de esta estadística puede ser simulada por *bootstrapping* de la distribución del objetivo bajo los contextos parentales AB|AA y AA|AB. Esto produce un valor p empírico bajo cada hipótesis. El valor p empírico bajo la hipótesis i^{th} se marcará y se define como

$$\hat{p}_i = P(T \geq t | \quad hypothesis \quad i) \tag{1}$$

donde *T* es la variable aleatoria, y vemos una realización de la estadística *t*. La distribución de *T* bajo la hipótesis *i* puede ser simulada con el *bootstrap*.

Valores p empíricos producirán posteriores distribuciones de $P(H_i|\text{“data”})$ vía formalización de “data” como el evento (una variable aleatoria) $\mathbb{1}_{T \geq t}$ con *T* definido en el espacio de probabilidad conjunta incluyendo todas las hipótesis y sus subhipótesis. Esto hace la anterior ecuación equivalente a $P(H_i|\mathbb{1}_{T \geq t})$ que por Bayes da

$$P(H_i|1_{T\geq t}) = P(1_{T\geq t}|H_i)\frac{P(H_i)}{P(1_{T\geq t})}$$

$$= \hat{p}_i \frac{p_{H_i}}{P(T\geq t)}$$

donde $\hat{p}_i$ como en la Ecuación, así $P(1_{T\geq t}) = \sum_i \hat{p}_i p_{H_i}$ y p_{H_i} es la previa en la hipótesis i.

Indiquemos (1,2a) como el caso donde la madre aporta 1 cromosoma y el padre aporta 2 con un error de copia M1. A los efectos de este debate, supongamos que un error de copia *m1* en un locus heterocigoto implica que AA, AB, y BB se producen cada uno con una probabilidad de 1/3. En un error de copia M2, un cromosoma está duplicado, por lo que para un locus heterocigoto, supongamos que AA y BB aparecen cada uno con una probabilidad de *1/2*.

El punto (A) puede mostrarse investigando la distribución del objetivo bajo las distintas hipótesis. Adviértase que (1,1) es el único caso donde $F_1 = F_2$ y ambas son mezclas de dos distribuciones distintas, que pueden ser simuladas utilizando SNPs homocigotos polares y no polares. Esta es una buena técnica para identificar la trisomía, pero resulta difícil calcular una confianza, porque es difícil simular su distribución. Por ejemplo, consideremos la mediana estadística $T = median_{AABB}\{x_i^X - x_i^Y\} - median_{BBAA}\{w_i^X - w_i^Y\}$, que es buena algorítmicamente separando (1,1) de (2a/b,1) o (1,2a/b). Nuevamente, no hay una confianza asociada, porque su distribución bajo la hipótesis de (1,2a/b) se simula del mismo modo que (1,1), es decir, si hay n_1 casos de AA|BB y n_2 casos de BB|AA, la distribución simulada es una distribución mezcla de AA|BB y BB|AA remuestreado con proporciones $n_1/(n_1 + n_2)$ y $n_2/(n_1 + n_2)$. Así, se esperará que *T* comparado con su distribución simulada bajo trisomía sea igual que *T* comparado con su distribución simulada bajo euploide. La siguiente explicación describe cómo solucionar este problema, con la improbable excepción del caso donde cada progenitor aporta dos copias de un cromosoma determinado al embrión.

En esta explicación, F_1 denota la distribución de los loci objetivo bajo el contexto parental AB|AA y F_2 la distribución de los loci objetivo bajo el contexto parental AA|AB.

1. (1,1): las distribuciones $F_1 = F_2$ y F_1 son una mezcla de ½ AA y ½ AB
2. (2b, 1): F1 es una mezcla de 3/2 AAA y 3/2 BBA. F2 es una mezcla de 3/2 AAA y 3/2 AAB.
3. (2a, 1): F1 es una mezcla de AAA ABA y BBA. Supondremos que la mezcla es 3/3 para cada uno, aunque puede no ser necesario para el método. F_2 es igual a una mezcla de 3/2 AAA y 7 AAB.
4. (1,2b) F1 es igual que F2 en el punto 2 por simetría y F2 es igual que F1 en el punto 2 por simetría.
5. (1,2a) F1 es igual que F2 en el punto 3 por simetría y F2 es igual que F3 en el punto 3 por simetría.
6. (2a, 2b) F1 es una mezcla de 3/3 cada uno de AAAA ABAA BBAA, F2 es una mezcla de 3/2 de AAAA AABB.
7. (2b, 2a) F1 es F2 del punto anterior y F2 es F1 del punto anterior por simetría.
8. (2a, 2a) F1 es una mezcla de 3/3 cada uno de AAAA ABAA BBAA, F_2 es igual que F1.
9. (2b, 2b) F1 es una mezcla de 3/3 AAAA, 3/2 BBAA. F2 tiene la misma distribución que F1.
10.

El enfoque algorítmico es como sigue:

• Hallar una buena F1 estadística de canales objetivo bajo cada contexto parental AA|AB, y una buena F2 estadística de canales objetivo bajo el contexto parental AB|AA. En una realización, tomemos t1 y t2 como las medias de $x_i^X - x_i^Y$ bajo AA|AB y AB|AA, respectivamente.

- Bajo la hipótesis *i,* producir distribuciones nulas conjuntas empíricas $(\hat{F}_1, \hat{F}_2)$ utilizando una mezcla de datos remuestreados de homocigotos polares cuando sea posible, generalmente es posible: de lo contrario utilizar remuestreado de heterocigotos.
- Comparar la distribución conjunta de (t1,t2) con la empírica, lo que da el valor p empírico.
- Calcular el valor p empírico como se describe en la primera parte del documento.
- Clasificar de acuerdo con la probabilidad posterior máxima y asignar la probabilidad posterior a la determinación.
- Para aumentar la potencia de este procedimiento, se pueden incluir distribuciones F_3, F_4 que corresponden a F1 y F2 pero intercambian los alelos A y B.

Consideremos ahora los casos en los que un progenitor no aporta cromosomas:

1. (0,0): F1 y F2 son ruido, pueden ser simuladas utilizando cualquier SNPs. En una realización, se podría utilizar el contexto AA|AA y BB|BB.
2. (0,1): F1 es una mezcla ½ de A y B, F2 es A
3. (0, 2a): F1 es AA y F2 es BB.
4. (0, 2b): F1 es AA y F2 es una mezcla de AA AB BB.
5. (1,0) sustituir F1 y F2 del caso de (0,1) por simetría.
6. (2a, 0) sustituir F1 y F2 del caso de (0,2a) por simetría.
7. (2b, 0) sustituir F1 y F2 del caso de (0,2b) por simetría.

Esbozo de confianza para la técnica no paramétrica

El análisis del algoritmo se basa en la idea de que para la hipótesis ith, Hi, se puede calcular la probabilidad de que alguna hipótesis Hj (otra o la misma) sea cierta con los datos P (Hi|data), lo que equivale a P ("determinaciones de algoritmo" Hi|data).

Utilizando los previos, se puede calcular P(data|Hi). En una realización, el algoritmo puede simplificarse utilizando el contexto parental 1. En otra realización, se pueden utilizar los tres contextos. Por consiguiente, se puede escribir el análisis del algoritmo que determina euploide cuando $\frac{|\hat{p}_q - q|}{\hat{\sigma}_{\hat{p}_q}}$ es menor que un umbral *t* donde $\hat{p}_q$ es la reestimación de *q* utilizando solamente el contexto parental 1 que son los homocigotos polares. Adviértase también que el algoritmo determina el estado de ploidía en base a un esquema de umbral modificado, donde el reestimado $\hat{p}_q$ es comparado con *q* y normalizado en base al error estándar estimado de $\hat{\sigma}_{\hat{p}_q}$. El algoritmo funciona en autosomas y cromosomas sexuales de este modo.

Determinar un contexto particular y suponer que Zi y Wj tienen la siguiente distribución:

$$Z_i = \mu_Z + \sigma_i^2 \quad \varepsilon_i \text{ y}$$

$$W_j = \mu_W + \sigma_j^2 \quad \varepsilon_j$$

donde el ε_i y ε_j se suponen I.I.D., y $\{\sigma_i\}_{i=1}^{n}$ son constantes. En la práctica, $z_1, \ldots z_{n_Z}$ y $w_1, \ldots w_{n_W}$ se observan realizaciones de las variables aleatorias $\{Z_i\}_{i=1}^{n_Z}$ and $\{W_i\}_{i=1}^{n_W}$.

Para analizar el algoritmo determinando el cuantil, suponer que el cuantil qth de ε es igual a 0. Esto es sin pérdida de generalidad porque, por ejemplo, la determinación del cuantil es invariable bajo escalamiento multiplicativo de Zi y Wj y añadiendo una constante a todos los Zi y Wj.

Suponer que todos σ_i^2 son iguales para simplificar, y tomemos z_q como el cuantil qth del Zi. Definir/ denotar el p_q por $p_q := P(W_j < z_q)$.

Luego, bajo la condición de euploide, dado que $\mu_Z = \mu_W$, para cada ε_i,

$$p_q = P(\mu_W + \varepsilon_i < \mu_Z) = q.$$

Donde el $P(\mu_W + \varepsilon_i < \mu_Z) = E(1_{(\mu_W + \varepsilon_i < \mu_Z)})$

Esquema de cálculos de probabilidad

Para entender la idea en general, consideremos un caso simplificado: supongamos que los σ_j son todos iguales y zq se conocen con exactitud. Entonces, el estimador de P_q marcado $\hat{p}_q$ que en general es $\hat{p}_q = \frac{1}{n_W}\sum_{i=1}^{n_W} 1_{W_i \le z_q}$, se simplificaría a $\hat{p}_q = \frac{1}{n_W}\sum_{i=1}^{n_W} 1_{W_i \le z_q}$.

En este caso, W_i son i.i.d., z_q es conocido y en consecuencia $\hat{p}_q$ es simplemente una media de I.I.D.Bernouillis. Este es un estimador más sencillo. El teorema límite central, que puede ser utilizado para obtener información exacta sobre la calidad de la aproximación, dice que $\frac{\hat{p}_q - p_q}{\sigma_{p_q}}$ (2) tiene una distribución normal aproximada.

Este método puede utilizarse para obtener confianzas, porque en euploidía, $p_q = q$, y en aneuploidía, si se supone que en aneuploidía tipo j[th] hay una diferencia δ_j entre Pq y q, (j = 1 significa contribuciones parentales (0,0), j = 1 significa contribuciones parentales (1,0), ...), $p_a - q > \delta_j$. En una realización, el cálculo de δ puede ser entre 0 y 0,5.

Supongamos ahora para simplificar, que todas las hipótesis se colapsan en H_g, la hipótesis de euploidía, y H_a la hipótesis de aneuploidía y señalan δ como la menor δ_j.

Definir $\hat{z} := \frac{\hat{p}_q - q}{\hat{\sigma}_{p_q}}$ (3)

Donde $\hat{\sigma}_{p_q}$ es una estimación de σ_{p_q}, por “bootstrap”, o por la fórmula de varianza de Bernouilli.

El algoritmo establece algún umbral t y determina H_g iff $|\hat{z}| < t$. Por consiguiente, en euploidía, utilizando la aproximación normal, $\hat{z}$ tiene una distribución normal estándar aproximada, así $P(H_g \text{ called} | \text{euploid condition}) = P(|\hat{z}| < t) \cong P(|N(0,1)| < t) \cong .99 \text{ for } t = 3$

Para t = 3, esta probabilidad es aproximadamente de ,99. Por consiguiente:

$P(H_a \text{ determinado} \mid \text{condición euploide}) \cong .01.$

A la inversa, en aneuploidía, $\hat{z}$ tiene una distribución normal con una media $\frac{\delta}{\hat{\sigma}_{p_q}}$ y una varianza de 1. Típicamente, σ_{p_q} está en el rango de 0,01, por tanto, de δ = (.01)c para una constante c. En algunas realizaciones c puede estar entre aproximadamente 1 y 10, y en otra realización, c puede estar entre aproximadamente 10 y 100.

$P(H_a \text{ called} \mid \text{aneuploid condition}) = P(|\hat{z}| < t) \cong P(|N(5,1)| < t)$ es pequeña. Para t = 3, esta probabilidad es aproximadamente de (1 - ,98)/2. Por tanto, $P(H_a \text{ called} \mid \text{aneuploid condition}) \cong 1 - .01.$

Otras posibles técnicas especializadas pueden utilizarse en el contexto de la determinación de ploidía, y la lista que se presenta en esta divulgación no pretende ser exhaustiva. Más abajo se indican otras técnicas.

Determinación de alelos

En el contexto del PGD durante la IVF, es muy necesario determinar el genoma del embrión. No obstante, genotipando una sola célula el resultado es frecuentemente una elevada tasa de pérdida de alelos, donde muchos alelos dan una lectura incorrecta o ninguna lectura. Se requieren unos datos genéticos exactos del embrión para

detectar genes vinculados a enfermedad con un alto nivel de confianza, y esas determinaciones pueden ser utilizadas entonces para seleccionar el mejor embrión para su implantación. Una realización de la presente divulgación, descrita aquí, implica inferir los datos genéticos de un embrión con la mayor exactitud posible. Los datos obtenidos pueden incluir los datos genéticos medidos, en el mismo conjunto de *n* SNPs, de un individuo objetivo, el padre del individuo y la madre del individuo. En una realización, el individuo objetivo puede ser un embrión. En una realización, se utilizan también los datos genéticos medidos de uno o más espermas del padre. En una realización, se utilizan también los datos genéticos medidos de uno o más hermanos del individuo objetivo. En una realización, el o los hermanos pueden ser considerados también individuos objetivo. Se describe aquí una forma de incrementar la fidelidad de las determinaciones de alelos en los datos genéticos de un individuo objetivo a los efectos de hacer predicciones clínicamente accionables. Adviértase que el método puede ser modificado para optimizarlo para otros contextos, como cuando el individuo objetivo no es un embrión, donde solo están disponibles datos genéticos de un progenitor, cuando se conocen ninguno, uno o ambos haplotipos parentales, o cuando se conocen datos genéticos de otros individuos relacionados y pueden ser incorporados.

Las presentes divulgaciones descritas en esta y otras secciones de este documento tienen el objetivo de incrementar la exactitud de la determinación de alelos en alelos de interés para un número determinado de SNPs, o alternativamente, reducir el número de SNPs necesarios, y con ello el coste, para alcanzar un nivel medio determinado de precisión en las determinaciones de SNP. A partir de esas determinaciones de alelos, en especial las de genes vinculados a enfermedad u otros fenotipos, se pueden establecer predicciones en cuanto a potenciales fenotipos. Esta información puede ser utilizada para seleccionar un embrión o embriones con cualidades deseables para su implantación. Dado que el PGD resulta muy caro, toda nueva tecnología o mejoría en los algoritmos PS, que permita conseguir un cálculo del genotipo objetivo con un nivel de precisión determinado, con menos potencia de cálculo, o menos SNPs medidos, significará una importante mejoría sobre la tecnología anterior.

Esta divulgación presenta varios métodos nuevos en los que se emplean datos genéticos medidos parentales y objetivo, y en algunos casos datos genéticos de hermanos, para determinar alelos con un alto grado de precisión, donde los datos de hermanos pueden proceder de hermanos nacidos, u otros blastómeros, y donde el objetivo es una sola célula. El método divulgado muestra la reducción a la práctica, por primera vez, de un método que puede aceptar, como entrada, datos genéticos no limpios medidos de varios individuos relacionados, y también determinar el estado genético más probable de cada uno de los individuos relacionados. En una realización, esto puede significar determinar la identidad de diversos alelos, así como ajustar por fases datos no ordenados, teniendo en cuenta cruces, y también el hecho de que todos los datos entrados pueden contener errores.

Los datos genéticos de un objetivo pueden describirse considerando los datos genéticos medidos del objetivo, y de los progenitores del objetivo, donde se supone que los datos genéticos de los progenitores son correctos. No obstante, es probable que todos los datos genéticos medidos contengan errores, y toda suposición a priori es probable que introduzca sesgos e inexactitudes en los datos. El método que se describe aquí muestra cómo determinar el estado genético más probable de un conjunto de individuos relacionados, donde se supone que ninguno de los datos genéticos es cierto. El método que se divulga aquí permite que la identidad de cada pieza de datos genéticos medidos esté influida por los datos genéticos medidos de cada uno de los otros individuos relacionados. Así, datos parentales medidos incorrectamente pueden ser corregidos si la evidencia estadística indica que es incorrecto.

En los casos en que los datos genéticos de un individuo, o de un conjunto de individuos relacionados, contienen una cantidad significativa de ruido, o errores, El método que se divulga aquí utiliza las similitudes esperadas entre datos genéticos de esos individuos relacionados, y la información contenida en los datos genéticos, para limpiar el ruido en el genoma objetivo, junto con errores que pueda haber en los datos genéticos de los individuos relacionados. Eso se hace determinando qué segmentos de cromosomas intervinieron en la formación de gametos, y dónde se produjeron cruces durante la meiosis, y por consiguiente qué segmentos de los genomas de individuos relacionados se espera que sean casi idénticos a secciones del genoma objetivo.

En determinadas situaciones este método puede ser utilizado para limpiar mediciones de pares de base ruidosos, pero también puede ser utilizado para deducir la identidad de pares de bases individuales o regiones completas de ADN que no se habían medido. En una realización, se pueden utilizar datos genéticos no ordenados como entrada, para el individuo objetivo, y/o para uno o más de los individuos relacionados, y los resultados contendrán los datos genéticos ajustados por fases y limpios de todos los individuos. Además, se puede calcular una confianza para cada determinación de reconstrucción realizada. En otra parte de esta divulgación pueden encontrarse otros debates sobre la creación de hipótesis, calcular las probabilidades de las diversas hipótesis, y usar esos cálculos para determinar el estado genético más probable del individuo.

Se presenta primero una explicación muy simplificada de determinación de alelos, haciendo suposiciones no realistas para ilustrar el concepto de la presente divulgación. Se presenta después un enfoque estadístico detallado que puede ser aplicado a la tecnología de hoy.

Un ejemplo simplificado

La Figura 9 ilustra el proceso de recombinación que se produce durante la meiosis para la formación de gametos en un progenitor. El cromosoma 101 de la madre del individuo se muestra en gris. El cromosoma 102 del padre del individuo se muestra en blanco. Durante este intervalo, conocido como Diploteno, durante la Profase I de la Meiosis,

es visible una tétrada de cuatro cromátidas 103. Se producen cruces entre cromátidas no hermanas de un par homólogo en los puntos conocidos como nódulos de recombinación 104. A efectos de ilustración, el ejemplo se centra en un solo cromosoma, y tres SNPs, de los que se supone que caracterizan a los alelos de tres genes. Para esa discusión se supone que los SNPs pueden ser medidos por separado en los cromosomas maternos y paternos. Este concepto puede aplicarse a muchos SNPs, muchos alelos caracterizados por múltiples SNPs, muchos cromosomas y a la tecnología de genotipado actual, donde los cromosomas maternos y paternos no pueden ser aislados individualmente antes del genotipado.

Debe prestarse atención a los puntos de potencial cruce entre los SNPs de interés. El conjunto de alelos de los tres genes maternos puede describirse como ($a_{m1,}$ $a_{m2,}$ a_{m8}) correspondiendo a SNPs (SNP1, SNP2, SNP3). El conjunto de alelos de los tres genes paternos puede ser descrito como ($a_{p1;}$ a_{p2}, a_{p3}). Consideremos los nódulos de recombinación formados en la Figura 1, y supongamos que hay solamente una recombinación para cada par de cromátidas recombinantes. El conjunto de gametos que se forman en este proceso tendrán alelos genéticos: (a_{m1}, a_{m2}, a_{p3}), (a_{m1}, a_{p2}, a_{p3}), (a_{p1}, a_{m2}, a_{m3}), (a_{p1}, a_{p2}, a_{m3}). En el caso sin cruces de cromátidas, los gametos tendrán alelos (a_{m1}, a_{m2}, a_{m3}), (a_{p1}, a_{p2}, a_{p3}). En el caso con dos puntos de cruce en las regiones pertinentes, los gametos tendrán alelos (a_{m1}, a_{p2}, a_{m3}), (a_{p1}, a_{m2}, a_{p3}). Estas ocho distintas combinaciones de alelos se referirán como el conjunto de alelos hipótesis, para ese progenitor concreto.

