(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 762 942**

(51) Int. CI.:

G16B 10/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **14.10.2015 PCT/US2015/055579**

(87) Fecha y número de publicación internacional: **21.04.2016 WO16061260**

(96) Fecha de presentación y número de la solicitud europea: **14.10.2015 E 15850218 (7)**

(97) Fecha y número de publicación de la concesión europea: **04.12.2019 EP 3207481**

(54) Título: **Reducción del error en las relaciones genéticas predichas**

(30) Prioridad:

14.10.2014 US 201462063849 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
26.05.2020

(73) Titular/es:

ANCESTRY.COM DNA, LLC (100.0%)
1300 W. Traverse Parkway
Lehi, UT 84043, US

(72) Inventor/es:

BARBER, MATHEW J.;
WANG, YONG;
NOTO, KEITH D.;
CHAHINE, KENNETH G. y
BALL, CATHERINE ANN

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

DESCRIPCIÓN

Reducción del error en las relaciones genéticas predichas

Antecedentes

1. Campo

Las realizaciones divulgadas se refieren a productos de programas informáticos, sistemas y métodos utilizados para identificar individuos en una población que están relacionados ancestralmente con base en los datos genéticos de los individuos.

2. Descripción de la técnica relacionada

Aunque los seres humanos son, genéticamente hablando, casi completamente idénticos, pequeñas diferencias en nuestro ADN son las responsables de gran parte de la variación entre individuos. Las extensiones de ADN que se determina que son relevantes para algún propósito se denominan haplotipos. Los haplotipos se identifican con base en polimorfismos de nucleótido único (SNP) consecutivos de longitud variable. Ciertos haplotipos compartidos por individuos sugieren una relación familiar entre esos individuos con base en un principio conocido como identidad por descendencia (IBD).

Debido a que identificar segmentos de ADN de IBD entre pares de individuos genotipados es útil en muchas aplicaciones, se han desarrollado numerosos métodos para realizar análisis de IBD (Purcell et al. 2007, Gusev A. et al., The Architecture of Long-Range Haplotypes Shared within and across Populations, Mol. Biol. Evol., 29(2):473-86, 2012; Browning S.R. y Browning B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, American Journal of Human Genetics, 91:1084-96, 2007; Browning S.R. y Browning B.L., Identity by descent between distant relatives: detection and applications, Annu. Rev. Genet., 46:617-33, 2012). Sin embargo, estos enfoques no se adaptan al crecimiento continuo de conjuntos de datos muy grandes. Por ejemplo, la implementación de GERMLINE existente está diseñada para tomar un único archivo de entrada que contiene todos los individuos para compararlos entre sí. Si bien es apropiado para el caso en el que todas las muestras se genotipan y analizan simultáneamente, este enfoque no es práctico cuando las muestras se recogen de forma incremental. El conjunto de programas GERMLINE (Gusev A. et al., Whole population, genome-wide mapping of hidden relatedness, Genome Res., 19:318-26, 2009) ofrece un "filtro ibs", que elimina los emparejamientos muy frecuentes (definidos por el cromosoma, así como la posición inicial y final en el cromosoma). Al igual que el enfoque de descubrimiento de segmentos emparejados de GERMLINE, el enfoque del "filtro ibs" está diseñado para ser rápido y, en consecuencia, es relativamente simplista. Los más precisos de estos métodos, como la IBD refinada, son mucho más precisos que el "filtro ibs" de GERMLINE, pero no se adaptan computacionalmente y sería difícil de integrar en un tubo analítico incluso si lo hicieran. Existen muchos métodos existentes que evalúan la evidencia de un segmento emparejado no solo por la anchura en centimorgan, como se hace dentro de GERMLINE. Los ejemplos incluyen IBD refinada, IBD rápida, SLRP y PARENTE. Destacan que las diferencias entre estos enfoques son una compensación entre la complejidad del modelo y la velocidad computacional (y la factibilidad).

Sumario

Se divulgan métodos, sistemas y productos de programas informáticos para estimar un grado de parentesco ancestral entre dos individuos. Los productos del programa informático incluyen TIMBER. Para estimar el parentesco ancestral de dos individuos, los métodos incluyen recibir datos de haplotipos de una población de individuos. Los datos de haplotipos incluyen una pluralidad de marcadores genéticos que se comparten entre los individuos de la población. Los datos de haplotipos luego se dividen en ventanas de segmento con base en los marcadores genéticos. En algunas realizaciones, los marcadores genéticos incluyen polimorfismos de un solo nucleótido (SNP) y los datos de haplotipos se dividen en ventanas de segmento *K* que incluyen un número igual *d* de SNP. En algunas realizaciones, los datos de haplotipos se dividen en 4105 ventanas de segmento de 96 SNP.

Para cada individuo, el método incluye emparejar segmentos de los datos de haplotipos que son idénticos entre el individuo y cualquier otro individuo en la población, en donde el emparejamiento se basa en los marcadores genéticos. Cada segmento emparejado tiene una primera anchura en centimorgan (cM) que excede una anchura cM umbral. En algunas realizaciones, la anchura cM umbral es de 5 cM. Cada segmento emparejado es parte de una o más de las ventanas de segmento. Los segmentos emparejados en cada ventana de segmento se cuentan. El recuento de segmentos emparejados en una ventana de segmento también se conoce como recuento de emparejamientos por ventana.

Para cada individuo, el método incluye estimar un peso asociado con cada ventana de segmento en función del recuento de segmentos emparejados en la ventana de segmento asociada. En algunas realizaciones, el peso asociado con una ventana de segmento disminuye a medida que aumenta el recuento de segmentos emparejados. El beneficio de disminuir los pesos para aumentar los recuentos de emparejamientos por ventana incluye la

reducción del efecto de los segmentos emparejados que probablemente no provienen del historial genealógico reciente (RGH) de los individuos, sino más bien de una ascendencia común más lejana en el ser humano, etnicidad o nivel subétnico.

Para cada individuo, el método incluye calcular una suma ponderada de anchuras cM por ventana para cada segmento emparejado en función de la primera anchura cM y los pesos asociados con las ventanas de segmento del segmento emparejado. Se estima un grado de parentesco ancestral entre dos individuos en función de la suma ponderada de las anchuras cM por ventana de cada segmento emparejado entre los dos individuos. En algunas realizaciones, el grado es igual a la suma ponderada de las anchuras cM por ventana. En algunas realizaciones, la suma ponderada de las anchuras cM por ventana es la suma de las primeras anchuras cM para cada ventana de segmento de un segmento emparejado entre los dos individuos multiplicado por los dos pesos para cada individuo asociado con estas ventanas de segmento.

En algunas realizaciones, TIMBER, una máquina de predicción de ascendencia que combina marcadores genéticos, es un procedimiento que funciona en un ordenador para refinar la lista de segmentos emparejados de cada individuo y priorizar los segmentos emparejados que probablemente procedan del historial genealógico reciente de los individuos. TIMBER utiliza los segmentos emparejados para eliminar el efecto de las ventanas de segmento "ruidosas" dentro de los datos de haplotipos que muestran un recuento "excesivo" de segmentos emparejados entre numerosas personas. En algunas realizaciones, un recuento es excesivo si el recuento es mayor que 10 o 20. Es menos probable que un segmento emparejado sea de un historial genealógico reciente si un segmento emparejado es principalmente parte de las ventanas de segmento "ruidosas". TIMBER estima los pesos de los datos del segmento emparejado y estima una suma ponderada de las anchuras cM por ventana de un segmento emparejado con base en el descuento de ventanas de segmento "ruidosas". TIMBER es computacionalmente eficiente y ampliable, permitiendo reevaluar una población entera de individuos cada vez que se agregan nuevos individuos a la población.

Breve descripción de los dibujos

La figura (Fig.) 1A ilustra un diagrama de flujo de un método para estimar un grado de parentesco ancestral entre dos individuos, de acuerdo con algunas realizaciones.
La figura 1B es un diagrama de bloques de un entorno informático para estimar un grado de parentesco ancestral entre dos individuos, de acuerdo con una realización.
La figura 2 ilustra un ejemplo de recuentos de emparejamientos por ventana en un fragmento del genoma para un individuo, de acuerdo con algunas realizaciones.
La figura 3 ilustra un ejemplo del histograma de recuentos de emparejamientos por ventana para todas las ventanas del genoma para un individuo, de acuerdo con algunas realizaciones.
La figura 4 es un ejemplo del histograma del recuento de emparejamientos por ventana para todas las ventanas de recuento que no son cero, donde el recuento máximo por ventana visible es 40, de acuerdo con algunas realizaciones.
La figura 5 es un ejemplo del histograma de recuentos de emparejamientos por ventana para todas las ventanas de recuento de emparejamientos distintas de cero y también bajas, de acuerdo con algunas realizaciones.
La figura 6 es un ejemplo del peso estimado por ventana (en el eje y) en función del posible recuento de emparejamientos por ventana (en el eje x), de acuerdo con algunas realizaciones.
La figura 7 es un ejemplo de la ponderación estimada por ventana (en el eje y) en función del posible recuento de emparejamientos por ventana (en el eje x), de acuerdo con algunas realizaciones.
La figura 8 ilustra un ejemplo para los pesos (línea continua) dados los recuentos por ventana (línea discontinua) del ejemplo original en la figura 1, de acuerdo con algunas realizaciones.
La figura 9 es un ejemplo de recuentos de emparejamientos por ventana en un fragmento del genoma para un individuo tanto pre-TIMBER (línea discontinua) como post-TIMBER (línea continua), de acuerdo con algunas realizaciones.
La figura 10 ilustra los resultados de TIMBER utilizando diferentes puntuaciones de anchura cM no ponderada, es decir, primeros filtros de anchura cM y puntuaciones de suma cM ponderada, incluidos los porcentajes emparejados de segmentos guardados para la meiosis conocida y desconocida, de acuerdo con algunas realizaciones.

Las figuras representan una realización solo con fines ilustrativos. Un experto en la materia reconocerá fácilmente a partir de la siguiente descripción que se pueden emplear realizaciones alternativas de las estructuras y métodos ilustrados en el presente documento sin apartarse de los principios descritos en el presente documento.

Descripción detallada

I. PANORAMA GENERAL

Se divulgan métodos, sistemas y productos de programas informáticos para estimar un grado de parentesco ancestral entre dos individuos. La estimación del parentesco ancestral de los individuos incluye la identificación y puntuación de segmentos emparejados idénticos por descendencia (IBD) entre los datos de haplotipos de estos

individuos. Para identificar segmentos de IBD, el método compara los marcadores genéticos entre los haplotipos de los individuos. En algunas realizaciones, los marcadores genéticos incluyen polimorfismos de un solo nucleótido (SNP). Los segmentos de dos individuos se consideran idénticos por estado (IBS) si los marcadores genéticos a lo largo de las secuencias de haplotipos de los individuos en estos segmentos son idénticos en los mismos loci a lo largo de los haplotipos. A lo largo de la divulgación, a menos que se declare lo contrario, "segmentos de haplotipos emparejados" o "segmentos emparejados" se refieren a segmentos de haplotipos idénticos compartidos entre dos o más individuos. En general, un segmento de IBS compartido entre dos individuos es idéntico por descendencia (IBD) si los individuos heredaron el segmento de IBS de un antepasado común, compartiendo el mismo origen ancestral. Así, cualquier segmento de IBD por definición también representa un segmento de IBS, mientras que lo contrario no suele ser verdadero, es decir, un segmento de IBS podría no representar un segmento de IBD. Además, muchos segmentos de IBD no proceden del historial genealógico reciente (RGH) de los individuos, sino más bien de una ascendencia común más lejana en el ser humano, etnicidad o nivel subétnico. El método divulgado permite priorizar segmentos emparejados que tienen más probabilidades de ser del RGH de los individuos sobre aquellos segmentos que son de una ascendencia común más lejana, perteneciendo así a su pasado ancestral lejano, es decir, historial genealógico no reciente (no RGH).

La figura 1A es un diagrama de flujo que ilustra un método 100 para estimar un grado de parentesco ancestral entre dos individuos, de acuerdo con algunas realizaciones. El método permite al usuario introducir segmentos que se clasifican como emparejados o descubiertos, es decir, que tienen una primera anchura en centimorgan (cM) superior a 5 cM. El método en forma del programa TIMBER luego usa esos segmentos emparejados para calcular una suma ponderada de anchuras cM por ventana. La suma ponderada de las anchuras cM por ventana tiene en cuenta el recuento de segmentos emparejados con otros individuos de la población en las ventanas de segmento, ventanas de segmento de ponderación a la baja que muestran un alto grado de segmentos emparejados para muchas personas.

En algunas realizaciones, las entradas al programa TIMBER son los segmentos emparejados por pares entre todos los individuos en una población almacenada en una base de datos. El programa TIMBER traslada en pesos los segmentos emparejados por pares para las ventanas de segmentos emparejables de cada individuo del genoma. El programa TIMBER luego utiliza los pesos para recalibrar o repuntuar los segmentos emparejados por pares originales.