La medición de los alelos del ADN embriónico es típicamente ruidosa. A efectos de este debate, tomemos un solo cromosoma del ADN embriónico, y supongamos que procede del progenitor cuya meiosis se ilustra en la Figura 9. Las mediciones de los alelos en este cromosoma pueden ser descritas en términos de un vector de variables indicadoras: $A = [A1\ A2\ A3]^T$ donde A1 = 1 si el alelo medido en el cromosoma embriónico es a_{m1}, A1 = -1 si el alelo medido en el cromosoma embriónico es a_{p1}, y A1 = 0 si el alelo medido no es a_{m1} ni a_{p1}. En base al conjunto de alelos hipótesis del progenitor supuesto, se puede crear un conjunto de ocho vectores que correspondan a todos los posibles gametos descritos más arriba. Para los alelos descritos más arriba, esos vectores serían $a1 = [1\,1\,1]^T$, $a_2 = [1\ 1\ -1]^T$ $a_3 = [1\text{-}1\,1]^T$ $a_4=[1\text{-}1\ -1]^T$ $a_5 = [\text{-}1\,1\,1]^T$ $a_6 = [\text{-}11\text{-}1]^T$ $a7 = [-1\ -1\ 1]^T$ $as = [-1\ -1\ -1]^T$. En esta altamente simplificada aplicación del sistema, los probables alelos del embrión pueden ser determinados realizando un simple análisis correlacional entre el conjunto hipótesis y los vectores medidos:

$$i^* = \arg\max_i A^T a_i,\ i = 1...8$$

Una vez hallado i*, se selecciona la hipótesis a_i* como el conjunto de alelos más probable en el ADN embriónico. Este proceso puede repetirse dos veces, con dos distintas suposiciones, a saber, que el cromosoma embriónico procede de la madre o del padre. Esta suposición que da la mayor correlación $A^T a_1$* se supondría correcta. En cada caso se utiliza un de conjunto de alelos hipótesis, basado en las mediciones del ADN respectivo de la madre o del padre.

Hay que advertir que en una realización, los SNPs que eran importantes debido a su asociación con fenotipos de enfermedad determinados, pueden ser denominados SNPs asociados a Fenotipo o PSNPs. En esta realización, se puede medir un gran número de SNPs entre los PSNPs, denominados SNPs no asociados a fenotipo (NSNPs), que se eligen a priori (por ejemplo, para desarrollar una matriz de genotipado especializada) seleccionando de la base de datos NCBI dbSNP esos RefSNPs que tienden a diferir sustancialmente entre individuos. Alternativamente, los NSNPs entre los PSNPs pueden ser elegidos para un par particular de progenitores porque los alelos de los progenitores son distintos. El uso de los SNPs adicionales entre los PSNPs permite determinar con un mayor nivel de confianza si se producen cruces entre los PSNPs. Es importante tener en cuenta que, aunque en esta anotación se hace referencia a “alelos” distintos, es simplemente por comodidad; los SNPs pueden no estar asociados a genes que codifican proteínas.

Un tratamiento más completo del método de determinación de alelos

En el ejemplo simplificado que se da más arriba, a los efectos de ilustración del concepto, se parte del supuesto de que los genotipos parentales están ajustados por fases y son perfectamente conocidos. No obstante, en muchos casos, este supuesto puede no sostenerse. Por ejemplo, en el contexto de genotipado de embriones durante la IVF, típicamente los datos genéticos medidos de los padres no están limpios ni ajustados por fases, todos los datos genéticos medidos de esperma del padre son sucios, y los datos genéticos medidos de uno o más blastómeros, biopsiados de uno o más embriones, son también sucios y no ajustados por fases. En teoría, el conocimiento de los datos genéticos sucios, no ajustados por fases derivados del embrión puede ser utilizado para ajustar por fases y limpiar los datos genéticos parentales. Además, en teoría el conocimiento del genotipo de un embrión puede utilizarse para ayudar a limpiar y ajustar por fases los datos genéticos de otro embrión. En algunos casos, los datos genéticos medidos de varios individuos objetivo hermanos pueden ser correctos para un conjunto determinado de alelos, mientras que los datos genéticos de un progenitor pueden ser incorrectos en esos mismos alelos. En teoría, el conocimiento de los individuos objetivo puede ser utilizado para limpiar los datos del progenitor.

En algunas realizaciones de la presente divulgación presentada aquí, se describen métodos que permiten limpiar y ajustar por fases los datos genéticos parentales utilizando el conocimiento de los datos genéticos del objetivo y otros individuos relacionados. En algunas realizaciones, se describen métodos que permiten también limpiar y ajustar por fases los datos genéticos utilizando el conocimiento de los datos genéticos de hermanos. En una realización de la presente divulgación, los datos genéticos de los padres, del individuo objetivo, y de uno o varios individuos relacionados, se utilizan como entrada, donde cada parte de los datos genéticos va asociada a una confianza, y el

conocimiento de las similitudes esperadas entre todos los genotipos es utilizado por un algoritmo que selecciona el estado genético más probable de todos individuos relacionados inmediatamente. El resultado de este algoritmo, el estado genético más probable de los individuos relacionados, puede incluir los datos de la determinación de alelos genéticos limpios y ajustados por fase. En algunas realizaciones de la presente divulgación, puede haber diversos individuos objetivo, y esos individuos objetivo pueden ser embriones hermanos. En algunas realizaciones de la presente divulgación, los métodos divulgados en la siguiente sección pueden ser utilizados para determinar la probabilidad estadística para una hipótesis alélica, con los apropiados datos genéticos.

En algunas realizaciones de la presente divulgación, la célula objetivo es un blastómero biopsiado de un embrión en el contexto del diagnóstico preimplantación (PGD) durante la fertilización in vitro (IVF). En algunas realizaciones, la célula objetivo puede ser una célula fetal, o ADN fetal extracelular en el contexto del diagnóstico prenatal no invasivo. Adviértase que este método puede aplicarse igualmente bien a situaciones en otros contextos. En algunas realizaciones de la presente divulgación, un dispositivo de cálculo, como un ordenador, se utiliza para ejecutar todos los cálculos que implique el método. En una realización de la presente divulgación, el método presentado aquí utiliza datos genéticos del individuo objetivo, de los padres del individuo objetivo, y posiblemente de uno o más espermas, y una o más células de hermanos para recrear, con gran exactitud, los datos genómicos del embrión, teniendo en cuenta con precisión los cruces. En una realización de la presente divulgación, el método puede ser utilizado para recrear datos genéticos para individuos objetivo en cromosomas aneuploides y también euploides. En una realización de la presente divulgación, se describe un método para determinar los haplotipos de células parentales, con datos parentales diploides dados, y datos genéticos diploides de uno o más blastómeros u otras células de hermanos, y posible, pero no necesariamente, una o más células de esperma del padre.

Descripción práctica de determinación de alelos

En la siguiente sección se da una descripción de un método de determinación del estado genético de uno o una serie de individuos objetivo. La descripción se realiza en el contexto de la determinación del genotipo del embrión, en el contexto de un ciclo de IVF, pero es importante advertir que el método descrito aquí puede aplicarse igualmente bien en otros contextos, para otros conjuntos de individuos relacionados, por ejemplo, en el contexto del diagnóstico prenatal no invasivo, cuando el individuo objetivo es un feto.

En el contexto de un ciclo de IVF, para un cromosoma particular, los datos resultado de la técnica de genotipado para *n* localizaciones de SNP, para objetivos distintos *k* (embriones o niños) se obtienen por la técnica de genotipado. Cada uno de los objetivos puede tener genotipos medidos para una o más muestras, y las mediciones pueden hacerse en amplificaciones de una sola célula, o de un pequeño número de células. Para cada SNP, cada medición de muestra consiste en mediciones de respuesta de canal (X,Y) (intensidad). El canal X mide la potencia de un alelo (A), y el canal Y mide la potencia del otro alelo (B). Si las mediciones han sido completamente exactas, en un SNP concreto, un alelo que es AA debe tener intensidades normalizadas (X,Y) (se utilizan unidades arbitrarias) de (100,0), un alelo que es AB debe tener intensidades de (50,50), y un alelo que es BB debe tener intensidades de (0,100), y en este caso ideal, sería posible derivar valores de alelos exactos, dadas las intensidades de canal (X,Y). No obstante, las mediciones de célula única objetivo están típicamente lejos de ser ideales, y no es posible determinar, con un alto grado de confianza, el verdadero valor del alelo con las respuestas de canal en bruto.

La determinación de alelos puede hacerse para cada cromosoma por separado. Este debate se centra en un cromosoma autosómico concreto con *n* SNPs. El primer paso es definir la nomenclatura de los datos de entrada. Los datos de entrada para el algoritmo pueden ser los datos de salida no limpios ni ordenados de ensayos de matriz de genotipado, pueden ser datos secuenciados, pueden ser datos de genotipo procesados parcial o completamente, pueden ser datos de genotipo conocidos de un individuo, o puede ser cualquier tipo de datos genéticos. Los datos pueden ser dispuestos en datos objetivo, datos parentales y gametos de esperma, pero no es necesario. En el contexto de la IVF, los datos objetivo se referirían a los datos genéticos medidos de blastómeros biopsiados de embriones, y pueden referirse también a datos genéticos medidos de hermanos nacidos. Los datos de esperma podrían referirse a datos medidos de un conjunto individual de cromosomas derivados de un progenitor, incluyendo esperma, cuerpos polares, huevos no fertilizados u otra fuente de materia genética monosómica. Los datos se disponen aquí en varias categorías para mejor comprensión, pero no es necesario.

En esta divulgación, los datos de entrada van marcados como sigue: D se refiere a un conjunto de datos genéticos de un individuo. D^T = (D^{T1} ,...,D^{Tk}) se refiere a los datos genéticos de objetivos distintos k (embriones/niños), D^s =(D^{S1} ,...,D^{S1}) se refiere a los datos de espermas distintos (1), (D^m) se refiere a los datos de la madre, y (D^F) se refiere a los datos del padre. Se puede escribir D = (D^T , D^s, D^M, D^F). Escrito de otro modo, por SNPs, donde el subíndice i se refiere al SNP i^{th} en el conjunto de datos, D = (D1,.. .,D_n), donde Di = (D^T;, D^Si, D^Mi, D^Fi).

Para distintos objetivos *k*, se puede escribir D^Ti = (D^{T1}i,D^{T2}i,...,D^{Tk}i). Cada objetivo distinto puede tener múltiples remuestras; una remuestra se refiere a una lectura de genotipo adicional efectuada en una muestra determinada. Para el objetivo distinto j^{th} se puede escribir $D^{Tj,1}$ = ($D^{Tj,2}$,$D^{Tj,2}$j,...,$D^{Tj,kj}$j) donde kj = número de muestras para el objetivo j. Para la remuestra r^{th} del objetivo j en el SNP i, se observará el conjunto de intensidades de canal $D^{Tj,r}{}_i$=($X^{Tj,r}{}_i$, $Y^{Tj,r}{}_i$).

Pueden considerarse diversos espermas, y para el SNP i se puede escribir D^Si =(D^{S1}i ,D^{S2}i,...,D^{S1}i) para objetivos distintos (1). Cada esperma distinto puede tener también múltiples remuestras. Así, para el esperma distinto jth

$D^{Sj}i=(D^{Sj,1}i,D^{sj,2}$ i,...,$D^{Sj,1j}i)$ donde 1j=número de remuestras para el esperma j. Para la remuestra r^{th} del esperma j en el SNPi, se observará el conjunto de intensidades de canal $D^{Sj,r}i= (X^{Sj,r}i,Y^{Sj,r}i)$.

Los datos genéticos de la madre, en el SNP i, son D^{M}_{i} =($D^{M,1}i$,$D^{M,2}$i,...,$D^{M,a}$i). Los datos genéticos de la madre pueden tener también múltiples remuestras, y para la remuestra r^{th} de la madre en el SNP i, se observará el conjunto de intensidades de canal $D^{M,r}i =(X^{M,r}i,Y^{M,r}i)$.

Los datos genéticos del padre, en el SNP i, son D^{F}; =($D^{F,1}i$, $D^{F,2}$ i, . . .,$D^{F,b}$i) Los datos genéticos del padre pueden tener también múltiples remuestras, y para la remuestra r^{th} del padre en el SNP i, se observará el conjunto de intensidades de canal $D^{F,r}i =(X^{F,r}i,Y^{F,r}i)$.

Nomenclatura de hipótesis

Para el SNPi, y el objetivo j, la hipótesis consiste en la hipótesis de origen de la madre y del padre; es decir, $H^{Tj}i$ = ($H^{Tj}i$,m, $H^{Tj}_{i,f}$),donde $H^{Tj}_{i,m}$ en {1,2}, $H^{Tj}_{i,f}$ en {1,2}, cada uno de los cuales denota el haplotipo parental de origen para cada valor. Para el esperma, hay solamente una hipótesis de origen paterna; es decir, $H^{Sj}i$ in {1,2}, indicando el origen paterno (suponiendo un esperma normal).

En general, se puede escribir:

H = (H1…,Hn), donde Hi = ($H^{T}i$, $H^{S}i$) y $H^{T}i$ = ($H^{T1}i$,H^{T2}i,.. .,$H^{Tk}i$) y $H^{S}i$ = ($H^{s1}i$,H^{S2}i ,…$H^{S1}i$, donde $H^{Tj}i$= ($H^{Tj}_{i,m}$, $H^{Tj}_{i,f}$).

En un ejemplo con 3 embriones y 1 esperma, una hipótesis de SNP particular para un segmento de cromosoma podría ser ((M1,P_2),(M_2,P2),(M_2,P1),S1). Hay en total $2^{(2k+1)n}$ distintas hipótesis H.

Cálculo de la probabilidad del genotipo objetivo P(g\D)

Para SNP i, objetivo j, si se halla *P(g\D)*, el más probable $\vec{g}_i^j = argmax_g P(g|D),$ se selecciona como la determinación de alelos, con confianza $c_i^j = P\left(\vec{g}_i^j \middle| D\right).$. Para derivar P(g|D), primero se toma g^{M},g^{F} como posibles progenitores ordenados en i^{th} SNP; es decir, $g^M, g^F \in \{AA, AB, BA, BB\}.$Hi es la hipótesis completa de SNP i. Así:

Aquí la probabilidad se ha dividido por las probabilidades locales de datos sobre SNP i, $\left(D_i, g_i^j, H_i\right)$

y las probabilidades de los datos en todos los demás SNPs dependen solo de la hipótesis Hi:

$$P\left(D_{1,\ldots,i-1}|H_i\right), P\left(D_{i+1,\ldots,n}|H_i\right).$$

La probabilidad en SNP i

$$P\left(D_i, g_i^j, H_i\right) = \sum_{g^M, g^F} P\left(D_i, g_i^j, H_i, g^M, g^F\right) =$$
$$= \sum_{g^M, g^F} P\left(D_i|g_i^j, g^M, g^F, H_i\right) P\left(g_i^j|g^M, g^F, H_i^{Tj}\right) P\left(g^M\right) P\left(g^F\right) P\left(H_i\right)$$

$P\left(g^M\right), P\left(g^F\right)$ son frecuencias de alelos parentales ordenados en ese SNP. En particular, si en ese SNP P(A) = p, P(AA) = p^2, P(AB) = P(BA) = p(1-p), P(BB) = $(1-p)^2$. Las frecuencias de alelos de SNP pueden ser calculadas por separado de grandes muestras de datos genómicos.

$P(H_i)$ es generalmente igual para todas las hipótesis Hi, y en todos los SNPs, excepto que para uno de los SNPs (esto puede elegirse arbitrariamente; se puede elegir un SNP en el medio, digamos en SNP n/2), la hipótesis está restringida, y se puede determinar el primer objetivo (M1,F1) a efectos de singularidad.

$P\left(g_i^j|g^M, g^F, H_i^{Tj}\right)$ es 1 o 0, dependiendo de la coincidencia del valor del alelo g_i^j con uno obtenido por una combinación de g^M, g^F, H_i^{Tj}, es decir, si definimos $a(g^M, g^F, h)$ = (un valor de alelo definido de forma única por los alelos maternos ordenados g^M, un alelo paterno ordenado g^F, y la hipótesis parental h), luego:

$$P(g_i^j | g^M, g^F, H_i^{Tj}) = I\{g_i^j = \alpha(g^M, g^F, H_i^{Tj})\}$$

Ahora $P(D_i | g_i^j, g^M, g^F, H_i)$ es la probabilidad de los datos con valores de alelos particulares, puesto que considerando los progenitores g^M,g^F y la hipótesis Hi, los valores de alelos para todos los objetivos, espermas y progenitores se determinan de forma única. En particular puede reescribirse como:

$$P(D_i | g_i^j, g^M, g^F, H_i) = P(D_i^T | g_i^j, g^M, g^F, H_i^T) P(D_i^S | g^F, H_i^S) P(D_i^M | g^M) P(D_i^F | g^F)$$

Para objetivos:

$$P(D_i^T | g_i^j, g^M, g^F, H_i^T) = P(D_i^{Tj} | g_i^j) \prod_{u \neq j} P(D_i^{Tu} | \alpha(g^M, g^F, H_i^{Tu}))$$

Para cada objetivo u, $P(D_i^{Tu} | g)$ es el producto de las probabilidades de todas las remuestras de ese objetivo $P(D_i^{Tu} | g) = \prod_r P(D_i^{Tu,r} | g)$.

De forma similar para el esperma:

$$P(D_i^S | g^F, H_i^S) = \prod_u P(D_i^{Su} | \alpha(g^F, H_i^{Su}))$$

Para cada esperma *u*, $P(D_i^{Su} | g)$ es el producto de las probabilidades de todas las remuestras de ese esperma $P(D_i^{Su} | g) = \prod_r P(D_i^{Su,r} | g)$.

Para los progenitores, se pueden multiplicar las probabilidades de las remuestras para cada progenitor:

$$P(D_i^M | g^M) = \prod_r P(D_i^{M,r} | g^M), \; P(D_i^F | g^F) = \prod_r P(D_i^{F,r} | g^F)$$

La parte de la probabilidad P(D|g) que queda por debatir para cada muestra de objetivo, esperma y progenitor, es el modelo de respuesta de plataforma calculada para esa muestra. Esto se debatirá posteriormente.

Probabilidad en los SNPs 1,...,i1

Para H_{i-1} todas las posibles hipótesis en SNP i-1

$$P(D_{1,\ldots,i-1} | H_i) = \sum_{H_{i-1}} P(D_{1,\ldots,i-1} | H_{i-1}) P(H_{i-1} | H_i)$$
$$= \sum_{H_{i-2}} P(D_{1,\ldots,i-2} | H_{i-1}) P(D_{i-1} | H_{i-1}) P(H_{i-1} | H_i) \qquad P(D_{1,\ldots,i-2} | H_{i-1}),$$

tiene el mismo formato que $P(D_{1,\ldots,i-1} | H_i)$, y puede ser calculado secuencialmente a partir de SNP 1. En particular, definir la matriz W^i como $W^i(h,1) = P(D_{1,\ldots,i-1} | h)$ donde h es la hipótesis sobre SNP i. Definir la matriz PD^i como

$PD^{i-1}(g,1) = P(D_{i-1} | g)$ donde g es la hipótesis sobre SNP i-1. Definir la matriz PC^i como

$PC^i(h,g) = P(g|h)$, la probabilidad de transición entre las hipótesis g a h, yendo de SNP i-1 a i.