En algunas realizaciones, los programas TIMBER, en donde los programas TIMBER se almacenan en la memoria y se configuran para ser ejecutados por uno o más procesadores de un dispositivo informático, Los programas TIMBER que incluyen instrucciones cuando se ejecutan mediante el dispositivo informático hacen que el dispositivo:

1. calcule los recuentos de segmentos emparejados en cada ventana de segmento de los datos de haplotipos, para cada individuo de la población, donde los segmentos emparejados están entre el individuo y cualquier otra persona en la base de datos,
2. calcule pesos para cada ventana individual y de segmento, y
3. calcule una suma ponderada de anchuras cM por ventana para cada segmento emparejado entre dos individuos en función de los pesos.

El método 100 se realiza en un dispositivo informático, tal como el dispositivo informático, como puede ser controlado por un código especialmente programado (instrucciones de programación informática) contenido, por ejemplo, en el programa TIMBER, en donde dicho código especialmente programado está presente o no de forma nativa en el dispositivo informático. Las realizaciones del dispositivo informático incluyen, aunque sin limitación, ordenadores de uso general, por ejemplo, un ordenador de sobremesa, un ordenador portátil, servidores informáticos, tabletas, dispositivos móviles o cualquier dispositivo informático similar. Una vez programado para ejecutar los métodos descritos en el presente documento, dicho dispositivo informático se convierte en un ordenador de propósito especial. Algunas realizaciones del método 100 pueden incluir menos pasos, o pasos adicionales o diferentes a los mostrados en la Figura 1A, y los pasos pueden realizarse en diferentes órdenes. Los pasos del método 100 se describen con respecto a los datos de ejemplo de haplotipos ilustrados en las Figuras (Figs.) 2 a 9.

La figura 1B es un diagrama de bloques de un entorno para usar un sistema informático 120 para estimar un grado de parentesco ancestral entre dos individuos, de acuerdo con algunas realizaciones. En la Figura 1B, se representan individuos 122 (es decir, un ser humano u otro organismo), un servicio de extracción de ADN 124 y un servicio de control de calidad (QC) de ADN y un servicio de preparación de emparejamientos 126.

Los individuos 122 proporcionan muestras de ADN para el análisis de sus datos genéticos. En alguna forma de realización, un individuo usa un kit de recogida de muestras para proporcionar una muestra, por ejemplo, saliva, a partir de la cual se pueden extraer los datos genéticos de manera fiable de acuerdo con los métodos convencionales. El servicio de extracción de ADN 124 recibe la muestra y genotipa los datos genéticos, por ejemplo, extrayendo el ADN de la muestra e identificando valores de SNP presentes dentro del ADN. El resultado es un genotipo diploide. El control de calidad del ADN y el servicio de preparación de emparejamientos 126 evalúa la calidad de los datos del genotipo diploide verificando varios atributos, como el genotipado por *call rate*, el genotipado de la tasa de heterocigosidad y la concordancia entre género genético y autoinformado. El sistema 120 recibe 102 los datos de haplotipos del servicio de extracción de ADN 124 y opcionalmente almacena los datos de haplotipos en

una base de datos 128 que contiene genotipos diploides de ADN sin fase, haplotipos por fases y otros datos genómicos. A menos que se indique lo contrario, los datos de haplotipos se refieren a cualquier información genética o genómica obtenida de los individuos 122, que se almacena opcionalmente en la base de datos 128.

En algunas realizaciones, el módulo de partición 130 divide 104 los datos de haplotipos en ventanas de segmento con base en los marcadores genéticos. En algunas realizaciones, el módulo de emparejamiento 132 empareja 106 segmentos de los datos de haplotipos que son idénticos entre el individuo y cualquier otro individuo en la población, donde cada segmento emparejado tiene una primera anchura cM que excede una anchura cM umbral y forma parte de una o más de las ventanas de segmento.

En algunas realizaciones, el módulo de estimación de recuento/peso 134 cuenta 108 los segmentos emparejados en cada ventana de segmento y estima 110 un peso asociado con cada ventana de segmento en función del recuento de segmentos emparejados en la ventana de segmento asociada.

El módulo de puntuación 136 calcula 112 entonces una suma ponderada de anchuras cM por ventana para cada segmento emparejado con base en la primera anchura cM y los pesos asociados con las ventanas de segmento del segmento emparejado. En algunas realizaciones, el módulo de puntuación 136 estima 114 un grado de parentesco ancestral entre dos individuos en función de la suma ponderada de las anchuras cM por ventana de cada segmento emparejado entre los dos individuos.

II. REALIZACIONES DE LOS PROGRAMAS DE PREDICCIÓN DE RELACIÓN ANCESTRAL

En algunas realizaciones, los segmentos emparejados entre una población de individuos se generan con base en los datos de haplotipos de los individuos. En algunas realizaciones, los segmentos emparejados se almacenan en una base de datos para su posterior recuperación por el programa TIMBER. El programa TIMBER está configurado para recibir todos los segmentos emparejados entre una población de individuos y priorizar los segmentos emparejados que provienen del historial genealógico reciente (RGH) de los individuos.

II.A. Emparejar segmentos de haplotipos

El método 100 incluye recibir 102 datos de haplotipos para una población de individuos, los datos de haplotipos que incluyen una pluralidad de marcadores genéticos compartidos entre los individuos, de acuerdo con algunas realizaciones. En algunas realizaciones, para identificar (emparejar) y puntuar segmentos de IBD entre los datos de haplotipos, el método 100 usa una reimplementación HADOOP® de un algoritmo de emparejamiento. El método se beneficia de ser computacionalmente rápido y ampliable para poblaciones de mayor tamaño. En algunas realizaciones, los segmentos de IBS emparejados con base en los datos de haplotipos de los individuos incluyen los segmentos de RGH y no RGH de los individuos. Los segmentos de IBS emparejados generalmente se denominan segmentos emparejados. En algunas realizaciones, los segmentos emparejados se generan usando métodos que son bien conocidos en la técnica. Utilizando, por ejemplo, el programa TIMBER, el método 100 luego prioriza los segmentos emparejados en segmentos que son más probables del RGH de dos individuos en comparación con su no RGH calculando una puntuación de TIMBER que cuantifica la probabilidad de que dos individuos compartan una relación ancestral reciente común.

El método para determinar la puntuación TIMBER se basa en el supuesto de que las ubicaciones (loci) de los segmentos emparejados del RGH de un individuo se distribuyen de manera uniforme en el genoma de un individuo. Por ejemplo, dividiendo los SNP de los cromosomas a través del genoma de los individuos en ventanas discretas, los recuentos de los segmentos en una ventana específica en el cromosoma 1 son independientes de los recuentos de los segmentos en una ventana en el cromosoma 14. En consecuencia, los emparejamientos entre segmentos de individuos en una ventana en el cromosoma 1 son, por lo tanto, independientes de los emparejamientos con segmentos en la ventana en el cromosoma 14, dando como resultado una distribución uniforme en todas las ventanas.

Además, los segmentos emparejados que no se originan a partir del RGH de un individuo exhiben picos en ciertas ventanas del genoma mientras se distribuyen uniformemente entre las ventanas restantes. En algunos casos, estos picos pueden atribuirse a razones particulares de la variación genética en estas ventanas, por ejemplo, a nivel de personas o a nivel de base de datos. A nivel de personas, los picos pueden ser el resultado de los segmentos de una ventana que muestra un alto nivel de similitud de secuencias en un grupo étnico particular, mientras que a nivel de base de datos, es posible que los individuos de la población en la base de datos no posean ninguna variación de SNP en una ventana particular en función de cómo se seleccionó la población. En particular, no está claro si la distribución de los recuentos de segmentos emparejados en una ventana se origina en el RGH de un individuo, y si los factores que confunden los picos locales de emparejamiento se deben a razones de no RGH desconocidos, que son difíciles de modelar. Para superar los problemas asociados con estos picos, el método introduce pesos para cada ventana para repuntuar todos los segmentos emparejados, en particular, los que contribuyen a los picos en las ventanas. En algunos casos, los segmentos emparejados ocurren comúnmente en una ubicación corta y específica del genoma de un individuo, mientras que emparejan un gran número de otros individuos, por ejemplo, más de 1.000 individuos.

El método 100 incluye dividir 104 los datos de haplotipos en ventanas de segmento con base en los marcadores genéticos, de acuerdo con algunas realizaciones. Los datos de haplotipos incluyen, aunque sin limitación, SNP observados a través del genoma del individuo. En algunas realizaciones, los datos de haplotipos incluyen los datos de haplotipos del genoma completo o parcial de un individuo. En algunas realizaciones, el método 100 divide los SNP observados en *K* ventanas de igual tamaño d, con cada ventana, por ejemplo, incluyendo 96 SNP. Otros ejemplos de tamaños de ventanas incluyen 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150 o cualquier número que se encuentre dentro del intervalo de 50 a 500. En algunas realizaciones, el tamaño de cada ventana, es decir, el número de SNP por ventana varía según las ventanas.

Por ejemplo, algunas ventanas pueden incluir solo 50 SNP dependiendo de la longitud de la secuencia de estos SNP, mientras que otras ventanas incluyen 96 SNP. En algunas realizaciones, las ventanas incluyen otros marcadores genéticos además de los SNP que se usan para identificar segmentos emparejados. Estos marcadores incluyen, aunque sin limitación, polimorfismos de longitud de fragmento de restricción, polimorfismos de longitud de secuencia simple, polimorfismo de longitud de fragmento amplificada, amplificación aleatoria de ADN polimórfico, repetición en tándem de números variables, polimorfismo de microsatélites de repetición de secuencias simples, repeticiones cortas en tándem, polimorfismos de una única característica, marcadores de ADN asociados al sitio de restricción, y similares.

En algunas realizaciones, el método 100 incluye el uso de datos de haplotipos por fases, es decir, datos para los cuales se ha estimado la fase, como entrada para identificar segmentos emparejados. Para esto, el método utiliza los datos de haplotipos para una población de *n* individuos. En algunas realizaciones, la entrada de los datos de haplotipos se representa como una matriz H *2n* x s con filas correspondientes a 2n haplotipos y columnas a s SNP. Al cortar verticalmente H en submatrices H_i de igual anchura no superpuestas de *d* columnas, cada submatriz H_i luego representa una ventana *i* de segmento diferente, donde $i = 0 \ldots K$ y $s = d \cdot K$. En algunas realizaciones, los datos de haplotipos incluyen n haplotipos implícitos no en fase de la población que utilizan los genotipos de la población para determinar posibles emparejamientos de haplotipos sin un emparejamiento explícito de haplotipos que requeriría conocer la fase de los haplotipos. El emparejamiento de haplotipos, por lo tanto, se refiere al emparejamiento de haplotipos implícito y explícito, donde el anterior se basa en datos genómicos no en fase de una población.

El método 100 incluye para cada individuo en la población, con base en los marcadores genéticos, emparejar 106 segmentos de los datos de haplotipos que son idénticos entre el individuo y cualquier otro individuo en la población, de acuerdo con algunas realizaciones. En algunas realizaciones, cada segmento emparejado tiene una primera anchura cM que excede una anchura cM umbral y forma parte de una o más de las ventanas de segmento. El emparejamiento 104 incluye la identificación de ventanas de segmento de emparejamiento de haplotipos exacto entre dos individuos en la población. Las ventanas de emparejamiento de haplotipos exacto se utilizan para anclar la identificación de todo el segmento emparejado que, en algunos casos, se extiende más allá del emparejamiento de ventana exacto inicial. En algunas realizaciones, el método 100 incluye extender el emparejamiento de ventana exacto hasta que se observan dos SNP homocigotos mal emparejados a cada lado del emparejamiento de ventana exacto original. Como resultado, el método 100 determina la anchura del segmento determinando el mínimo y el máximo de las ubicaciones inicial y final de las ventanas sin desajustes homocigotos y extendiendo el emparejamiento de ventana exacto.

En caso de que la anchura del segmento determinado exceda una anchura del umbral de anchura cM umbral, el método 100 identifica el segmento correspondiente como un segmento emparejado. En algunas realizaciones, la anchura cM umbral es de 5 cM. En algunas realizaciones, el umbral de anchura cM umbral es 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 cM o cualquier valor mayor de 5 cM. En algunas realizaciones, el método 100 empareja 104 segmentos de individuos incluidos en una base de datos de individuos que incluye los datos de haplotipos del genoma de cada individuo.

II.B. Recuento de emparejamientos por ventana de segmentos emparejados

El método 100 incluye para cada individuo en el recuento de la población 108 los segmentos emparejados en cada ventana de segmento, de acuerdo con algunas realizaciones. Para cada individuo en la población y cada ventana, el método 100 determina un recuento de emparejamientos por ventana de segmentos emparejados. Un recuento particular de emparejamientos por ventana k_i se refiere al número de segmentos emparejados identificados dentro de la población con *i* indicando la ventana del recuento. En algunas realizaciones, dado que cada segmento emparejado entre dos individuos abarca varias ventanas, el método 100 traslada los segmentos emparejados de todos los individuos en los identificadores de ventana y cuenta cuántas veces cada ventana forma parte de un segmento emparejado. Los segmentos emparejados de parientes cercanos no se incluyen en el recuento de emparejamientos por ventana, lo que aumenta la probabilidad de que todos los segmentos emparejados compartan niveles similares de incertidumbre sobre si el segmento emparejado forma parte o no del RGH de los individuos. En algunas realizaciones, un pariente cercano se define como un individuo con quien el individuo relacionado tiene una puntuación de TIMBER en bruto que iguala o excede un umbral de parientes cercanos predefinido. En algunas realizaciones, el umbral de parientes cercanos predefinido es de 50 cM. En algunas realizaciones, el umbral de

parientes cercanos predefinido es 30, 40, 50, 60, 70, 80, 90, 100 cM o cualquier valor del intervalo de 30 a 200 cM. Una puntuación de TIMBER en bruto entre dos individuos se define como la suma de las primeras anchuras cM de segmentos emparejados sobre todos los segmentos emparejados entre los dos individuos. La puntuación se conoce como puntuación "en bruto", dado que las anchuras cM utilizadas son las primeras anchuras cM que no están ponderadas hacia abajo.