Entonces se puede decir $W^i = PC^i \times (PD^{i-1} \cdot W^{i-1})$ con la condición inicial $W^1(g) = P(start@g)$. Esto puede ser una constante elegida arbitrariamente.

Por tanto, primero hallar $W^2 = PC^2 X(PD^1 \cdot W^1)$, , luego W^3, y así sucesivamente hasta W^i.

$PC^i(H_i, H_{i-1}) = P(H_{i-1}|H_i)$ es la probabilidad de transición dependiendo de la probabilidad de cruce entre SNPs i-1, i. Es importante recordar que la hipótesis *Hc* (y del mismo modo para H_{i-1}) consiste en la hipótesis para todos los objetivos y el esperma Hi = (H^Ti, H^si).

Hipótesis H^T_i= (H^{T1}_i,H^{T2}_i,..,H^{TK}_i) son las hipótesis objetivo para los objetivos k, donde cada hipótesis objetivo está compuesta por las hipótesis originales de la madre y el padre H^{Tj}_i = ($H^{Tj}_{i,m}$, $H^{Tj}_{i,f}$). La hipótesis H^T_i= (H^{T1}_i,H^{T2}_i,..,H^{TK}_i) es la hipótesis original del padre para 1 esperma. Luego

$$P(H_{i-1}|H_i) = \prod_j P(H^{Tj}_{i-1,m}|H^{Tj}_{i,m}) \prod_j P(H^{Tj}_{i-1,f}|H^{Tj}_{i,f}) \prod_j P(H^{Sj}_{i-1,f}|H^{Sj}_{i,f}) \quad P(g|h) = \begin{cases} cp & g \neq h \\ 1-cp & g = h \end{cases}$$

y donde cp es la probabilidad de cruce entre SNPs i,i-1, y puede ser calculada por separado a partir de datos HAPMAP.

$PD^{i-1}(H_{i-1}) = P(D_{i-1}|H_{i-1})$ es la probabilidad de los datos sobre SNP i-1, dada esta hipótesis H_{t-1}, y puede ser calculada sumando todos los valores de alelos parentales ordenados, similar al desglose descrito anteriormente.

$$\begin{aligned} &P(D_{i-1}|H_{i-1}) \\ &= \sum_{g^M} \sum_{g^F} P(D_{i-1}|H_{i-1}, g^M, g^F) P(g^M) P(g^F) \\ &= \sum_{g^M} \sum_{g^F} P(D^T_{i-1}|g^M, g^F, H^T_{i-1}) P(D^S_{i-1}|g^F, H^S_{i-1}) P(D^M_{i-1}|g^M) P(D^F_{i-1}|g^F) P(g^M) P(g^F) \end{aligned}$$

Probabilidad sobre SNPs *i +1,….,n*

La derivación en esta sección es similar a la anterior, excepto que se parte del otro extremo, es decir, si definimos $V^i(h,1) = P(D_{i+1,\ldots,n}|h)$, donde h es la hipótesis sobre SNP I, tenemos $V^i = PC^{i+1} \times (PD^{i+1} \cdot V^{i+1})$

Con la condición inicial $V^n(g) = P(end@g)$ (solo una constante igual para todo, no es importante). Por tanto, hallar primero $V^{n-1} = PC^n X(PD^n \cdot V^n)$, y sucesivamente bajar hasta V^i.

Cálculo de hipótesis P(h\D)

Derivar las hipótesis exactas de objetivo o esperma no es parte integrante de la determinación de alelos, pero puede ser muy útil para la comprobación de resultados y otras aplicaciones. El procedimiento es muy similar a derivar probabilidades de genotipo, y se esboza aquí. En particular, para SNP i, objetivo j, e hipótesis h definida como hipótesis particular para SNP i, objetivo j,

$$P(h|D) \sim \sum_{H_i, H^{Tj}_i = h} P(D, H_i) = \sum_{H_i, H^{Tj}_i = h} P(D_{1,\ldots,i-1}|H_i) P(D_{i+1,\ldots,n}|H_i) P(D_i|H_i) P(H_i)$$

Donde todas las partes son derivadas como se describe en otra parte de este documento.

Cálculo del genotipo parental P(g\D)

La derivación del genotipo parental exacto no es parte integrante de la determinación de alelos, pero puede ser muy útil para la comprobación de resultados y otras aplicaciones. El procedimiento es muy similar a derivar probabilidades de genotipo, y es esbozado aquí. En particular, para SNP i, objetivo j, tomemos genotipo materno g^M

$$P(g^M|D) \sim \sum_{H_i, g^F} P(D, H_i, g^M, g^F)$$

$$= \sum_{H_i, g^F} P(D_{1,\ldots,i-1}|H_i) P(D_{i+1,\ldots,m}|H_i) P(D_i|H_i, g^M, g^F) P(H_i) P(g^M) P(g^F)$$

donde todas las partes son derivadas como se describe en otro lugar en este documento.

Cálculo del modelo de respuesta de plataforma $P(D^T\backslash g)$

El modelo de respuesta puede derivarse por separado para cada muestra y cada cromosoma. El objetivo es calcular P((X,Y)|g) donde g = AA, AB, BB.

Primero hacer discreto el rango de la respuesta de intensidad X, Y en T bins B^X,B^Y, derivado como T percentiles espaciados igualmente de datos sobre canales respectivos (T<=20). Entonces se puede calcular P((X,Y)|g) como $P((X,Y)|g) \sim f(b_x, b_y, g)\ for\ X \in b_x, Y \in b_y,$ donde $f(b_x, b_y, g)$ se calcula a partir de los datos. En una realización, los datos pueden proceder de datos resultado de una matriz de genotipado de SNP Illumina y/o datos secuenciados, que tienen distintos modelos. En otras realizaciones, los datos pueden proceder de otras matrices de genotipado, de otros métodos de secuenciación u otras fuentes de datos genéticos.

Modelo para datos Illumina

A partir de datos parentales, calcular el genotipo materno G^M, el genotipo paterno G^F y derivar la frecuencia parental de la muestra $\hat{p}(gm, gf)$ para gm,gf = AA,AB,BB.

Calcular la frecuencia de alelos: $P(g) \sim f(g) = \sum_{gm,gf} P(g|gm, gf) * \hat{p}(gm, gf)$ Definir S^{AA} como el subconjunto de SNPs de los datos objetivo S para el contexto parental AA|AA, es decir, S^{AA} = {S|G^M = AA, G^F = AA}, y S^{BB} como el subconjunto de SNPs de los datos objetivo S para el contexto parental BB|BB, es decir, S^{BB} = {S|G^M = BB, G^F = BB. El valor de alelos de SNPs en S^{AA} ha de ser AA, y de forma similar BB para S^{BB}.

Cálculo conjunto

Definir $f^{Joint}(b_x,b_y,AA)$ como la frecuencia de intensidades de la muestra de bin conjunto en S^{AA}. Esto es una estimación de P((X,Y)|AA).

Definir $f^{Joint}(b_x,b_y,BB)$ como la frecuencia de intensidades de la muestra de bin conjunto en S^{BB}. Esto es una estimación de P((X,Y)|BB).

Definir $f^{joint}(b_x,b_y,;)$ como la frecuencia de intensidades de la muestra de bin conjunto en S. Esto es una estimación de P((X,Y)).

Se sabe que $b((X,Y)) = \sum^{\theta=AA,AB,BB} b((X,Y)|\theta) * b(\theta)$

por lo que podríamos escribir $P((X,Y)|AB) = \frac{P((X,Y)) - P(AA)P((X,Y)|AA) - P(BB)P((X,Y)|BB)}{1 - P(AA) - P(BB)}$

y es posible calcular P((X,Y)|AB) como sigue:

$$f^{joint}(b_x, b_y, AB) = \frac{f^{joint}(b_x b_y) - f(AA) f^{joint}(b_x b_y AA) - f(BB) f^{joint}(b_x b_y BB)}{1 \quad f(AA) \quad f(BB)}.$$

Ahora las funciones f^{joint} (b_x,b_y,g) son una posible estimación de P ((X,Y)|g).

Cálculo marginal

Definir $f^{marginal}(b_x,:,g)$ como la frecuencia de bin marginal de las intensidades de canal X en S^g, para g=AA,BB. Esto es una estimación de P(X|g).

Definir $f^{marginal}(:,b_y,g)$ como la frecuencia de bin marginal de las intensidades de canal Y en S^g, para g=AA,BB. Esto es una estimación de P(Y|g).

Si se supone que las respuestas de canal son independientes (y pueden no serlo), para g=AA,BB, se podría escribir:

$$f^{marginal}(b_x,b_y,g) = f^{marginal}(b_x,:,g) * f^{marginal}(:,b_y,g)$$

Y como anteriormente:

$$f^{marginal}(b_x,b_y,AB) = \frac{f^{marginal}(b_x,b_y,:) - f(AA) f^{marginal}(b_x,b_y,AA) - f(BB) f^{marginal}(b_x,b_y,BB)}{1 - f(AA) - f(BB)}.$$

Ahora la función $f^{marginal}(b_x,b_y,g)$ es otra posible estimación de P((X,Y)|g).

Cálculos combinados

En algunas realizaciones, por ejemplo, cuando f^{joint} está demasiado controlado por los datos, y $f^{marginal}$ es demasiado suave, es decir, no teniendo en cuenta la dependencia del canal, es posible utilizar el cálculo combinado, juntando esos dos para dar:

$$f(b_x,b_y,g) = c * f^{joint}(b_x,b_y,g) + (1-c) * f^{marginal}(b_x,b_y,g),$$

para c = 0,5 (una constante arbitraria).

Modelo para datos secuenciales

Los datos secuenciales son distintos a los datos derivados de matrices de genotipado. Cada SNP se da por separado, junto con diversas localizaciones en torno a ese SNP (típicamente más o menos 400-500), por intensidad para los cuatro canales A,C,T,G. Los datos secuenciales incluyen también la determinación “salvaje” de homocigoto para todas esas localizaciones. Típicamente, la mayoría de las localizaciones no SNP son homocigotos y corresponden al valor de alelos de la determinación salvaje. En una realización es posible suponer que, para localizaciones no SNP, esa determinación salvaje es la “cierta”.

Determinar datos de intensidad no SNP, “datos de localización” se pueden utilizar para ayudar a crear el modelo de respuesta. Los datos de localización tienen el formato LD =(LD1,.. .,LD_n) para n localizaciones, donde LDi=(L^Ai,L^Ci, L^Ti, L^Gi), A,C,T,G intensidades en la localización i. Correspondiente a los datos de determinación salvaje es WD = (W_1,...,W_n), donde Wi es uno de A,C,T,G. Idealmente, si un alelo concreto, digamos C, está presente en la localización i, el valor de intensidad L^Ci debería ser elevado. Si el valor del alelo no está presente, su intensidad debería ser muy baja, idealmente 0. Así, por ejemplo para TT, se puede esperar tener intensidades para (A, T, C, G) = (baja, alta, baja, baja) = (no, sí, no, no). Para AT, cabría esperar tener (alta, alta, baja, baja) = (sí, sí, no, no).

Teniendo esto en cuenta, es posible $f(b_x,b_y,AA) = YD(b_x) * ND(b_y)$, (sí en A, no en B)

$f(b_x,b_y,AB) = YD(b_x) * YD(b_y)$, (sí en A, sí en B)

$f(b_x,b_y,BB) = ND(b_x) * YD(b_y)$, (no en A, sí en B)

donde *YD(b)* es el “sí/presente” y ND(b) es el “no/ausente” una distribución de bin discreta dimensional derivada de los datos. YD puede ser derivado de los datos en el conjunto Y = {todas las intensidades de canal especificadas por determinación salvaje}. ND puede ser derivado de los datos en el conjunto N = {todas las intensidades de canal NO especificadas por determinación salvaje}. Por ejemplo, si los datos de intensidad en una localización particular son (1a, 1c, 1t, 1g) y la determinación salvaje es T, esto irá hacia el conjunto Y, y 1a, 1c, 1g irá hacia el conjunto N.

Si se supone independencia de canal e idéntica distribución (modelo I.I.D.), las distribuciones YD, ND son solo frecuencia de muestra simple de datos en el conjunto Y y el N respectivamente.

No obstante, los cuatro canales pueden ser subamplificados o sobreamplificados, y en consecuencia no son independientes. En una realización, es posible crear un canal dependiente y distribución idéntica (modelo D.I.D.), escalando la intensidad por intensidad de canal máxima en esa localización y aplicando el modelo I.I.D.

Resultados

En esta sección se comentan los resultados de este método de determinación de alelos, aplicada a datos reales, trabajando con un conjunto de datos genéticos medidos de individuos relacionados. Los datos de entrada

consistieron en el resultado en bruto de una matriz de genotipado Illumina Infmium. Los datos incluían 22 cromosomas, de 1000 SNP cada uno, para un conjunto de individuos relacionados, incluyendo:

2 niños (con 2 muestras por niño),

3 embriones (2 muestras por embrión),

ambos progenitores (la madre y el padre, 2 muestras genómicas por cada progenitor)

3 espermas (1 muestra de cada uno)

Resultado de la determinación de objetivo

La tasa global de aciertos dada para niños, donde las mediciones genómicas efectuadas en 30 muestras de masa de tejido fueron consideradas “ciertas”, fue del 98,55%. La tasa de aciertos varió para los distintos contextos, y se muestra en la siguiente tabla:

(m1m2|f1f2)	Tasa aciertos	deDesviación estándar
AA|AA	0,9963	σ = 0,1822
AA|AB	0,9363	σ = 0,0933
AA|BB	0,9995	σ = 0,0365
AB|AA	0,9665	σ = 0,0956
AB|AB	0,9609	σ = 0,1313
AB|AA	0,9635	σ = 0,1013
BB|AA	0,9980	σ = 0,0337
BB|AB	0,9940	σ = 0,1088
BB|BB	0,9983	σ = 0,2112

La tasa de aciertos varió por cromosomas, y osciló de aproximadamente el 99,5% a aproximadamente el 96,4%. Los cromosomas 16, 19 y 22 estuvieron por debajo de aproximadamente el 98%. Obsérvese que las tasas de acierto para los SNPs derivados del padre fueron de aproximadamente el 99,82%, y las tasas de acierto para las SNPs derivados de la madre fueron de aproximadamente el 93,75%. Las mejores tasas de acierto para los SNPs derivados del padre se deben al mejor ajuste por fases del padre, gracias a los datos genéticos ajustados por fases disponibles por genotipado de esperma.

La tasa de aciertos por bin de confianza se refiere a la tasa de aciertos para el conjunto de determinaciones de alelos de las que se espera que tengan un intervalo de confianza determinado. La tasa de aciertos global para todos los datos fue aproximadamente una tasa de aciertos del 98,55%. La tasa de aciertos para las determinaciones de alelos de las que se esperaba que tuvieran confianzas superiores a aproximadamente 90%, lo que corresponde a aproximadamente el 96,2% de todas las determinaciones de alelos realizadas, fue del 99,63%. La tasa de aciertos para todas las determinaciones de alelos para las que se esperaba que tuvieran confianzas superiores a aproximadamente el 99%, lo que corresponde a aproximadamente el 90,37% de los datos, fue aproximadamente el 99,9%. Las tasas de aciertos para bins de confianza individuales indican que las confianzas previstas son muy exactas, dentro de los límites de la significación estadística. Por ejemplo, para las determinaciones de alelos con confianzas previstas de entre aproximadamente el 80% y aproximadamente el 90% la tasa de aciertos real fue aproximadamente el 85,0%. Para las determinaciones de alelos con confianzas esperadas de aproximadamente el 70% al 80%, la tasa de aciertos real fue de aproximadamente el 76,2%. Para las determinaciones de alelos con confianzas previstas de entre aproximadamente el 96% y 97%, la tasa de aciertos real fue de aproximadamente el 96,3%. Para las determinaciones de alelos con confianzas previstas de entre aproximadamente el 94% y 95%, la tasa de aciertos real fue de aproximadamente el 93,9%. Para las determinaciones de alelos con confianzas previstas de entre aproximadamente el 99,1% y 99,2%, la tasa de aciertos real fue aproximadamente el 99,4%. Para las determinaciones de alelos con confianzas previstas de entre aproximadamente el 99,8% y 99,9%, la tasa de aciertos real fue de aproximadamente el 99,7%. Las Figuras 10A y 10B y las Figuras 11A y 11B presentan representaciones gráficas de tasas de aciertos de objetivo realizadas, con barras de confianza, frente a tasa de aciertos prevista por confianza. La Figura 10A representa gráficamente la tasa de aciertos real frente a la confianza prevista para bins de una anchura de tres y un tercio porcentual, y la Figura 11A representa gráficamente la tasa de aciertos real frente a

la confianza prevista para bins de una anchura de una mitad porcentual. La línea diagonal representa el caso ideal donde la tasa de aciertos real es igual a la confianza prevista. La Figura 10B muestra la población relativa de los diversos bins de la Figura 10A, y la Figura 11B muestra la población relativa de los diversos bins de la Figura 11A. Los bins con una población o frecuencia mayores se espera que presenten una desviación menor.

Como control, se llevó a cabo el mismo experimento, pero utilizando mediciones genómicas tomadas de datos en bloque, en lugar de mediciones de una sola célula, como los datos genéticos objetivo medidos. En este caso, la tasa global de aciertos fue de aproximadamente el 99,88%.

Probabilidad de hipótesis con cruces

El método descrito aquí puede determinar también si se ha producido un cruce en la formación de los embriones. Dado que la exactitud de las determinaciones de alelos depende de conocer la identidad de los alelos vecinos, cabría esperar que en determinaciones de alelos cerca de un cruce, donde los alelos vecinos pueden no ser del mismo haplotipo, la confianza de esas determinaciones puede descender. Esto puede verse en las Figuras 12A-12B. La Figura 12A muestra la representación gráfica de la confianza de alelos promediada sobre los SNPs vecinos para un cromosoma típico. Se representan dos conjuntos distintos de datos, E5 y E5GEN, obtenidos del mismo individuo objetivo, pero utilizando métodos distintos. Una brusca caída en la confianza en torno a una región determinada de un cromosoma indica que se ha producido un cruce en la localización durante la meiosis que dio origen al individuo objetivo. La Figura 12B muestra una representación en línea del cromosoma, con una estrella indicando la localización en que la hipótesis de ploidía ha determinado que se produjo un cruce. En la Figura 12B, es posible observar dos cruces, un cruce en el homólogo materno en torno a SNP 350, y un cruce en el homólogo paterno en torno a SNP 820. La línea marcada “E5” se dió cuando el método se aplica en datos objetivo de una sola célula, y la línea marcada “E5GEN” se dió cuando el método se aplicó a datos genómicos medidos en masa de tejido. El hecho de que las líneas sean similares indica que el método reconstruye con precisión los datos genéticos del objetivo de una sola célula, específicamente, la localización del cruce.

Variar el número y las confianzas de los datos de entrada

En una realización de la presente divulgación, es posible utilizar datos genómicos de la madre y del padre, y datos genéticos de una sola célula medidos de los blastómeros y el esperma. En otra realización de la presente divulgación, es posible también utilizar datos genómicos de un niño nacido de los mismos padres, como información adicional para ayudar a incrementar la precisión de la determinación de la información genética del objetivo de una sola célula. En un experimento, se utilizaron los datos genómicos de ambos progenitores, junto con las mediciones genéticas de una sola célula de 2 células objetivo de embrión, y la tasa media de aciertos en el objetivo fue de aproximadamente el 95%. Se llevó a cabo un experimento similar utilizando los datos genómicos de ambos progenitores, los datos genómicos de un hermano, y la información genética de objetivo de una sola célula procedente de una célula, y la precisión añadida de los datos genéticos de hermanos incrementó la tasa de aciertos en la célula objetivo a aproximadamente el 99%.

En otra realización de la presente divulgación, es posible utilizar los datos genéticos de cero, uno, dos, tres, cuatro o cinco o más espermas como entrada para el método. En algunas realizaciones de la presente divulgación es posible utilizar los datos genéticos de uno, dos, tres, cuatro, cinco o más de cinco embriones hermanos como entrada para el método. En general, cuanto mayor es el número de entradas, mayor es la precisión de las determinaciones de alelos objetivo. También cuanto mayor es la precisión de las mediciones de las entradas, tanto mayor es la precisión de las determinaciones de alelos objetivo.

Se llevó a cabo otro experimento con distintos conjuntos de entradas de blastómeros y esperma, en la forma de mediciones de blastómeros de una sola célula, y mediciones de esperma de una sola célula. La siguiente tabla muestra que cuanto mayor es el número de entradas, tanto mayor es la tasa de aciertos de alelos y la tasa de aciertos de hipótesis en el objetivo. Adviértase que “núm. espermas” indica el número de espermas utilizado en la determinación; “núm. emb” corresponde al número total de embriones hermanos usados en la determinación, incluyendo el objetivo; BK28 es un conjunto de datos particular.

BK28 tasa de aciertos de alelos (%)				
	núm. espermas			
núm. emb	0	1	2	3
3	93,46 95.18	95,18	95,69	95,86
4	95,06 96.13	96,13	96,59	96,75
5	95,93 96.67	96,67	97,00	97,15
BK28 tasa de aciertos hipótesis (%)				
	núm. espermas			

núm. emb	0	1	*2*	*3*
3	98,49 99.72	99,73	99,73	99,74
4	99,70 99.72	99,72	99,73	99,73
5	99,64 99.65	99,65	99,52	99,68

Amplificación de ADN genómico

La amplificación del genoma se puede lograr por múltiples métodos, incluyendo: PCR (LM-PCR) mediada por ligación, cebador de oligonucleótido degenerado PCR (DOP-PCR), y amplificación de desplazamiento múltiple (MDA). De los tres métodos, DOP-PCR produce de forma fiable grandes cantidades de ADN a partir de pequeñas cantidades de ADN, incluyendo copias individuales de cromosomas; este método puede ser el más apropiado para el genotipado de datos diploides parentales, donde la fidelidad de los datos es crítica. MDA es el método más rápido, produciendo una amplificación de cien veces de ADN en unas pocas horas; este método puede ser el más apropiado para genotipar células embriónicas, o en otras situaciones donde el tiempo es esencial.

La amplificación de fondo es un problema en cada uno de esos métodos, dado que cada método amplificaría potencialmente contaminando el ADN. Cantidades muy pequeñas de contaminación pueden envenenar irreversiblemente el ensayo y dar datos falsos. Por tanto, es crítico utilizar entornos de laboratorio limpios, donde los flujos pre y post amplificación estén completamente separados físicamente. Flujos de amplificación de ADN limpios y libres de contaminación son ahora rutinarios en la biología molecular industrial, y simplemente requiere prestar mucha atención al detalle.

Ensayo de genotipado e hibridación

El genotipado del ADN amplificado puede hacerse por muchos métodos, incluyendo sondas de inversión molecular (MIPs) como Genflex Tag Array de Affymetrix, microarrays como la matriz Affymetrix's 500K o las matrices Illumina Bead Arrays, o ensayos de genotipado de SNP tales como el ensayo AppliedBioscience's TaqMan. Todos estos son ejemplos de técnicas de genotipado. La matriz The Affymetrix 500K, MIPs/GenFlex, TaqMan y el ensayo Illumina requieren todos cantidades de microgramos de ADN, por lo que el genotipado de una sola célula con cualquier flujo de trabajo requiere algún tipo de amplificación.

En el contexto del diagnóstico preimplantación durante la IVF (Fertilización in vitro), son significativas las limitaciones de tiempo inherentes, y los métodos que se pueden desarrollar en un día pueden proporcionar una ventaja clara. El protocolo estándar del ensayo MIPs es un proceso de tiempo relativamente intensivo que se tarda en completar típicamente de 2,5 a 3 días. Los ensayos de matrices 500K y el Illumina tienen un tiempo de aplicación relativamente más rápido: aproximadamente de 1,5 a 2 días para generar datos altamente fiables en el protocolo estándar. Estos dos métodos son optimizables, y se calcula que el tiempo de aplicación para el ensayo de genotipado para la matriz 500k y/o el ensayo Illumina podría reducirse a menos de 24 horas. Aún es más rápido en el ensayo TaqMan que puede completarse en 3 horas. Para todos estos métodos, la reducción del tiempo del ensayo puede tener como resultado una reducción en la calidad de los métodos, no obstante, esto es exactamente de lo que trata la presente divulgación.

Naturalmente, en situaciones donde la medida del tiempo es crítica, como genotipar un blastómero durante la IVF, los ensayos más rápidos tienen una clara ventaja sobre los ensayos más lentos, mientras que en los casos en los que no hay tanta presión de tiempo, como cuando en el fenotipado del ADN parental antes de la IVF se ha iniciado, otros factores predominarán en la elección del método apropiado. Cualquier técnica que se desarrolle hasta el punto de permitir un alto rendimiento de genotipado suficientemente rápido podría ser utilizada en el genotipado de material genético para su uso con este método.

Métodos para la simultánea amplificación de locus objetivo y amplificación de genoma completo

Durante la amplificación de genoma completo de pequeñas cantidades de material genético, por ligación mediada PCR (LM-PCR), amplificación de desplazamiento múltiple (MDA), o se producen otros métodos, pérdidas de loci de forma aleatoria e inevitable. Con frecuencia es deseable amplificar el genoma completo no específicamente, pero garantizar que un locus particular se amplifica con mayor exactitud. Es posible realizar simultáneamente la focalización de locus y la amplificación de genoma completo.

En una realización, es posible combinar la reacción específica de cadena de polimerasa (PCR) para amplificar loci particulares de interés, con cualquier método de amplificación de genoma completo generalizado. Esto puede incluir, entre otros, preamplificación de loci particulares antes de la amplificación generalizada por MDA o LM-PCR, la adición de cebadores PCR específicos a cebadores universales en el paso PCR de LM-PCR, y la adición de cebadores PCR específicos para cebadores degenerados en MDA.

Respuesta de plataforma

Hay muchos métodos que pueden utilizarse para medir datos genéticos. Ninguno de los métodos conocidos actualmente en la técnica es capaz de medir los datos genéticos con una precisión del 100%, sino que más bien hay siempre errores, o sesgos estadísticos en los datos. Puede esperarse que el método de medición presente ciertos

sesgos previsibles estadísticamente en la medición. Cabe esperar que determinados conjuntos de ADN, amplificados por ciertos métodos, y medidos con ciertas técnicas, puedan proporcionar mediciones que son distintas cualitativa y cuantitativamente de otros conjuntos de ADN, amplificados por otros métodos, y/o medidos con distintas técnicas. En algunos casos esos errores pueden ser debidos al método de medición. En algunos casos este error puede deberse al estado del ADN. En algunos casos este sesgo puede ser debido a la tendencia de algunos tipos de ADN a responder de forma distinta a un método de medición genética determinado. En algunos casos, las mediciones pueden diferir de formas que se correlacionan con el número de células utilizadas. En algunos casos, las mediciones pueden diferir en base a la técnica de medición, por ejemplo, qué técnica de secuenciación o técnica de genotipado de matriz se usa. En algunos casos, distintos cromosomas pueden amplificarse en distinta medida. En algunos casos, cierto alelo puede ser más o menos probable que se amplifiquen. En algunos casos, el error, sesgo o respuesta diferencial pueden ser debidos a una combinación de factores. En muchos o todos estos casos, la previsibilidad estadística de estas diferencias de medición, denominada la “respuesta de plataforma”, puede ser utilizada para corregir esos factores, y puede dar datos con una precisión maximizada, y en los que cada medición va asociada a una confianza apropiada.

La respuesta de plataforma puede ser descrita como una caracterización matemática de las características de entrada/salida de una plataforma de medición genética, como Taqman o Infinium. La entrada al canal es el material genético amplificado con cualquier material genético recocido marcado con fluorescente. La salida del canal podrían ser determinaciones de alelos (cualitativas) o mediciones numéricas en bruto (cuantitativas), dependiendo del contexto. Por ejemplo, en el caso en que la salida numérica en bruto de la plataforma se reduce a determinaciones de genotipo cualitativas, la respuesta de plataforma puede ser una matriz de transición errónea que describe la probabilidad condicional de ver una determinación particular de genotipo de salida, dada una entrada particular de genotipo verdadero. En una realización, en la que la salida de la plataforma se deja como mediciones numéricas en bruto, la respuesta de plataforma puede ser una función de densidad de probabilidad condicional que describe la probabilidad de las salidas numéricas dada una entrada de genotipo verdadero particular.

En algunas realizaciones de la presente divulgación, el conocimiento de la respuesta de plataforma puede ser utilizado para corregir estadísticamente el sesgo. En algunas realizaciones de la presente divulgación, el conocimiento de la respuesta de plataforma puede ser utilizado para incrementar la precisión de los datos genéticos. Esto puede hacerse realizando una operación estadística sobre los datos, que actúa de forma opuesta a la tendencia al sesgo del proceso de medición. Puede incluir atribuir la confianza apropiada a un dato determinado, de forma que al combinarlo con otros datos, la hipótesis considerada la más probable es la de mayor probabilidad de corresponder al estado genético real del individuo en cuestión.

Otras notas

Como se ha indicado previamente, considerando el beneficio de esta divulgación, hay más realizaciones que pueden implementar uno o más de los sistemas, métodos y características divulgados aquí.

En algunas realizaciones de la presente divulgación, se puede utilizar un método estadístico para eliminar el sesgo en los datos debido a la tendencia de los alelos maternos a amplificarse de forma desproporcionada respecto a los otros alelos. En algunas realizaciones de la presente divulgación, se puede utilizar un método estadístico para eliminar el sesgo en los datos debido a la tendencia de los alelos paternos a amplificarse de forma desproporcionada respecto a los otros alelos. En algunas realizaciones de la presente divulgación, se puede utilizar un método estadístico para eliminar el sesgo en los datos debido a la tendencia de determinadas sondas a amplificar determinados SNPs de forma desproporcionada respecto a otros SNPs.

Imaginemos el espacio bidimensional donde la coordenada x es la intensidad del canal x, y la coordenada y es la intensidad del canal y. En este espacio, cabe esperar que las medias de contexto estarían en la línea definida por las medias de los contextos BB|BB and AA|AA. En algunos casos, puede observarse que las medias promediadas de los contextos no están en esta línea, sino que están sesgadas de forma estadística; esto puede ser denominado “sesgo fuera de línea”. En algunas realizaciones de la presente divulgación, se puede utilizar un método estadístico para corregir el sesgo fuera de línea en los datos.

En algunos casos, los puntos extendidos en la representación gráfica de las medias de contexto podrían ser causados por translocación. Si se produce una translocación, se podría esperar ver anormalidades solo en los extremos del cromosoma. Por tanto, si el cromosoma se divide en segmentos, y las representaciones gráficas de la media de contextos de cada segmento se representan, esos segmentos que se encuentran ahí de una translocación puede esperarse que respondan como una trisomía o monosomía verdaderas, mientras que los segmentos restantes parecen disómicos. En algunas realizaciones de la presente divulgación, se puede utilizar un método estadístico para determinar si se ha producido translocación en un cromosoma concreto considerando las medias de contexto de distintos segmentos del cromosoma.

En algunos casos, puede ser deseable incluir un amplio número de individuos relacionados en el cálculo para determinar el estado genético más probable de un objetivo. En algunos casos, aplicar el algoritmo con todos los individuos relacionados deseados puede no ser factible debido a limitaciones de potencia de cálculo o tiempo. La potencia de cálculo necesaria para calcular los valores de alelos más probables para el objetivo aumenta exponencialmente con el número de espermas, blastómeros y otros genotipos de entrada de individuos relacionados. En una realización, estos problemas pueden superarse utilizando un método denominado “subsetting”

(subconjuntos), donde los cálculos pueden dividirse en conjuntos menores, realizarlos por separado y luego combinarlos. En una realización de la presente divulgación, se pueden tener los datos genéticos de los progenitores junto con los de diez embriones y diez espermas. En esta realización, se podrían aplicar varios subalgoritmos menores con, por ejemplo, tres embriones y tres espermas, y luego combinar los resultados. En una realización, el número de embriones hermanos utilizados en la determinación puede ser de uno a tres, de tres a cinco, de cinco a diez, de diez a veinte, o más de veinte. En una realización el número de espermas cuyo contenido genético es conocido puede ser de uno a tres, de tres a cinco, de cinco a diez, de diez a veinte, o más de veinte. En una realización cada cromosoma puede ser dividido en de dos a cinco, de cinco a diez, de diez a veinte, o más de veinte subconjuntos.

En una realización de la presente divulgación, cualquiera de los métodos descritos aquí puede ser modificado para permitir que objetivos múltiples deriven del mismo individuo objetivo. Esto puede mejorar la exactitud del modelo, ya que mediciones genéticas múltiples pueden proporcionar más datos con los que se puede determinar el genotipo objetivo. En métodos previos, un conjunto de datos genéticos objetivo sirvió como los datos primarios que se comunicaron, y los otros sirvieron como datos para la doble comprobación de los datos genéticos objetivo primarios. Esta realización de la presente divulgación es una mejora sobre los métodos previos porque diversos conjuntos de datos genéticos, cada uno de ellos medido de material genético tomado del individuo objetivo, se consideran en paralelo, y así ambos conjuntos de datos genéticos objetivo sirven para ayudar a determinar qué sección de los datos genéticos parentales, medidos con alta precisión, compone el genoma embriónico. En una realización de la presente divulgación, el individuo objetivo es un embrión, y las distintas mediciones de genotipo se efectúan en diversos blastómeros biopsiados. En otra realización, se podrían utilizar también múltiples blastómeros de distintos embriones, del mismo embrión, células de niños nacidos o una combinación de lo anterior.

En algunas realizaciones de la presente divulgación, los métodos descritos aquí pueden usarse para determinar el estado genético de un feto en desarrollo prenatalmente y de forma no invasiva. La fuente del material genético a usar en la determinación del estado genético del feto pueden ser células fetales, como glóbulos rojos nucleados fetales, aislados de la sangre materna. El método puede implicar obtener una muestra de sangre de la madre gestante. El método puede incluir aislar un glóbulo rojo fetal utilizando técnicas visuales, basadas en la idea de que una determinada combinación de colores va asociada exclusivamente a un glóbulo rojo nucleado, y una combinación similar de colores no va asociada a ninguna otra célula presente en la sangre materna. La combinación de colores asociada a los glóbulos rojos nucleados puede incluir el color rojo de la hemoglobina en torno al núcleo, color que puede potenciarse por tinción, y el color del material nuclear puede teñirse, por ejemplo, azul. Aislando las células de la sangre materna y extendiéndolas sobre una placa, y luego identificando los puntos en los que se ve tanto el rojo (de la hemoglobina) como el azul (del material nuclear) se puede identificar la localización de glóbulos rojos nucleados. Entonces se pueden extraer esos glóbulos rojos nucleados utilizando un micromanipulador, y usar técnicas de genotipado y/o secuenciación para medir aspectos del genotipo del material genético en esas células. En una realización de la presente divulgación, se puede utilizar entonces una técnica basada en la informática, como las descritas en esta divulgación, para determinar si las células son o no realmente de origen fetal.

En una realización de la presente divulgación, se puede utilizar entonces una técnica basada en la informática como las descritas en esta divulgación para determinar el estado de ploidía de uno de un conjunto de cromosomas en esas células. En una realización de la presente divulgación se puede utilizar entonces una técnica basada en la informática como las descritas en esta divulgación, para determinar el estado genético de las células. Aplicada a los datos genéticos de la célula, PARENTAL SUPPORT™ podría indicar si un glóbulo rojo nucleado es de origen fetal o materno, identificando si la célula contiene un cromosoma de la madre y uno del padre, o dos cromosomas de la madre.

En una realización, se puede teñir el glóbulo rojo nucleado con una tinción que solo resulte fluorescente en presencia de hemoglobina fetal y no de hemoglobina materna, y así eliminar la ambigüedad respecto a si un glóbulo rojo deriva de la madre o del feto. Algunas realizaciones de la presente divulgación pueden implicar la tinción o marcar de otra forma el material nuclear. Algunas realizaciones de la presente divulgación pueden implicar marcar específicamente material nuclear fetal utilizando anticuerpos específicos de célula fetal. Algunas realizaciones de la presente divulgación pueden incluir aislar, mediante diversos métodos posibles, una o varias células, algunas o todas ellas de origen fetal. Algunas realizaciones de la presente divulgación pueden incluir amplificar el ADN en esas células, y utilizar un microarray de genotipado de alto rendimiento, como la matriz Illumina Infmium, para genotipar el ADN amplificado. Algunas realizaciones de la presente divulgación pueden incluir utilizar el ADN parental medido o conocido para inferir los datos genéticos más precisos del feto. En algunas realizaciones, se puede asociar una confianza a la determinación de uno o más alelos, o al estado de ploidía del feto. Algunas realizaciones de la presente divulgación pueden implicar teñir el glóbulo rojo nucleado con una tinción que solo resulta fluorescente en presencia de hemoglobina fetal y no de hemoglobina materna, y así eliminar la ambigüedad sobre si un glóbulo rojo nucleado deriva de la madre o del feto.

Hay muchas otras formas de aislar las células fetales de la sangre materna, o ADN fetal de sangre materna, o enriquecer muestras de material genético fetal en presencia de material genético materno. Algunos de esos métodos se mencionan aquí, pero no se pretende que la lista sea exhaustiva. Se mencionan aquí a efectos de comodidad algunas técnicas apropiadas: utilizar anticuerpos marcados con fluorescencia o de otro modo, cromotografía de exclusión de tamaño, etiquetas de afinidad marcadas magnéticamente o de otro modo, diferencias epigenéticas, como metilación diferencial entre las células maternas y fetales en alelos específicos, centrifugación de gradiente de

densidad seguido de reducción CD45/14 y selección CD71-positiva de células CD45/14 negativas, gradientes de Percoll individuales o dobles con distintas osmolaridades, o el método de la lectina específica de galactosa.

Una realización de la presente divulgación podría ser como sigue: una mujer gestante quiere saber si su feto tiene el Síndrome de Down, y si tendrá fibrosis quística. Un médico extrae su sangre, y tiñe la hemoglobina con un marcador de forma que aparezca claramente roja, y tiñe el material nuclear con otro marcador para que resulte claramente azul. Sabiendo que los glóbulos rojos maternos son típicamente anucleares, mientras que una elevada proporción de células fetales contienen un núcleo, puede aislar visualmente un número de glóbulos rojos nucleados identificando las células que presentan un color tanto rojo como azul. El médico recoge esas células de la placa con un micromanipulador y las envía a un laboratorio para que amplifique y genotipe diez células individuales. Viendo las mediciones genéticas, el PARENTAL SUPPORT™ puede determinar que seis de las diez células son de sangre materna, y cuatro de ellas son células fetales. Si un niño ha nacido ya de una madre gestante, se puede utilizar también PARENTAL SUPPORT™ para determinar que la célula fetal es distinta de las células del niño que ha nacido realizando determinaciones de alelos fiables sobre las células fetales y mostrando que son distintas de las del niño nacido. Los datos genéticos medidos de las células fetales son de muy mala calidad, y contienen muchas pérdidas de alelos, debido a la dificultad en el genotipado de células individuales. El médico puede utilizar el ADN fetal medido junto con las mediciones de ADN fiables de los padres para inferir el genoma del feto con gran precisión utilizando Parental Support. El médico puede determinar tanto el estado de ploidía del feto, como la presencia o ausencia de varios genes de interés vinculados a enfermedad.