En algunas realizaciones, el método para determinar el recuento de emparejamientos por ventana para cada individuo incluye los siguientes pasos:

1. inicializar un vector de recuento de emparejamientos por ventana $\{k_i\}_{i = 0... K}$ a cero recuentos (un valor k_i por cada ventana de emparejamiento en el genoma, por ejemplo, con K igual a 4105 ventanas, y cada ventana incluyendo 96 marcadores SNP; y
2. para cada segmento emparejado:

   (a) saltar el segmento emparejado para parientes cercanos y pasar al siguiente segmento emparejado,
   (b) trasladar el segmento emparejado en un vector de índices de ventana $\{i\}$ (que abarca el segmento emparejado) y
   (c) incrementar las entradas del vector de recuento de emparejamientos por ventana $\{k_i\}$ por uno para los valores respectivos del vector de índices de ventana $\{i\}$.

3. eliminar cualquier entrada del vector de recuento de emparejamientos por ventana $\{k_i: k_i > 0\}_{i = 0... K}$ que tenga un recuento de segmentos emparejados de cero.

La figura 2 ilustra un ejemplo de recuentos de emparejamientos por ventana en una porción del genoma de un individuo en particular en la población. El eje x muestra el número de una ventana en particular *i* a lo largo de la secuencia de haplotipos, y el eje y representa el número total de segmentos emparejados *ki* para una ventana particular *i.* En este ejemplo, las ventanas que se muestran van del número 1 al número 145. La figura 2 muestra ventanas con cero segmentos emparejados, mientras que estas ventanas están excluidas del vector de recuento de emparejamientos por ventana $\{k_i: k_i > 0\}_{i = 0... K}$. Los mayores recuentos de emparejamientos se observan para la ventana número 1 y número 2 en este ejemplo, llegando a cerca de 60 segmentos emparejados.

La figura 3 ilustra un ejemplo del histograma de los recuentos de emparejamientos por ventana $\{c: c = 0 ... C_{max}, C_{max} \leq n\}$ para todas las ventanas en todo el genoma para un individuo, que solo muestra una parte del genoma del individuo. El histograma indica la frecuencia de observación de segmentos emparejados contra toda la población *n* en cada ventana genómica, donde *n* por lo tanto limita el recuento máximo por ventana C_{max}. Solo las ventanas que incluyen menos de 305 segmentos emparejados se muestran con el recuento de ventanas disminuyendo rápidamente para aumentar los valores de recuento de emparejamientos por ventana. En este ejemplo, las ventanas que incluyen cero segmentos emparejados no se muestran, ya que estas ventanas no se analizan más en el método 100 de cálculo de la puntuación de TIMBER.

Como se muestra en la figura 2, ciertas ventanas o regiones de segmentos dentro del genoma del individuo muestran un recuento muy grande de segmentos emparejados, ya sea debido a un alto nivel de segmentos emparejados que son no RGH por razones desconocidas o debido a una distribución realmente alta de segmentos de RGH emparejados en estas ventanas. El método 100 intenta diferenciar entre estas dos posibilidades evaluando el recuento de emparejamientos por ventana para cada individuo de la población. Además, el método 100 determina una puntuación de TIMBER para cada segmento emparejado ponderando a la baja las ventanas emparejadas de un segmento emparejado que tienen una probabilidad disminuida de originarse a partir del RGH entre los dos individuos emparejados. Esta ponderación a la baja, en algunas realizaciones, incluye determinar un peso para cada individuo para un segmento emparejado en una ventana.

II.C. Estimación de pesos $\{w_i^A\}$

En algunas realizaciones, el método 100 incluye estimar 110 un peso asociado con cada ventana de segmento en función del recuento de segmentos emparejados en la ventana de segmento asociada, de acuerdo con algunas realizaciones. En algunas realizaciones, la estimación 110 incluye determinar para cada individuo de la población un peso para cada ventana que cuenta al menos un segmento emparejado. En algunas realizaciones, el peso se aproxima por la probabilidad de que el emparejamiento en esa ventana para ese individuo proporcione evidencia de RGH. Esta probabilidad está relacionada con el recuento de segmentos emparejados en una ventana. Una ventana con un recuento de segmentos emparejados extremadamente alto para un individuo es muy poco probable debido al RGH que el individuo comparte con otros individuos en la población. Factores desconocidos distintos de RGH pueden explicar un recuento de segmentos muy alto como se describió anteriormente.

Para estimar los pesos, el método 100 determina la probabilidad de RGH, Prob(RGH|C = c), dado el recuento medido c de segmentos emparejados en una ventana. La variable aleatoria *C* representa todos los recuentos posibles de segmentos emparejados en una ventana y se supone que es idéntica en todas las ventanas. Midiendo

los recuentos reales *c* en ventanas particulares, el método 100 determina la probabilidad de RGH con la condición de que *C* es igual a *c* para esta ventana. Para determinar Prob(RGH|C = c), el método 100 utiliza el teorema de Bayes que proporciona:

$$\text{Prob}(\text{RGH}|C = \text{c}) = \frac{\text{Prob}\ (C = \text{c}|RGH)\text{Prob}\ (\text{RGH})}{\text{Prob}\ (C = \text{c})} \qquad (1),$$

donde Prob (C = c|RGH) es la probabilidad de tener c segmentos emparejados en la ventana y todos los segmentos emparejados se deben al RGH de los individuos, Prob (C = c) es la probabilidad de tener c segmentos emparejados en la ventana, independientemente de que el segmento emparejado sea de RGH o no RGH, y Prob (RGH) es la probabilidad de que el segmento emparejado sea del RGH de los individuos. Con base en el teorema de Bayes, el método 100 determina estimaciones de Prob (C = c|RGH), Prob (RGH) y Prob (C = c), donde

$$\text{Prob}\ \widehat{(C = c|\text{RGH})},$$

$$\widehat{\text{Prob}(\text{RGH})},$$

$\widehat{Prob(C = c)}$ son las estimaciones de Prob(C = c|*RGH),* Prob(RGH) y Prob(C = c), respectivamente. El método 100 luego determina el peso w_i^A para un individuo A en una ventana específica *i* de acuerdo con:

$$w_i^A = \frac{\text{Prob}\ \widehat{(C = k_i|\text{RGH})}\text{Prob}(\text{RGH})}{\text{Prob}\widehat{(C = k_i)}} \qquad (2).$$

En algunas realizaciones, ya que es difícil estimar Prob(C = c|RGH), el método 100 genera al menos dos estimaciones ligeramente diferentes de Prob(C = c|RGH), y luego seleccionando la estimación de al menos dos estimaciones que resultan en la mayor ponderación a la baja de C para determinar el peso w_i^A.

II.C.1. Determinar $\widehat{Prob(C = c)}$ una estimación

En algunas realizaciones, el método 100 incluye determinar una estimación de la distribución de probabilidad de C, $\widehat{Prob(C = c)}$, que proporciona la probabilidad de que una ventana tenga un recuento de segmentos emparejados c dado para todos los recuentos posibles. Estas realizaciones proporcionan una predicción más precisa de la probabilidad en el número de personas que un individuo empareja en una ventana en función de la probabilidad de emparejamiento de cada individuo en la población. Por ejemplo, el valor de $\widehat{Prob(C = 20)}$ es la probabilidad de contar 20 segmentos emparejados en una ventana dada para todos los datos de haplotipos, incluyendo todos los segmentos emparejados de la población. $\widehat{Prob(C = c)}$ incluye contribuciones de segmentos emparejados que son de RGH y no RGH de los individuos de la población. La figura 4 ilustra $\widehat{Prob(C = c)}$de un individuo de la población con base en el genoma completo del individuo. Dado $\widehat{Prob(C = c)}$ y Prob (C = c|RGH), el método 100 puede cuantificar la probabilidad de que una ventana transmita información sobre RGH para un recuento *c* dado. En algunas realizaciones, ambas distribuciones estiman la distribución de recuentos para recuentos que son mayores que cero, es decir, ventanas que tienen al menos un segmento emparejado dentro de ellas.

Las figuras 3 y 4 ilustran $\widehat{Prob(C = c)}$ en forma de un histograma de recuentos de emparejamientos por ventana para todas las ventanas de recuento distintas de cero sin contar ninguna ventana que incluya cero segmentos. En algunas realizaciones, solo se cuentan las ventanas con cero segmentos por al menos las dos razones siguientes: 1) diferentes procesos biológicos u observacionales son la causa probable de descubrir un segmento emparejado dentro de una ventana en comparación con el número de segmentos emparejados descubiertos dentro de esa ventana para una población dada; y 2) el peso de las ventanas debe basarse en segmentos emparejados que están realmente presentes y no faltantes en una población. Así, estas realizaciones evitan asignar efectivamente un peso de uno a ventanas con cero segmentos emparejados descubiertos. En comparación, la figura 2 ilustra los valores de los recuentos para una ventana específica (y no la frecuencia con la que se observa ese recuento en todas las ventanas que no son cero).

Para determinar una distribución a la probabilidad de que se observe un recuento particular en una ventana, es decir, Prob(C = c), el método 100 se ajusta a una distribución de recuentos observados en todas las ventanas que no son cero ilustradas como un histograma en la figura 4.

En algunas realizaciones, el método 100 usa una primera distribución beta-binomial para aproximar la distribución de los recuentos por ventana, $Prob(\widehat{C} = c)$, que representa la probabilidad de que cada individuo se empareje con otro individuo en la población. La ventaja de usar una distribución beta-binomial es que es capaz de explicar la heterogeneidad subyacente en la probabilidad de emparejar individuos sin identificar la razón de la heterogeneidad. Tal y como se ilustra en la Figura 4, la distribución beta-binomial proporciona un buen ajuste para los recuentos por ventana en un individuo. En comparación, si se supone que todos los individuos de la población se emparejan con las mismas oportunidades que cualquier otro individuo de la población, la distribución binomial proporciona un modelo de los recuentos observados en todas las ventanas que no son cero. Una distribución binomial normalmente se usaría para modelar la probabilidad de que varios eventos sean exitosos, si la probabilidad conocida de éxito se comparte entre todos los eventos independientes.

La figura 4 ilustra además un ejemplo de ajuste de una distribución binomial a los recuentos por ventana en un individuo en comparación con una distribución beta-binomial. En general, se prefiere la distribución beta-binomial para modelar $Prob(\widehat{C} = c)$. En particular, la figura 4 ilustra un ejemplo del histograma del recuento de emparejamientos por ventana para todas las ventanas de recuento distintos de cero donde el recuento máximo de emparejamientos por ventana que se muestra es 40. El modelo de distribución binominal $Prob(\widehat{C} = c)$ se muestra como una línea discontinua con la línea continua que indica el ajuste de la distribución beta-binomial a los datos del histograma.

En algunas realizaciones, el método 100 determina dos parámetros α y *β de* la distribución beta-binomial para determinar el ajuste óptimo entre la distribución beta-binomial y el recuento de emparejamientos por ventana de segmentos emparejados para un individuo en función de toda la población de individuos. Dado que la distribución beta-binomial se define para recuentos desde cero hasta un recuento máximo *n* que equivale al tamaño de la población, el método 100 usa un vector de recuento de emparejamientos por ventana modificado que se obtiene restando uno de cada elemento de $\{k_i: k_i > 0\}_{i = 0 \ldots K}$. El número de observaciones utilizadas para determinar la distribución beta-binomial es igual al número de ventanas *K* a través de los datos de haplotipos. Por ejemplo, los datos de haplotipos se dividen en 4105 ventanas, cada ventana incluye 96 marcadores. La distribución beta-binomial conjunta $f(\{k_i\}|n, \alpha, \beta, k_i > 0)$ es dada por:

$$f(\{k_i\}|n, \alpha, \beta, k_i > 0) = \prod_{i=1}^{K} \binom{n}{k_i - 1} \frac{\mathrm{B}(k_i - 1 + \alpha, n - k_i + 1 + \beta)}{\mathrm{B}(\alpha, \beta)} \qquad (3),$$

donde $\{k_i: k_i > 0\}_{i = 0 \ldots K}$ es el vector de recuentos por ventana (para las ventanas K con al menos un segmento emparejado), *n* es el tamaño de la población, *a* y *β* son parámetros de la distribución, y B es la función beta. Prob (C = c) para un tamaño de población *n* entonces es dado por:

$$\mathrm{Prob}(C = \mathrm{c} \mid \mathrm{c} > 0) = \binom{n}{c} \frac{\mathrm{B}(c + \alpha, n - c + \beta)}{\mathrm{B}(\alpha, \beta)} \qquad (4).$$

El método 100 utiliza los datos de haplotipos de todas las personas que se emparejan con una persona para determinar los parámetros *a* y *β* de la distribución asociada a este individuo. Esta determinación se basa en parte en el supuesto de que el recuento de cada ventana es independiente del recuento de cualquier otra ventana. Como ejemplos típicos, este supuesto proporciona una buena estimación de los datos reales. En algunos casos, esta suposición podría no ser verdadera, y otras funciones de distribución pueden proporcionar mejores aproximaciones a los datos.