En algunas realizaciones de la presente divulgación, se pueden cambiar diversos parámetros sin alterar la esencia de la presente divulgación. Por ejemplo, los datos genéticos pueden obtenerse utilizando cualquier plataforma de genotipado de alto rendimiento, o se pueden obtener a partir de cualquier método de genotipado, o pueden ser simulados, inferidos o conocidos de otra forma. Se pueden utilizar diversos lenguajes computacionales para codificar los algoritmos descritos en esta divulgación, y se podrían utilizar diversas plataformas computacionales para realizar los cálculos. Por ejemplo, los cálculos pueden realizarse utilizando ordenadores personales, superordenadores, una plataforma de computación masivamente paralela, o incluso plataformas de computación no basadas en silicona, como un número suficientemente grande de personas provistas de ábacos.

Algunas de las matemáticas en esta divulgación establecen hipótesis sobre un número limitado de estados de aneuploidía. En algunos casos, por ejemplo, solo se espera que cero, uno o dos cromosomas sean originarios de cada progenitor. En algunas realizaciones de la presente divulgación, las derivaciones matemáticas pueden ser ampliadas para tener en cuenta otras formas de aneuploidía, como la cuadrosomía, donde tres cromosomas son originarios de un progenitor, la pentasomía, etc., sin cambiar los conceptos fundamentales de la presente divulgación.

En algunas realizaciones de la presente divulgación, un individuo relacionado puede referirse a cualquier individuo que esté relacionado genéticamente, y por tanto comparta bloques de haplotipo con el individuo objetivo. Algunos ejemplos de individuos relacionados incluyen: padre biológico, madre biológica, hijo, hija, hermano, hermana, medio hermano, media hermana, abuelo, abuela, tío, tía, sobrino, sobrina, nieto, nieta, primo, clon, el propio individuo, y otros individuos con relación genética con el objetivo conocida. El término “individuo relacionado” incluye también cualquier embrión, feto, esperma, huevo, blastómero o cuerpo polar derivados de un individuo relacionado.

En algunas realizaciones de la presente divulgación, el individuo objetivo puede referirse a un adulto, un menor, un feto, un embrión, un blastómero, una célula o un conjunto de células de un individuo, o de una línea celular, o cualquier conjunto de material genético. El material objetivo puede estar vivo, muerto, congelado o en estasis.

En algunas realizaciones de la presente divulgación, donde el individuo objetivo se refiere a un blastómero utilizado para diagnosticar un embrión, puede haber casos causados por mosaicismo en los que el genoma del blastómero analizado no se corresponda exactamente con los genomas de todas las demás células del embrión.

En algunas realizaciones de la presente divulgación, es posible utilizar el método divulgado aquí en el contexto del genotipado y/o cariotipado del cáncer, donde una o más células cancerosas es considerada el individuo objetivo, y el tejido no canceroso del individuo afectado con cáncer es considerado ser el individuo relacionado. El tejido no canceroso del individuo afectado por el objetivo podría proporcionar el conjunto de determinaciones de genotipo del individuo relacionado, que permitiría la determinación del número de copias cromosómicas de la célula o células cancerosas utilizando los métodos divulgados aquí.

En algunas realizaciones de la presente divulgación, dado que todas las personas vivas, o que han estado vivas, contienen datos genéticos, los métodos son igualmente aplicables a cualquier ser humano, animal o planta, vivo o muerto, que herede o hubiera heredado cromosomas de otros individuos.

Es también importante advertir que los datos genéticos embriónicos que pueden ser generados midiendo el ADN amplificado de un blastómero pueden ser usados para muchos fines. Por ejemplo, pueden usarse para detectar aneuploidía, disomía uniparental, determinar el sexo del individuo, así como también para realizar diversas predicciones fenotípicas basadas en alelos asociados a fenotipo. Actualmente, en los laboratorios de IVF, debido a las técnicas utilizadas, sucede con frecuencia que un blastómero puede proporcionar solamente material genético suficiente para hacer pruebas sobre un trastorno, como la aneuploidía, o una enfermedad monogénica determinada. Dado que el método divulgado aquí tiene un primer paso común de medir un amplio conjunto de SNPs de un blastómero, con independencia del tipo de predicción a realizar, un médico, progenitor u otro agente no se ve

forzado a elegir un número limitado de trastornos para los que efectuar el cribado. Por el contrario, existe la opción de cribar para tantos genes y/o fenotipos como permitan los conocimientos médicos. Con el método divulgado, una ventaja para identificar trastornos determinados para los que efectuar el cribado, antes de genotipar el blastómero es que si se decide que determinados loci son especialmente importantes, se puede seleccionar un conjunto más apropiado de SNPs con mayor probabilidad de cosegregación con el locus de interés, incrementando así la confianza de las determinaciones de alelos de interés.

En algunas realizaciones, los sistemas, métodos y técnicas de la presente divulgación pueden ser utilizados para reducir las probabilidades de que un embrión implantado, obtenido por fertilización in vitro, sufra aborto espontáneo.

En algunas realizaciones de la presente divulgación, los sistemas, métodos y técnicas de la presente divulgación pueden ser usados conjuntamente con otros procedimientos de cribado de embriones o comprobación prenatal. Los sistemas, métodos y técnicas de la presente divulgación se emplean en métodos para aumentar la probabilidad de que los embriones y fetos obtenidos por fertilización in vitro sean implantados con éxito y sigan viables durante todo el periodo de gestación. Además, los sistemas, métodos y técnicas de la presente divulgación se emplean en métodos que pueden reducir la probabilidad de que los embriones y fetos obtenidos por fertilización in vitro y que son implantados, no tengan un riesgo específico de enfermedad congénita.

En algunas realizaciones, los sistemas, métodos y técnicas de la presente divulgación se utilizan en métodos para reducir la probabilidad de la implantación de un embrión con riesgo específico de una enfermedad congénita, comprobando por lo menos una célula extraída de embriones tempranos concebidos mediante fertilización in vitro, y transferir al útero materno solamente aquellos embriones en los que se determine que no han heredado la enfermedad congénita.

En algunas realizaciones, los sistemas, métodos y técnicas de la presente divulgación son usados en métodos para reducir la probabilidad de implantación de un embrión con riesgo específico de una anormalidad cromosómica, comprobando por lo menos una célula extraída de embriones tempranos concebidos por fertilización in vitro, y transferir al útero materno solamente los embriones en los que se determine que carecen de anormalidades cromosómicas.

En algunas realizaciones, los sistemas, métodos y técnicas de la presente divulgación se utilizan en métodos para aumentar la probabilidad de implantación de un embrión obtenido por fertilización in vitro, sea transferido y tenga un riesgo reducido de conllevar una enfermedad congénita.

En algunas realizaciones, la enfermedad congénita es una malformación, defecto del tubo neural, anormalidad cromosómica, síndrome de Down (o trisomía 21), trisomía 18, espina bífida, paladar hendido, enfermedad de Tay Sachs, anemia falciforme, talasemia, fibrosis quística, enfermedad de Huntington, síndrome del maullido, y/o síndrome X frágil. Las anormalidades cromosómicas pueden incluir, entre otras, síndrome de Down (cromosoma 21 extra), síndrome de Turner (45X0) y síndrome de Klinefelter (un macho con 2 cromosomas X).

En algunas realizaciones, la malformación puede ser malformación de un miembro. Las malformaciones de miembros pueden incluir entre otras, amelia, ectrodactilia, focomelia, polimelia, polidactilia, sindactilia, polisindactilia, oligodactilia, braquidactilia, acondroplasia, aplasia o hipoplasia congénitas, síndrome de la banda amniótica y disostosis cleidocraneal.

En algunas realizaciones, la malformación puede ser una malformación congénita del corazón. Las malformaciones congénitas del corazón pueden incluir, entre otras, ducto arterioso patente, defecto septal atrial y tetralogía de fallot.

En algunas realizaciones, la malformación puede ser una malformación congénita del sistema nervioso. Las malformaciones congénitas del sistema nervioso incluyen, entre otras, defectos del tubo neural (ej., espina bífida, meningocele, meningomielocele, encefalocele y anencefalia), malformación de Arnold-Chiari, malformación de Dandy-Walker, hidrocefalia, microencefalia, megalencefalia, lisencefalia, polimicrogiria, holoprosencefalia, y agénesis del corpus callosum.

En algunas realizaciones, la malformación puede ser una malformación congénita del sistema gastrointestinal. Las malformaciones congénitas del sistema gastrointestinal incluyen, entre otras, estenosis, atresia y ano imperforado.

En algunas realizaciones, los sistemas, métodos y técnicas de la presente divulgación se utilizan en métodos para aumentar la probabilidad de implantar un embrión obtenido por fertilización in vitro con un riesgo reducido de conllevar predisposición para una enfermedad genética.

En algunas realizaciones, la enfermedad genética es monogénica o multigénica. Las enfermedades genéticas incluyen, entre otras, el síndrome de Bloom, la enfermedad de Canavan, fibrosis quística, disautonomía familiar, síndrome de Riley-Day, anemia de Fanconi (grupo C), enfermedad de Gaucher, enfermedad de almacenamiento de glicógeno, enfermedad de orina de jarabe de arce, mucolipidosis IV, enfermedad de Niemann-Pick, enfermedad de Tay-Sachs, beta talasemia, anemia falciforme, alfa talasemia, deficiencia de factor XI, ataxia de Friedrich, MCAD, enfermedad de Parkinson juvenil, connexin 26, SMA, síndrome de Rett, fenilcetonuria, distrofia muscular de Becker, distrofia muscular de Duchennes, síndrome de X frágil, hemofilia A, demencia tipo Alzheimer – aparición precoz, cáncer de mama/ovario, cáncer de colon, diabetes MODY, enfermedad de Huntington, distrofia muscular miotónica, enfermedad de Parkinson –aparición precoz, síndrome de Peutz-Jeghers, enfermedad de riñón poliquístico, distonía de torsión.

Combinaciones de los aspectos de la presente divulgación

Como se ha señalado anteriormente, considerando el beneficio de esta divulgación, hay más aspectos y realizaciones que pueden implementar uno o más de los sistemas, métodos y características divulgados aquí. Más abajo se incluye una corta lista de ejemplos ilustrando situaciones en las que los diversos aspectos de la presente divulgación pueden combinarse de distintas maneras. Es importante tener en cuenta que la lista no pretende ser exhaustiva; son posibles muchas otras combinaciones de los aspectos, métodos, características y realizaciones de la presente divulgación.

La clave de un aspecto de la presente divulgación es el hecho de que las técnicas de determinación de ploidía que utilizan datos parentales ajustados por fases del objetivo pueden ser mucho más exactas que las que no utilizan tales datos. No obstante, en el contexto de la IVF no es algo trivial el ajuste por fases de los datos genotípicos medidos obtenidos de masa de tejido parental. Se describe en esta divulgación un método para diferenciar los datos parentales ajustados por fases de los datos genéticos parentales no ajustados por fases, junto con los datos genéticos no ajustados por fases de uno o más embriones, cero o más hermanos, y cero o más espermas. En este método para ajustar por fases los datos parentales se supone que los datos genéticos del embrión son euploides en un cromosoma determinado. Naturalmente, puede no ser posible determinar el estado de ploidía del cromosoma específico, para asegurarse de la euploidía, utilizando un método que requiera como entrada datos parentales ajustados por fases, antes de que se hayan ajustado por fases los datos genéticos, lo que representa un problema de “boot strapping”.

En algunas realizaciones de la presente divulgación, se divulga un método en el que se utiliza una técnica para la determinación del estado de ploidía, para realizar una determinación preliminar del estado de ploidía en un cromosoma determinado para un conjunto de células derivado de uno o más embriones. Luego el método que se describe aquí para determinar los datos parentales ajustados por fases puede ser ejecutado, utilizando solamente los datos de cromosomas embriónicos que han sido determinados, con elevada confianza utilizando el método preliminar, como euploide. Una vez ajustados por fases los datos parentales, se puede aplicar el método de determinación del estado de ploidía que requiere datos parentales ajustados por fases para obtener determinaciones de ploidía de alta precisión. Los resultados de este método pueden ser utilizados por sí solos, o pueden ser combinados con otros métodos de determinación de ploidía.

Algunas de las técnicas especializadas para la determinación del número de copias descritas en esta divulgación, por ejemplo, la técnica de la “presencia de homólogos”, se basan en datos genómicos parentales ajustados por fases. Algunos métodos de ajuste por fases de los datos, como algunos de los descritos en esta divulgación, funcionan bajo el supuesto de que los datos de entrada son de material genético euploide. Cuando el objetivo es un feto o un embrión, resulta especialmente probable que uno o más cromosomas no sean euploides. En una realización de la presente divulgación, una o varias técnicas de determinación de ploidía que no están basadas en los datos parentales ajustados por fases pueden ser utilizadas para determinar qué cromosomas son euploides, tales datos genéticos de esos cromosomas euploides pueden ser usados como parte de un algoritmo de determinación de alelos que proporciona datos parentales ajustados por fases, que podrán entonces ser usados en la técnica de determinación del número de copias que requieren datos parentales ajustados por fases.

En una realización de la presente divulgación, un método para determinar el estado de ploidía de por lo menos un cromosoma en un individuo objetivo incluye obtener datos genéticos del individuo objetivo, y de ambos progenitores del individuo objetivo, y de uno o más hermanos del individuo objetivo, donde los datos genéticos incluyen datos relativos por lo menos a un cromosoma; determinar un estado de ploidía de por lo menos un cromosoma en el individuo objetivo y en el hermano o hermanos del individuo objetivo utilizando una o más técnicas especializadas, donde ninguna de las técnicas especializadas requiere como entrada datos genéticos ajustados por fases; determinar datos genéticos ajustados por fases del individuo objetivo, y de los progenitores del individuo objetivo, y de uno o más hermanos del individuo objetivo, utilizando un método basado en la informática, y los datos genéticos obtenidos del individuo objetivo, y de los padres del individuo objetivo, y de uno o más hermanos del individuo objetivo que fueron determinados como euploides en ese cromosoma; y redeterminar el estado de ploidía de por lo menos un cromosoma del individuo objetivo, utilizando una o más técnicas especializadas, por lo menos alguna de las cuales requiere datos genéticos ajustados por fases como entrada, y los datos genéticos ajustados por fases determinados del individuo objetivo, y de los progenitores del individuo objetivo, y de uno o más hermanos del individuo objetivo. En una realización, la determinación del estado de ploidía puede ser llevado a cabo en el contexto de la fertilización in vitro, y donde el individuo objetivo es un embrión. El estado de ploidía determinado del cromosoma en el individuo objetivo puede ser utilizado para tomar una decisión clínica respecto al individuo objetivo.

Primero, los datos genéticos pueden obtenerse del individuo objetivo y de los progenitores del individuo objetivo, y posiblemente de uno o más individuos que son hermanos del individuo objetivo. Estos datos genéticos de individuos pueden ser obtenidos de diversas formas, que se describen en otra parte de esta divulgación. Los datos genéticos del individuo objetivo pueden ser medidos utilizando herramientas o técnicas tomadas de un grupo que incluye, entre otras, Sondas de Inversión Molecular (MIP), Microarrays de Genotipado, el Ensayo de Genotipado de SNP TaqMan, el Sistema de Genotipado Illumina, otros ensayos de genotipado, hibridación in situ fluorescente (FISH), secuenciación, otras plataformas de genotipado de alto rendimiento, y combinaciones de lo anterior. Los datos genéticos del individuo objetivo pueden medirse analizando sustancias tomadas de un grupo que incluye, entre otras, una o más células diploides del individuo objetivo, una o más células haploides del individuo objetivo, uno o

más blastómeros del individuo objetivo, material genético extracelular hallado en el individuo objetivo, material genético extracelular del individuo objetivo hallado en sangre materna, células del individuo objetivo halladas en sangre materna, material genético del que se sabe que procede del individuo objetivo y combinaciones de lo anterior. Los datos genéticos del individuo relacionado pueden ser medidos analizando sustancias tomadas de un grupo que incluye, entre otras, masa de tejido diploide del individuo relacionado, una o más células diploides del individuo relacionado, una o más células haploides tomadas del individuo relacionado, uno o más embriones creados a partir de un gameto o gametos del individuo relacionado, uno o más blastómeros tomados de uno de tales embriones, material genético extracelular hallado en el individuo relacionado, material genético del que se sabe que procede del individuo relacionado y combinaciones de lo anterior.

Segundo, puede crearse un conjunto de por lo menos una hipótesis de estado de ploidía para uno o más cromosomas del individuo objetivo y de los hermanos. Cada hipótesis del estado de ploidía puede referirse a un posible estado de ploidía del cromosoma de los individuos.

Tercero, utilizando una o más de las técnicas especializadas, como las comentadas en esta divulgación, puede determinarse una probabilidad estadística para cada hipótesis de estado de ploidía en el conjunto. En este paso, la técnica especializada es una técnica especializada que no requiere datos genéticos ajustados por fases como entrada. Algunos ejemplos de técnicas especializadas que no requieren datos genéticos ajustados por fases como entrada incluyen, entre otras, la técnica de permutación, la técnica de media de cromosoma completo, y la técnica de presencia parental. La matemática en que se basan las diversas técnicas especializadas apropiadas se describe en otra parte en esta divulgación.

Cuarto, si en el tercer paso se ha utilizado más de un método especializado, el conjunto de probabilidades determinadas puede entonces combinarse y normalizarse. El conjunto de los productos de las probabilidades de cada hipótesis en el conjunto de hipótesis resulta entonces como las probabilidades combinadas de las hipótesis.

Quinto, el estado de ploidía más probable, y para cada uno del o de los individuos hermanos, se determina como el estado de ploidía asociado con la hipótesis cuya probabilidad es la mayor.

Sexto, un método basado en la informática, como el método de determinación de alelos divulgado en este documento, u otros aspectos del método de PARENTAL SUPPORT™, junto con datos genéticos parentales no ordenados, y los datos genéticos de hermanos de los que se ha hallado en el quinto paso que son euploides, en ese cromosoma, puede ser utilizado para determinar el estado alélico más probable del individuo objetivo, y de los individuos hermanos. En algunas realizaciones, los individuos objetivo pueden ser tratados de igual modo, algorítmicamente, que los hermanos. En algunas realizaciones, el estado alélico de un hermano puede ser determinado dejando que el individuo objetivo actúe como hermano, y el hermano como un objetivo. En algunas realizaciones, el método basado en la informática debería también proporcionar el estado alélico de los progenitores, incluyendo los datos genéticos haplotípicos. En algunas realizaciones de la presente divulgación, el método basado en la informática utilizado puede determinar también el estado genético ajustado por fases más probable del o los progenitores y de los otros hermanos.

Séptimo, un nuevo conjunto de por lo menos una hipótesis de estado de ploidía puede crearse para uno o más cromosomas del individuo objetivo y de los hermanos. Como anteriormente, cada una de las hipótesis de estado de ploidía puede referirse a un posible estado de ploidía del cromosoma de los individuos.

Octavo, utilizando una o más de las técnicas especializadas, como las comentadas en esta divulgación, se puede determinar una probabilidad estadística para cada hipótesis de estado de ploidía en el conjunto. En este paso, por lo menos una de las técnicas especializadas es una técnica especializada que requiere como entrada datos genéticos ajustados por fases, como la técnica de la “presencia de homólogos”.

Noveno, el conjunto de probabilidades determinadas puede ser entonces combinado como se describe en el cuarto paso. Por último, el estado de ploidía más probable para el individuo objetivo, en ese cromosoma, se determina como el estado de ploidía que está asociado con la hipótesis de mayor probabilidad. En algunas realizaciones, el estado de ploidía se determinará solamente si la hipótesis de mayor probabilidad supera un determinado umbral de confianza y/o probabilidad.