En algunas realizaciones, el método 100 utiliza la estimación de máxima verosimilitud para determinar los dos parámetros α y β de la distribución beta-binomial conjunta $f(\{k_i\}|n, \alpha, \beta)$. El vector de recuento de emparejamientos por ventana observado $\{k_i: k_i > 0\}_{i = 0 \ldots K}$ representa a lo sumo *K* parámetros fijos de la función de verosimilitud, $\mathcal{L}(\alpha,\beta|\{k_i\},n) = f(\{k_i\}|n,\alpha,\beta)$. Los parametros *α* y *β* son números reales mayores que cero que maximizan el logaritmo promedio de la función de probabilidad.

En algunas realizaciones, el método 100 usa un método de optimización 100 para estimar la verosimilitud máxima $\mathcal{L}(\alpha,\beta|\{k_i\},$ n) de la distribución beta-binomial conjunta para $\{k_i\}$ y *n.* Varios métodos de optimización son bien conocidos en la técnica, cada uno de los cuales puede usarse para la estimación de máxima verosimilitud. En algunas realizaciones, el método 100 aplica el algoritmo de optimización simplex "Nelder-Mead" (Nelder J.A. y Mead R., A simplex algorithm for Function Minimization, Computer J., 7:308-13, 1965) como el método de optimización 100. Los parámetros a y β se establecen inicialmente en valores iniciales indicados por $\tilde{\alpha}$ y $\tilde{\beta}$ proporcionados por:

$$\tilde{\alpha} = \frac{nE1 - E2}{n\left(\frac{E2}{E1} - E1 - 1\right) + E1} \qquad (5),$$

$$\tilde{\beta} = \frac{(n - E1) \cdot \left(n - \frac{E2}{E1}\right)}{n\left(\frac{E2}{E1} - E1 - 1\right) + E1} \qquad (6),$$

donde $E1 = \frac{\sum_{i=1}^{K}(k_i - 1)}{n}$ y $E2 = \frac{\sum_{i=1}^{K}(k_i - 1)^2}{n}$.

El parámetro α y β obtenido por el método de optimización 100 se denota por $\tilde{\alpha}$ y $\tilde{\beta}$ Por lo tanto, todos los individuos tienen una distribución beta-binomial estimada para describir $\text{Prob}\,(C = \text{c} \mid \text{c} > 0) = \binom{n}{c} \frac{\text{B}(c+\hat{\alpha}, n-c+\hat{\beta})}{\text{B}(\hat{\alpha},\hat{\beta})}$, donde $\tilde{\alpha}$ y $\tilde{\beta}$ son específicos de cada individuo en una población de tamaño n, ilustrado en el ejemplo de la figura 4.

II.C.2. Determinar $\text{Prob}(\widehat{\text{C} = \text{c}|\text{RGH}})$ *estimaciones*

En algunas realizaciones, el método 100 determina una estimación de la distribución de probabilidad $\text{Prob}\widehat{(\text{C} = \text{c}|\text{RGH})}$ ajustando una segunda distribución beta-binomial a los recuentos de emparejamientos por ventana que solo incluyen ventanas con un recuento de emparejamientos bajo y excluye ventanas con un recuento de emparejamientos más alto. La primera distribución beta-binomial se basa en el ajuste a $Prob(\widehat{C = c})$, como se ha descrito anteriormente.

En algunas realizaciones, el método 100 excluye cualquier recuento de emparejamientos por ventana que exceda un valor umbral *V.* Excluir ventanas con recuentos de emparejamientos más altos se basa en la lógica de que las ventanas con recuentos de emparejamientos bajos se deben principalmente a segmentos emparejados de los RGH de los individuos, mientras que las ventanas de recuento de emparejamientos altos probablemente incluyen segmentos emparejados que son una mezcla de RGH y no RGH. En general, $\text{Prob}(\widehat{\text{C} = \text{c}|\text{RGH}})$ puede estimarse a partir de Prob (C = c), ya que la distribución posterior representa la suma de dos distribuciones condicionales Prob (C = c|RGH) y Prob (C = c|no RGH). Sin embargo, ninguna de estas distribuciones condicionales se conoce con confianza o puede determinarse fácilmente a partir de Prob(C = c), en parte porque no se puede distinguir entre segmentos emparejados que son de los RGH de los individuos y segmentos emparejados que no lo son. Intentos de estimar $\text{Prob}(\widehat{\text{C} = \text{c}|\text{RGH}})$ directamente de Prob(c = c), por lo tanto, a menudo resultan en estimaciones pobres y, a veces, muy pobres, mientras se añaden varios problemas computacionales.

Además, Prob(C = c) está bien estimado por una distribución beta-binomial como se describió anteriormente. Para estimar Prob(C = c|RGH), por lo tanto, el método 100 ajusta una segunda distribución beta-binomial a los recuentos de emparejamientos por ventana con base en ventanas que tienen un recuento de emparejamientos inferior o igual al valor umbral V.

Tal y como se ilustra en la Figura 5, ambas distribuciones ajustadas de Prob (C = c) y Prob (C = c|RGH) son similares entre sí, si la distribución de recuentos para las ventanas de alto valor que no están incluidas en el ajuste de $\text{Prob}(\widehat{\text{C} = \text{c}|\text{RGH}})$ es coherente con la distribución estimada solo a partir de las ventanas de recuento de emparejamientos bajas. La figura 5 muestra la primera distribución beta-binomial con base en todas las ventanas de recuento distintas de cero (mostradas por la línea continua), la segunda distribución beta-binomial con base solo en las ventanas de recuento de emparejamientos bajos (mostrados por la línea discontinua) y el histograma con base solo en los datos de las ventanas de recuento de emparejamientos bajas. El recuento máximo de emparejamientos por ventana que se muestra es 40.