En una realización de este método, en el tercer paso pueden utilizarse las tres técnicas especializadas siguientes en la determinación del estado de ploidía inicial: la técnica de permutación, la técnica de la media de cromosoma completo, y la técnica de la presencia parental. En una realización de la presente divulgación, en el paso octavo se puede utilizar el siguiente conjunto de técnicas especializadas en la determinación de ploidía final: la técnica de permutación, la técnica de la media de cromosoma completo, la técnica de la presencia parental, y la técnica de la presencia de homólogos. En algunas realizaciones de la presente divulgación se pueden utilizar distintos conjuntos de técnicas especializadas en el tercer paso. En algunas realizaciones de la presente divulgación pueden utilizarse en el octavo paso diferentes conjuntos de técnicas especializadas. En una realización de la presente divulgación, es posible combinar varios de los aspectos de la presente divulgación, de forma que se podría realizar tanto la determinación de alelos, como la determinación de aneuploidía utilizando un algoritmo.

En una realización de la presente divulgación, el método divulgado se utiliza para determinar el estado genético de uno o más embriones, a los efectos de la selección de embriones en el contexto de la IVF. Esto puede incluir la recogida de ovocitos de la futura madre, y fertilizar esos ovocitos con esperma del futuro padre para crear uno o más

embriones. Esto puede incluir realizar una biopsia de embrión para aislar un blastómero de cada uno de los embriones. Puede incluir amplificar y genotipar los datos genéticos de cada uno de los blastómeros. Esto puede incluir obtener, amplificar y genotipar una muestra de material genético diploide de cada uno de los progenitores, así como uno o más espermas individuales del padre. Puede incluir incorporar los datos diploides y haploides medidos de la madre y del padre, junto con los datos genéticos medidos del embrión de interés en un conjunto de datos. Puede incluir utilizar uno o más de los métodos estadísticos divulgados en esta patente para determinar el estado más probable del material genético en el embrión considerando los datos genéticos medidos o determinados. Puede incluir la determinación del estado de ploidía del embrión de interés. Puede incluir la determinación de la presencia de diversos alelos vinculados a enfermedad conocidos en el genoma del embrión. Puede incluir realizar predicciones fenotípicas sobre el embrión. Puede incluir generar un informe que se envía al médico de la pareja, para que puedan tomar una decisión informada sobre qué embrión o embriones transferir a la futura madre.

Otro ejemplo podría ser una situación donde una mujer de 44 años sometida a IVF tiene problemas para concebir. La pareja hace que se recojan sus ovocitos y se fertilicen con esperma del hombre, produciendo nueve embriones viables. Se recoge un blastómero de cada embrión, y los datos genéticos de los blastómeros se miden utilizando una Illumina Infmium Bead Array. Mientras tanto, se miden los datos diploides de tejido tomado de ambos progenitores utilizando también la Illumina Infmium Bead Array. Se miden los datos haploides del esperma del padre utilizando el mismo método. El método divulgado aquí se aplica a los datos genéticos de los nueve blastómeros, a los datos genéticos diploides maternos y paternos, y a tres espermas del padre. Los métodos descritos aquí se utilizan para limpiar y ajustar por fases todos los datos genéticos utilizados como entradas, y para realizar determinaciones de ploidía para todos los cromosomas en todos los embriones, con altas confianzas. Se observa que seis de los nueve embriones son aneuploides, y tres son euploides. Se genera un informe que presenta estos diagnósticos y es enviado al médico. El médico, junto con los futuros padres, decide transferir dos de los tres embriones euploides, uno de los cuales se implanta en el útero materno.

Otro ejemplo puede referirse a una mujer gestante que ha sido inseminada artificialmente con el esperma de un donante, y está embarazada. Desea minimizar el riesgo de que el feto que lleva tenga una enfermedad genética. Un especialista le extrae sangre y se aplican las técnicas descritas en esta divulgación para aislar tres glóbulos rojos fetales nucleados, y también se toma una muestra de tejido de la madre y del padre. El material genético del feto y de la madre y el padre es amplificado convenientemente, y se procede a su genotipado utilizando la matriz Illumina Infmium Bead, y los métodos descritos aquí para limpiar y ajustar por fases el genotipo parental y el fetal con gran precisión, así como para realizar determinaciones de ploidía para el feto. Se averigua que el feto es euploide, y se prevén susceptibilidades fenotípicas por el genotipo fetal reconstruido, y se genera un informe que se envía al médico de la madre, para que puedan decidir qué acciones deben emprender.

Otro ejemplo podría ser una situación donde un criador de caballos de carreras desea incrementar la probabilidad de que los potros engendrados por su caballo campeón sean también campeones. Hace que la yegua deseada sea fecundada por el semental por IVF, y utiliza los datos genéticos del semental y la yegua para limpiar los datos genéticos medidos de los embriones viables. Los datos genéticos embriónicos limpios permiten al criador seleccionar para su implantación los embriones con mayores probabilidades de producir un caballo de carreras deseable.

Un método para determinar un estado de ploidía de por lo menos un cromosoma en un individuo objetivo incluye obtener datos genéticos del individuo objetivo y de uno o más individuos relacionados; crear un conjunto de por lo menos una hipótesis de estado de ploidía para cada cromosoma del individuo objetivo; determinar una probabilidad estadística para cada hipótesis de estado de ploidía en el conjunto, considerando los datos genéticos obtenidos y utilizando una o más técnicas especializadas; combinar, para cada hipótesis de estado de ploidía, las probabilidades estadísticas determinadas por la o las técnicas especializadas; y determinar el estado de ploidía de cada uno de los cromosomas en el individuo objetivo, en base a las probabilidades estadísticas combinadas de cada una de las hipótesis de estado de ploidía.

Un método para determinar datos alélicos de uno o más individuos objetivo, y uno o ambos progenitores de los individuos objetivo, en un conjunto de alelos, incluye obtener datos genéticos del o de los individuos objetivo y de uno o de ambos progenitores; crear un conjunto de por lo menos una hipótesis alélica para cada uno de los alelos de los individuos objetivo y para cada uno de los alelos de los progenitores; determinar una probabilidad estadística para cada hipótesis alélica en el conjunto considerando los datos genéticos obtenidos; y determinar el estado alélico para cada uno de los alelos en uno o más individuos objetivo y el o los progenitores basado en las probabilidades estadísticas de cada una de las hipótesis alélicas.

Un método para determinar un estado de ploidía de por lo menos un cromosoma en un individuo objetivo incluye obtener datos genéticos del individuo objetivo, de ambos progenitores del individuo objetivo, y de uno o más hermanos del individuo objetivo, donde los datos genéticos incluyen datos relacionados con por lo menos un cromosoma; determinar un estado de ploidía de por lo menos un cromosoma en el individuo objetivo y en uno o más hermanos del individuo objetivo utilizando una o más técnicas especializadas, donde ninguna de las técnicas especializadas requiere datos genéticos ajustados por fases como entrada; determinar datos genéticos ajustados por fases del individuo objetivo, de los progenitores del individuo objetivo, y del o de los hermanos del individuo objetivo, utilizando un método basado en la informática, y los datos genéticos obtenidos del individuo objetivo, de los progenitores del individuo objetivo y del hermano o hermanos del individuo objetivo que fueron determinados como

euploides en ese cromosoma; y redeterminar el estado de ploidía de por lo menos un cromosoma del individuo objetivo, utilizando una o más técnicas especializadas, por lo menos una de las cuales requiere datos genéticos ajustados por fases como entrada, y los datos genéticos ajustados por fases determinados del individuo objetivo, de los progenitores del individuo objetivo y del hermano o hermanos del individuo objetivo.

REIVINDICACIONES

1. Un método para la determinación de un estado de ploidía de por lo menos un cromosoma en un individuo objetivo, comprendiendo este método:

obtener datos genéticos del individuo objetivo y de uno o más individuos relacionados utilizando un medio para medir datos genéticos, donde el medio se selecciona de un grupo que comprende sondas de inversión molecular, microarrays de genotipado, un ensayo de genotipado, hibridación de fluorescencia in situ (FISH), secuenciación, otras plataformas de genotipado de alto rendimiento, y combinaciones de lo anterior, y donde los datos genéticos son las respuestas medidas en varios loci de polimorfismo de un solo nucleótido (SNP) en por lo menos un cromosoma, comprendiendo el o los individuos relacionados mencionados uno o ambos progenitores del individuo objetivo;

crear un conjunto de por lo menos una hipótesis de estado de ploidía para por lo menos un cromosoma del individuo objetivo, donde cada una de las hipótesis del estado de ploidía es un posible estado de ploidía de por lo menos un cromosoma, donde 0, 1, o 2 copias del cromosoma proceden de cada progenitor;

utilizar dos o más técnicas especializadas que son algoritmos aplicados sobre los datos genéticos obtenidos para determinar, para cada técnica especializada utilizada, una probabilidad estadística de cada hipótesis de estado de ploidía en el conjunto, considerando los datos genéticos obtenidos, donde las técnicas especializadas se seleccionan de entre:

la técnica de la presencia de homólogos, técnica que utiliza datos genéticos obtenidos de ambos progenitores, donde uno de los progenitores es heterocigoto en un SNP, y el otro es homocigoto en ese SNP, donde la técnica de presencia de homólogos comprende:

1 ajustar por fases los datos genéticos obtenidos de los progenitores y calcular los umbrales de ruido por cromosoma;

2 segmentar por lo menos un cromosoma;

3 calcular las tasas de pérdidas de SNP por segmento para los genotipos parentales de interés;

4 calcular las tasas de pérdidas de SNP para cada progenitor en por lo menos un cromosoma y las probabilidades de hipótesis en cada segmento;

5 combinar las probabilidades en los segmentos de cromosoma de producir una probabilidad de datos, dada la hipótesis de cadena parental para cromosomas completos; y

6 comprobar las determinaciones no válidas y calcular una probabilidad para cada hipótesis de estado de ploidía;

la técnica de permutación, técnica que compara la relación entre distribuciones de los datos genéticos obtenidos del individuo objetivo para distintos genotipos parentales utilizando un algoritmo estadístico para determinar la probabilidad de cada hipótesis de estado de ploidía, considerando los datos genéticos obtenidos; y

la técnica de la presencia parental, que detecta, independientemente para cada progenitor, para un cromosoma determinado, si hay o no contribución del genoma de ese progenitor, basado en distancias entre conjuntos de genotipos parentales en el punto más amplio en las curvas de función de distribución acumulativa, que representa gráficamente las distribuciones observadas de los datos genéticos obtenidos para distintos genotipos parentales, y asigna probabilidades a cada hipótesis de estado de ploidía calculando un resumen estadístico para cada progenitor y comparando con modelos de datos para casos en los que un cromosoma parental está presente, y casos en los que no está presente un cromosoma parental;

combinar para cada hipótesis de estado de ploidía, las probabilidades estadísticas determinadas por las dos o más técnicas especializadas, para determinar probabilidades estadísticas combinadas; y

determinar el estado de ploidía para por lo menos un cromosoma en el individuo objetivo, basado en las probabilidades estadísticas combinadas de cada una de las hipótesis de estado de ploidía, donde el estado de ploidía con la mayor probabilidad estadística combinada se determina que es el estado de ploidía de por lo menos un cromosoma.

2. El método de la reivindicación 1, donde los individuos relacionados comprenden además uno o más abuelos del individuo objetivo, uno o más hermanos del individuo objetivo y combinaciones de lo anterior.

3. El método de la reivindicación 1, donde los datos genéticos obtenidos comprenden por lo menos uno de: polimorfismos de un solo nucleótido medidos de una matriz de genotipado, datos de secuencias de ADN, y combinaciones de ellos.

4. El método de la reivindicación 1, donde el individuo objetivo es un embrión, y la determinación del estado de ploidía se realiza a los efectos de selección de embrión durante la fertilización in vitro.

5. El método de la reivindicación 1, donde el individuo objetivo es un feto, y la determinación del estado de ploidía se realiza a los efectos del diagnóstico prenatal no invasivo.

6. El método de la reivindicación 1, donde los datos genéticos obtenidos no están ajustados por fases, y donde los individuos relacionados comprenden a ambos progenitores del individuo objetivo, y donde el método comprende, además:

determinar los datos genéticos ajustados por fases de ambos progenitores del individuo objetivo utilizando un método basado en la informática; y

determinar los datos genéticos ajustados por fases del individuo objetivo utilizando un método basado en la informática.

7. El método de la reivindicación 1, donde los datos genéticos obtenidos comprenden datos genéticos reconstruidos ajustados por fases de uno o ambos progenitores del individuo objetivo.

8. El método de la reivindicación 1, donde por lo menos una de las técnicas especializadas es específica para un cromosoma sexual.

9. El método de la reivindicación 1, donde la determinación del estado de ploidía de cada uno de los cromosomas en el individuo objetivo incluye el cribado para un estado cromosómico seleccionado del grupo compuesto por euploidía, nulisomía, monosomía, disomía uniparental, trisomía, error de copia emparejada, error de copia no emparejada, tetrasomía, otra aneuploidía, translocación no balanceada, supresiones, inserciones, mosaicismo y combinaciones de lo anterior.

10. El método de la reivindicación 1, donde se utilizan tres técnicas especializadas: la técnica de la presencia de homólogos, la técnica de permutación y la técnica de la presencia parental, y donde, para cada hipótesis de estado de ploidía, las probabilidades estadísticas determinadas por esas tres técnicas especializadas se combinan para determinar probabilidades estadísticas combinadas.

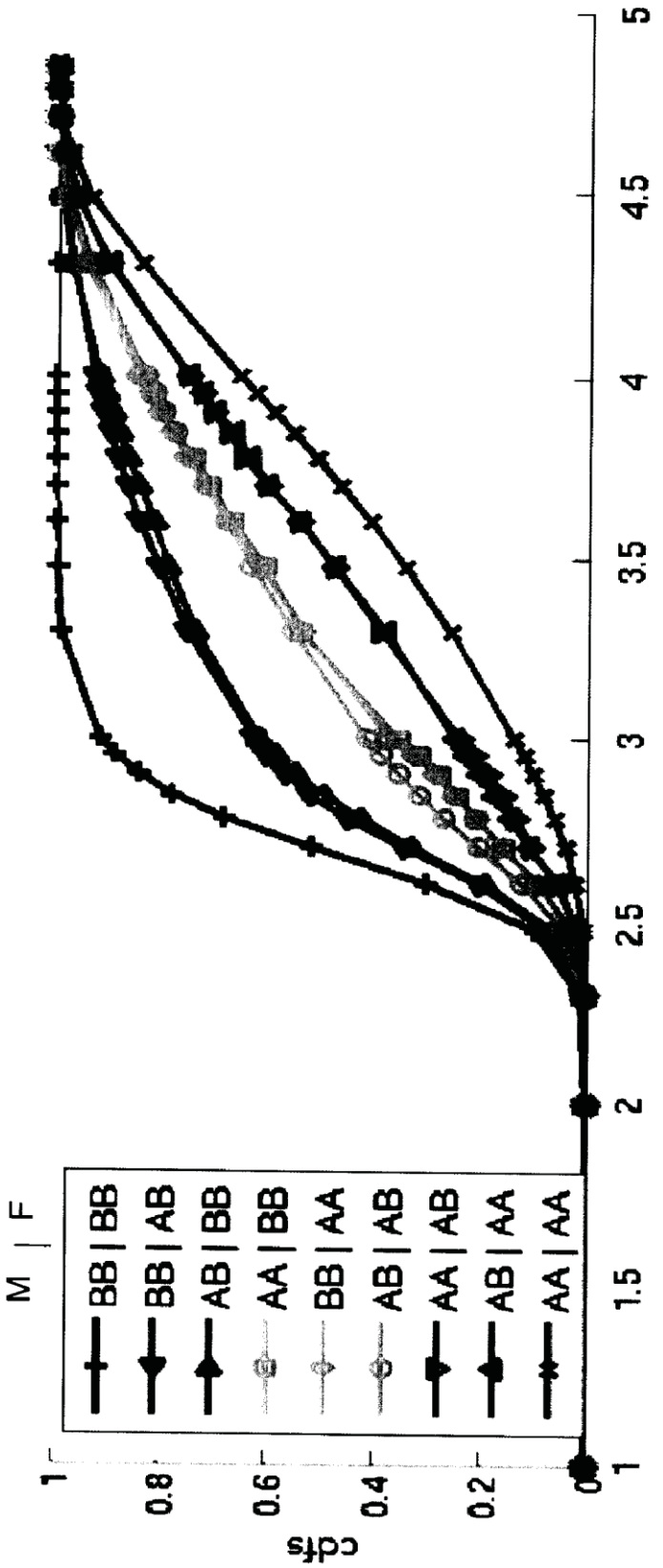
M | F
BB | BB
BB | AB
AB | BB
AA | BB
BB | AA
AB | AB
AA | AB
AB | AA
AA | AA
cdfs
0
0.2
0.4
0.6
0.8
1
1
1.5
2
2.5
3
3.5
4
4.5
5
Respuestas log en el canal A

Fig 1

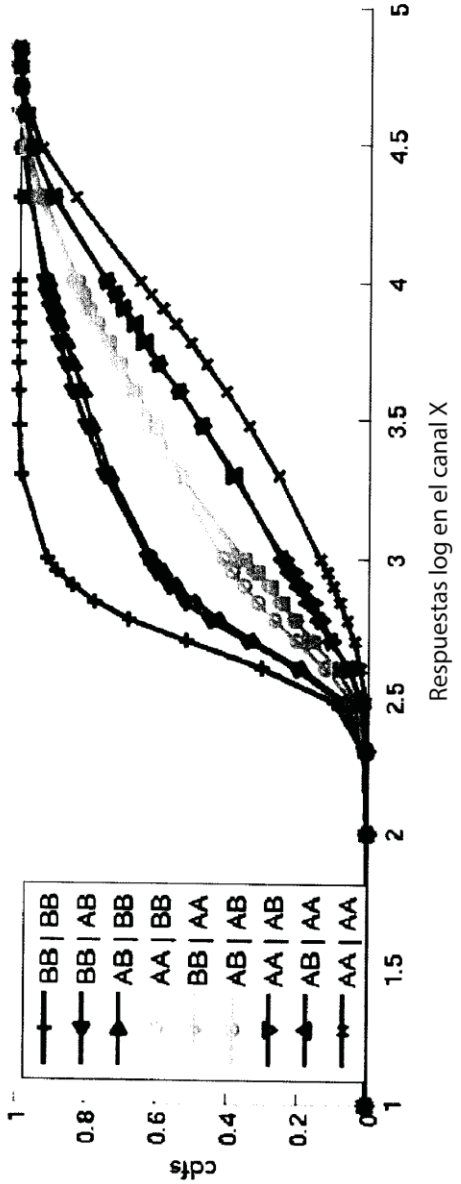
cdfs
Respuestas log en el canal X
BB | BB
BB | AB
AB | BB
AA | BB
BB | AA
AB | AB
AA | AB
AB | AA
AA | AA

Fig 2A

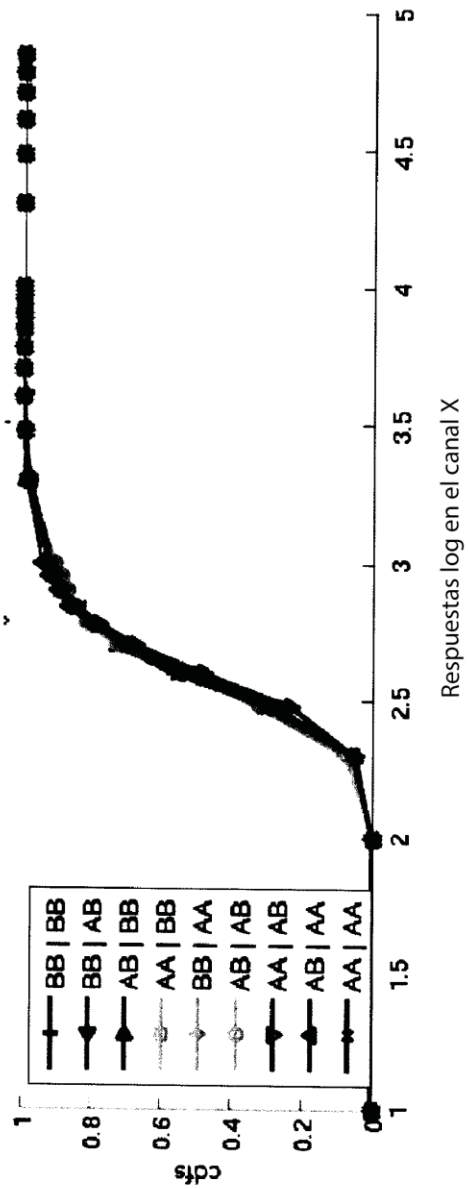
cdfs
Respuestas log en el canal X
BB | BB
BB | AB
AB | BB
AA | BB
BB | AA
AB | AB
AA | AB
AB | AA
AA | AA

Fig 2B

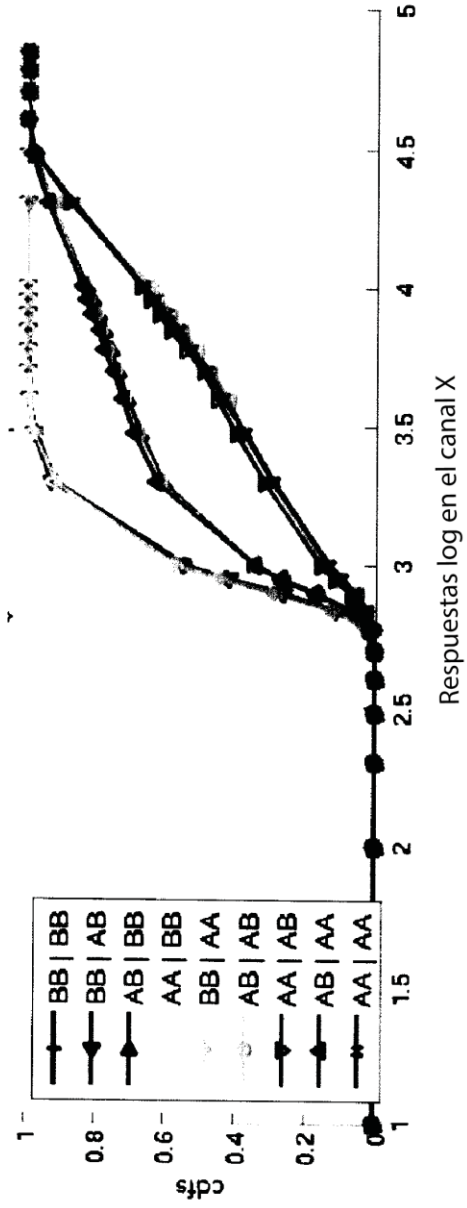
cdfs
Respuestas log en el canal X
BB | BB
BB | AB
AB | BB
AA | BB
BB | AA
AB | AB
AA | AB
AB | AA
AA | AA

Fig 2C

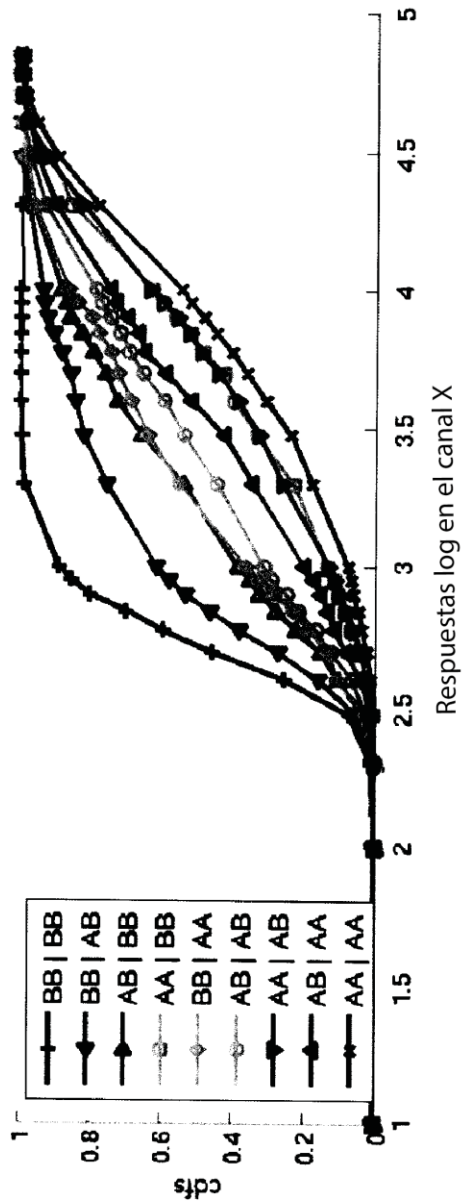
cdfs
Respuestas log en el canal X
BB | BB
BB | AB
AB | BB
AA | BB
BB | AA
AB | AB
AA | AB
AB | AA
AA | AA

Fig 2D

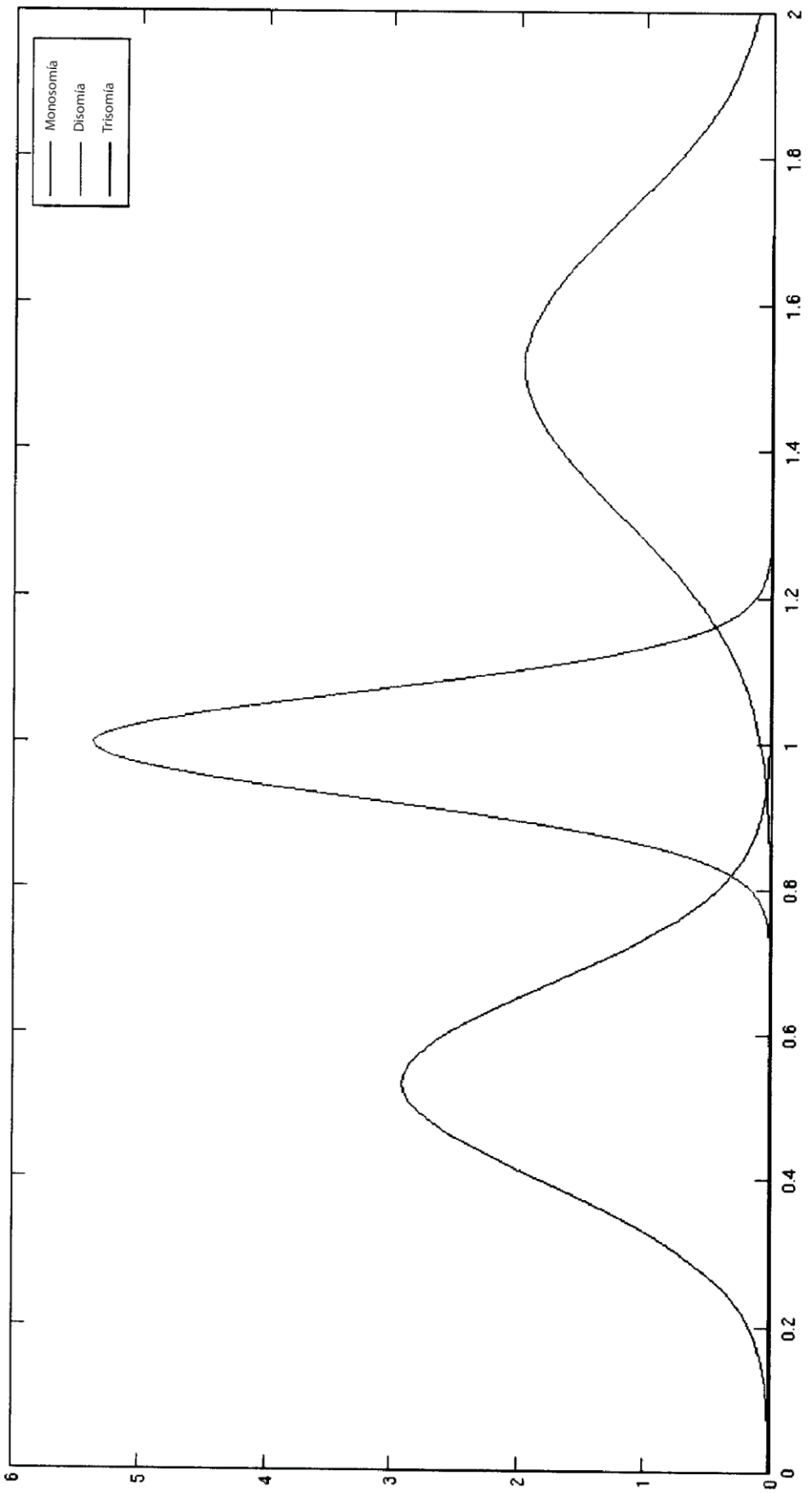
Monosomía
Disomía
Trisomía
0
1
2
3
4
5
6
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

Fig 3

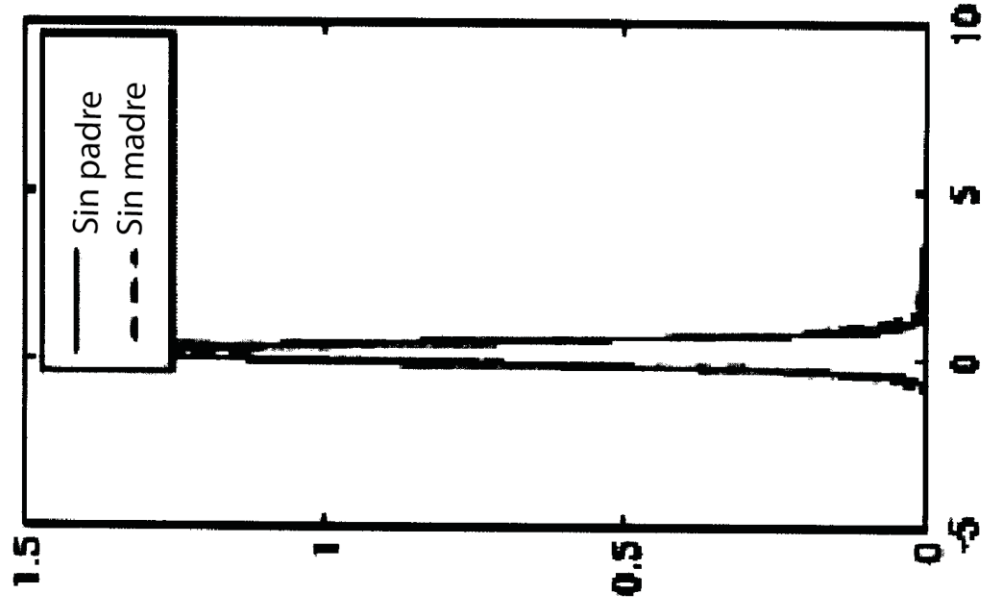
Sin padre
Sin madre
1.5
1
0.5
0
-5
0
5
10
x

Fig 4B

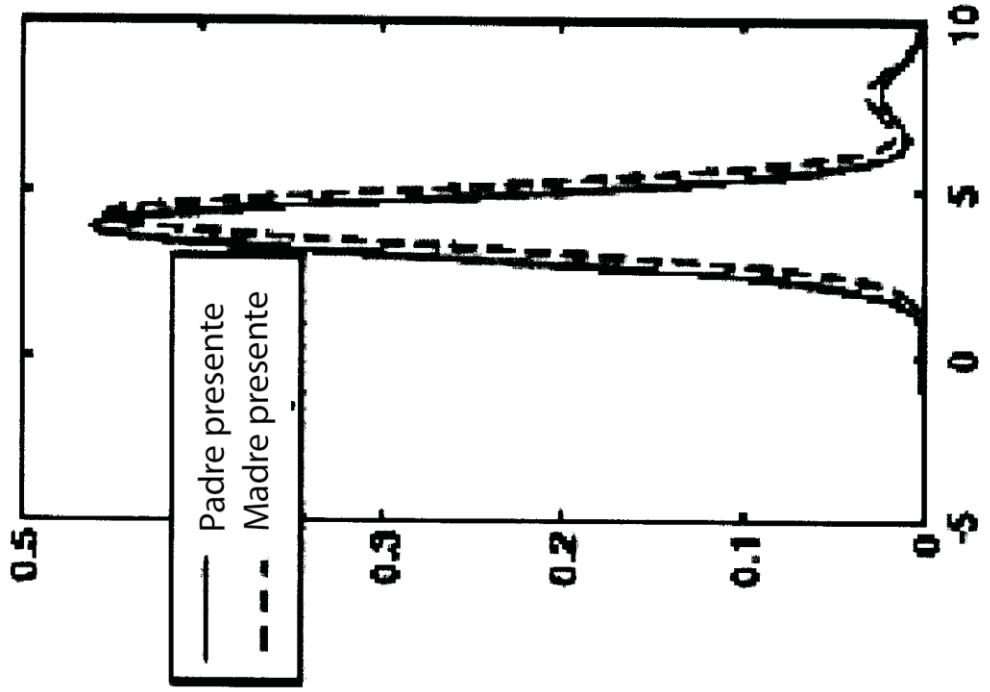
Padre presente
Madre presente
0.5
0.3
0.2
0.1
0
-5
0
5
10
x

Fig 4A

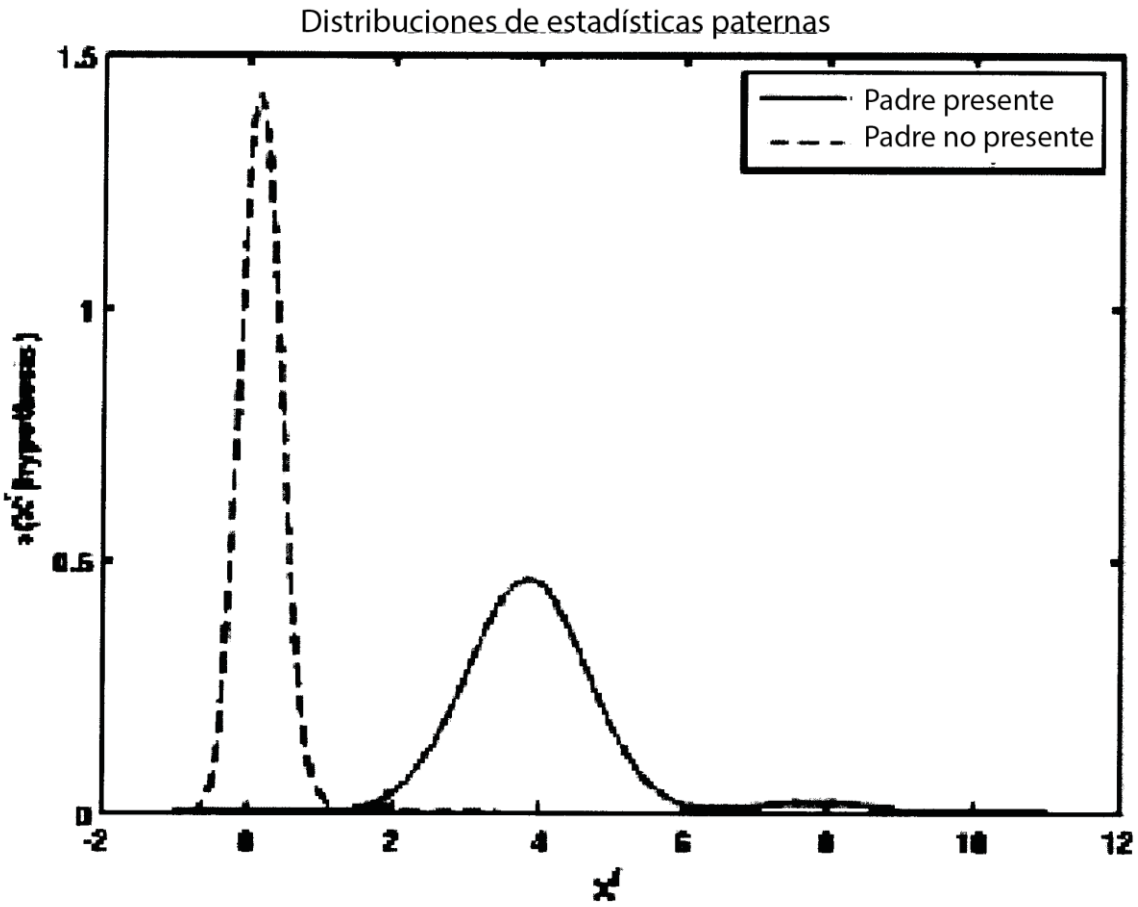
Distribuciones de estadísticas paternas
Padre presente
Padre no presente
1.5
1
0.5
0
-2
0
2
4
6
8
10
12

Fig 5

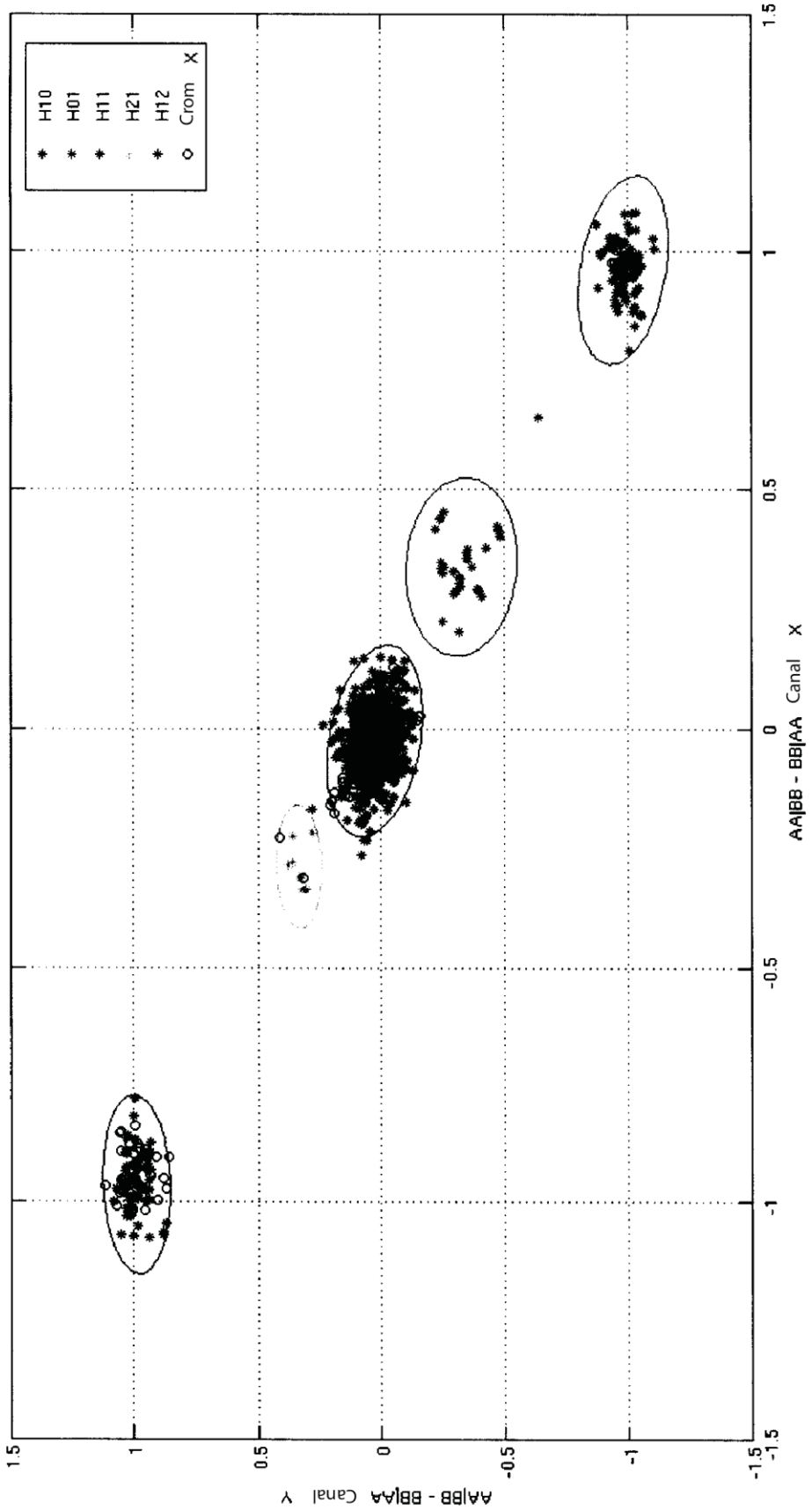
H10
H01
H11
H21
H12
Crom X
AA|BB - BB|AA Canal Y
AA|BB - BB|AA Canal X
1.5
1
0.5
0
-0.5
-1
-1.5

Fig 6

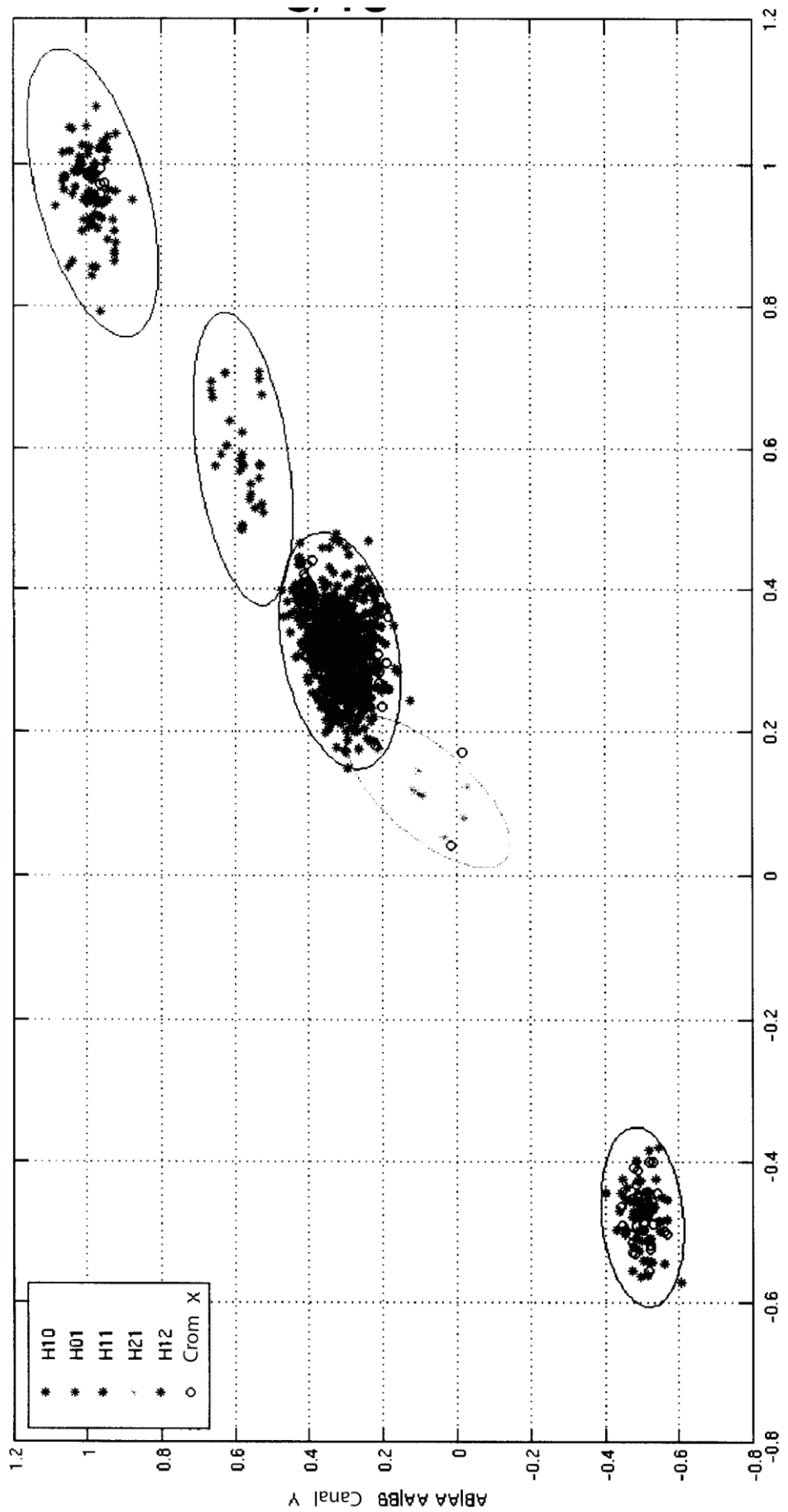
H10
H01
H11
H21
H12
Crom X
AB|BB BB|AA Canal X
ABIAA AAIBB Canal Y

Fig 7

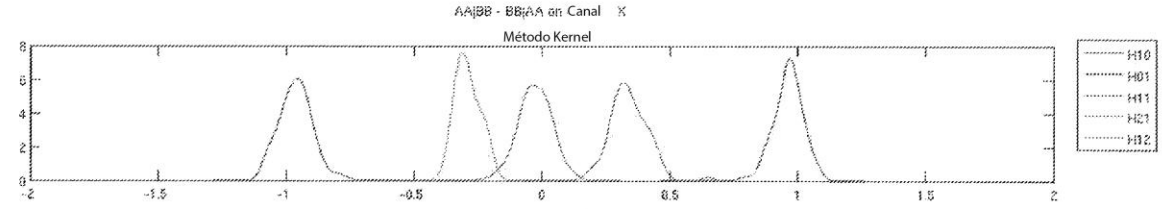

Fig 8A

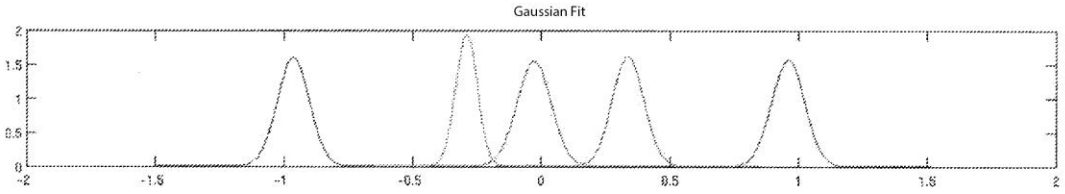

Fig 8B

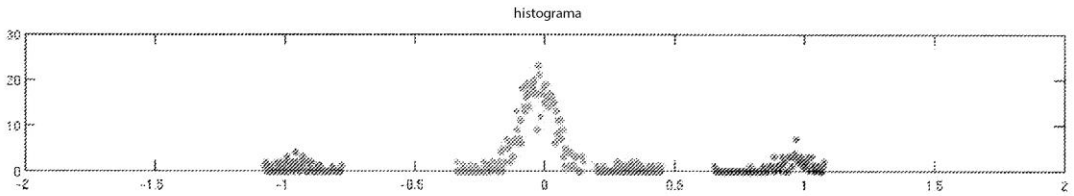

Fig 8C

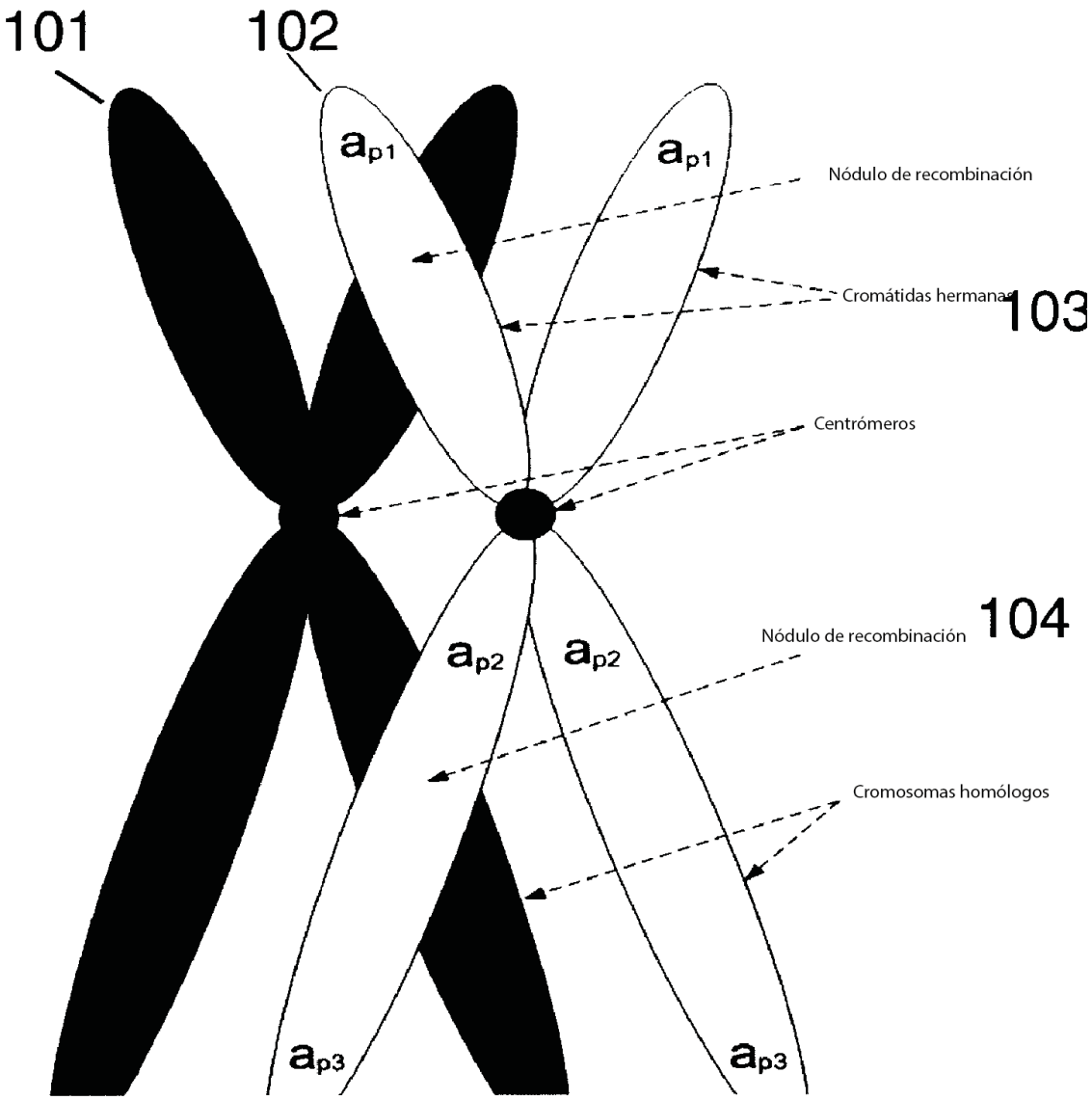
101
102
a_{p1}
a_{p1}
Nódulo de recombinación
Cromátidas hermanas
103
Centrómeros
a_{p2}
a_{p2}
Nódulo de recombinación
104
Cromosomas homólogos
a_{p3}
a_{p3}

Fig 9

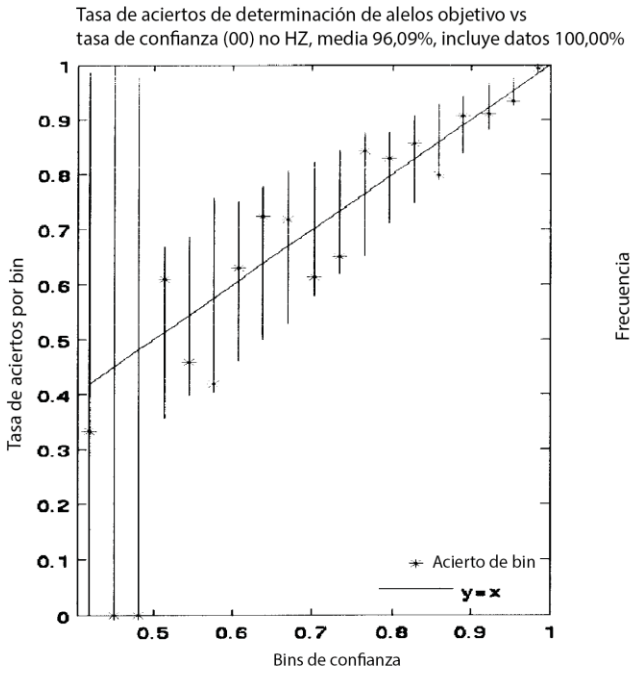

Fig 10A

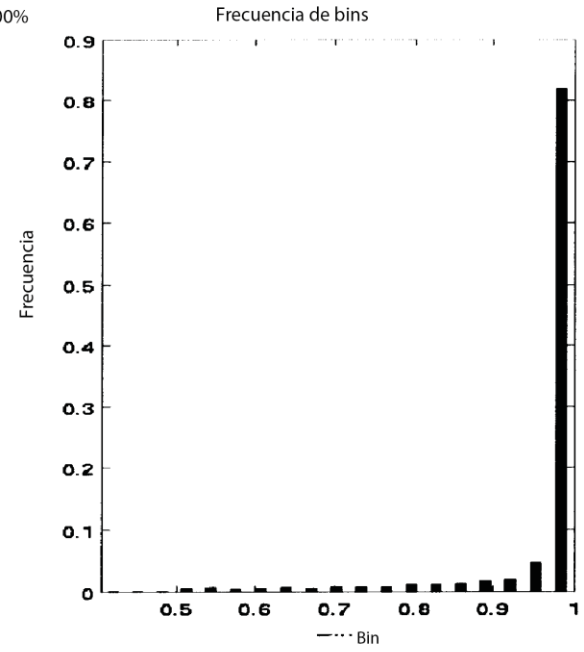

Fig 10B

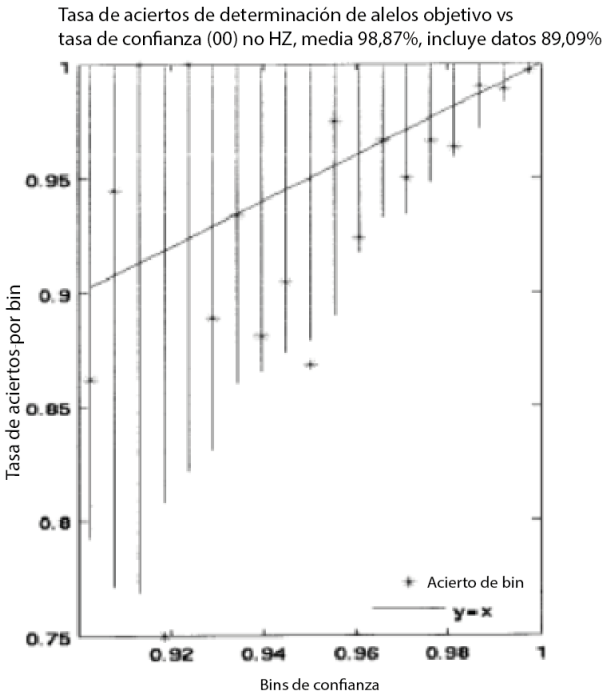

Fig 11A

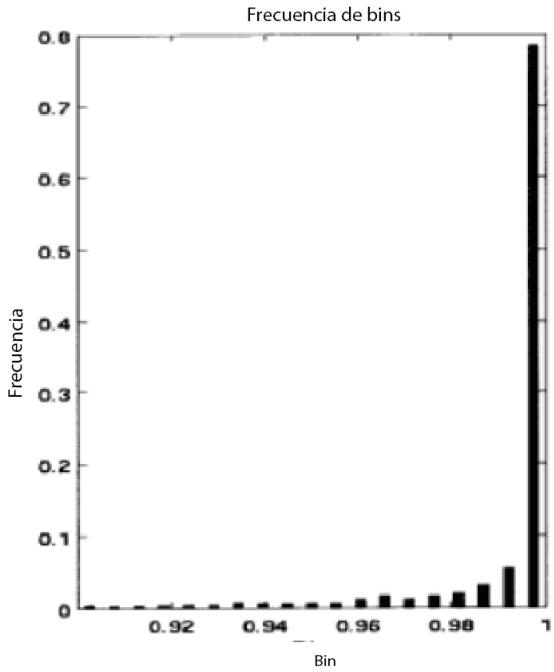

Fig 11B

Fig 12A

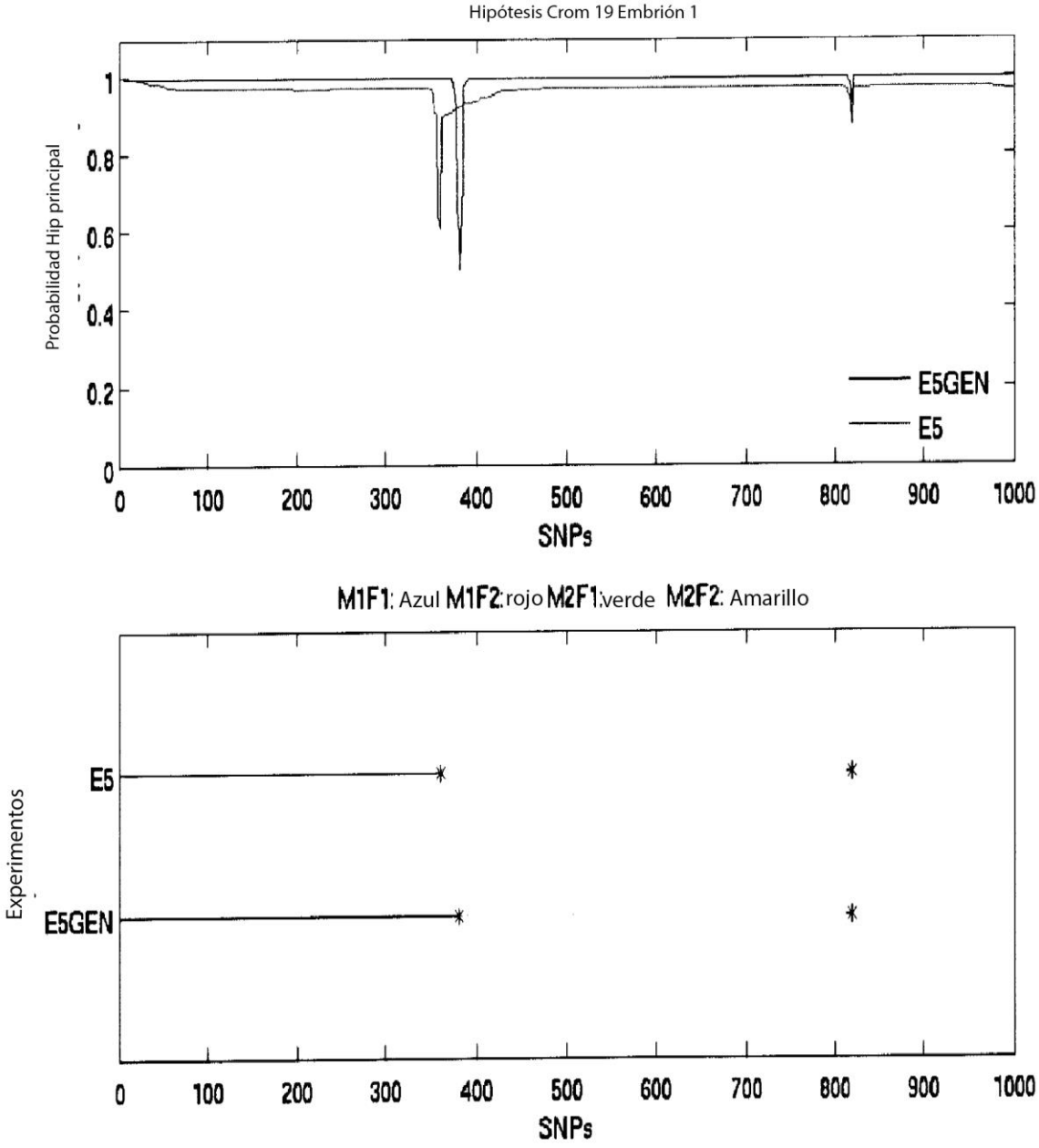

Fig 12B