En algunas realizaciones, puesto que la estimación $\text{Prob}(\widehat{\text{C} = \text{c}|\text{RGH}})$ es probable que sea sensible al último valor especificado por el usuario V, el método 100 determina al menos dos estimaciones de Prob (C = c|RGH, V) utilizando al menos dos valores diferentes de *V.* El método 100 luego selecciona la estimación que resulta en pesos más pequeños para las ventanas, ponderar a la baja más eficazmente los recuentos de segmentos emparejados de los individuos como se describe con más detalle a continuación. En algunas realizaciones, el método 100 determina solo una estimación de $\text{Prob}(\widehat{\text{C} = \text{c}|\text{RGH}})$. En algunas realizaciones, *V* se especifica como el valor máximo de un valor umbral mínimo V_{min} y un cuantil de todos los recuentos en ventanas con al menos un segmento emparejado. En algunas realizaciones, el valor umbral mínimo V_{min} se establece en 11. En algunas realizaciones, el valor umbral

mínimo V_{min} es 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 20, 25, 30 o más de 30. En algunas realizaciones, se especifican al menos dos cuantiles para determinar las al menos dos estimaciones de $\widehat{\mathrm{Prob}(\mathrm{C} = \mathrm{c}|\mathrm{RGH})}$. En algunas realizaciones, dos cuantiles especificados son 75 % y 90 %, respectivamente. En algunas realizaciones, el primer cuantil especificado es igual al 50 %, 55 %, 60 %, 65 %, o 70 %, y el segundo cuantil especificado equivale al 65 %, 70 %, 75 %, 80 % u 85 %. En algunas realizaciones, los dos cuantiles se especifican de modo que la diferencia entre los dos cuantiles esté en el intervalo del 10 %-20 %, y el cuantil más pequeño sea igual al 50 %, 55 %, 60 %, 65 %, 70 %, 75 % o cualquier valor en el intervalo del 40 %-80 %.

En algunas realizaciones, para estimar Prob(C = c|RGH, *V)* el método 100 usa una distribución beta-binomial conjunta $f(\{m_i\}|n, \alpha, \beta V, m_i > 0)$ que depende de V, *α* y β y viene dada por:

$$f(\{m_i\}|n,\alpha,\beta,V,m_i>0) = \frac{\prod_{i=1}^{M}\binom{n}{m_i-1}\frac{\mathrm{B}(m_i-1+\alpha,\,n-m_i+1+\beta)}{\mathrm{B}(\alpha,\beta)}}{\mathrm{Prob}(0<C\leq V|\alpha,\beta)} \qquad (7).$$

donde $\{m_i: 0 < m_i \leq V\}_{i=0\ldots M}$ es el vector de recuentos por ventana (para las M ventanas con al menos un segmento emparejado y menos o igual que los recuentos de V), *n* es el tamaño de la población, α y β son los parámetros para la función beta B, y Prob (0 < *C* ≤ V|α, β) es la probabilidad de que un recuento por segmento emparejado de ventana sea mayor que cero y menor o igual que V condicional a *a* y β

La distribución de probabilidad de C condicional a RGH viene dada por:

$$\mathrm{Prob}(C=\mathrm{c}|\mathrm{RGH},V,c>0) = \binom{n}{c}\frac{\mathrm{B}(c+\alpha,\,n-c+\beta)}{\mathrm{B}(\alpha,\beta)}\Bigg/\mathrm{Prob}(0<C\leq V|\alpha,\beta) \qquad (8),$$

en donde Prob (0 < *C* ≤ V|α, β) representa un factor de normalización que es función de *a* y β y dado por:

$$\mathrm{Prob}(0<C\leq V|\alpha,\beta) = \sum_{c=1}^{V}\binom{n}{c}\frac{\mathrm{B}(c+\alpha,\,n-c+\beta)}{\mathrm{B}(\alpha,\beta)} \qquad (9).$$

En algunas realizaciones, el método 100 utiliza la estimación de máxima verosimilitud para determinar los dos parámetros α y β de la distribución beta-binomial conjunta *f({mi}|n, α, β, V)* usando los mismos algoritmos de optimización o similares que se usan para estimar Prob(C = *c)* como se describió anteriormente. En algunas realizaciones, el método 100 usa los mismos valores iniciales o similares para *a* y *β* que se usan para estimar Prob(C = c). El uso de los mismos valores iniciales para la primera y la segunda distribución beta-binomial conjunta minimiza el efecto de los valores iniciales sobre las diferencias entre estas dos distribuciones.

II.C.3. Determinar $\widehat{\mathrm{Prob}(\mathrm{RGH})}$ *una estimación*

En algunas realizaciones, el método 100 luego determina una estimación de $\widehat{\mathrm{Prob}(\mathrm{RGH})}$ con base en $\widehat{Prob(C = c)}$ y $\widehat{\mathrm{Prob}(\mathrm{C} = \mathrm{c}|\mathrm{RGH})}$. En algunas realizaciones, $\widehat{\mathrm{Prob}(\mathrm{RGH})}$ está configurado para ser máximo de $\frac{1}{\frac{\mathrm{Prob}\,(\mathrm{C}=\mathrm{c})\mathrm{RGH})}{\widehat{\mathrm{Prob}(\mathrm{C}=\mathrm{c})}}}$, que es igual a $\frac{\widehat{\mathrm{Prob}(\mathrm{C}=\mathrm{m})}}{\mathrm{Prob}(\overline{\mathrm{C}=\mathrm{m}}|\mathrm{RGH})}$, donde m es el punto en el cual $\frac{\mathrm{Prob}(\overline{C=c}|RGH)}{\widehat{\mathrm{Prob}(\mathrm{C}=\mathrm{c})}}$ está maximizado. Así, el método 100 solo estima $\widehat{\mathrm{Prob}(\mathrm{RGH})}$ implícitamente evaluando la relación $\widehat{\mathrm{Prob}(\mathrm{C} = \mathrm{c}|\mathrm{RGH})} \;/\; \widehat{Prob(C = c)}$. En algunas realizaciones, el método 100 determina al menos dos estimaciones de Prob (RGH, V) con base en las al menos dos Prob(C = c|RGH, V). En algunas realizaciones, el método 100 determina dos $\widehat{\mathrm{Prob}(\mathrm{RGH},\mathrm{V})}$, estimaciones con base en los dos valores umbral especificados por el usuario *V,* donde cada *V son* determinados en dos cuantiles especificados de todos los recuentos en ventanas con al menos un segmento emparejado como se describe anteriormente. Determinar la estimación de Prob(RGH) como una razón de dos probabilidades asegura que la estimación, así como los pesos correspondientes, caigan en el intervalo de cero a uno, ya que el peso también se determina como una razón de tres probabilidades que se encuentran dentro del intervalo de cero a uno. El peso sería indefinido, si $\widehat{Prob(C = c)}$ es cero. La estimación descrita anteriormente usando una distribución beta-binomial asegura que cualquier valor de $\widehat{Prob(C = c)}$ es mayor que cero.

II.C.4. Estimación de pesos temporales $\{W_c\}$ *para cada estimación de* Prob(C = c|RGH)

El método 100 determina entonces los pesos temporales con base en las estimaciones de $Prob(\widehat{C = c})$, $\text{Prob}(\widehat{\text{C = c}}|\text{RGH})$, y $\text{Prob}(\widehat{\text{RGH}})$. Estos pesos son temporales, dado que el método 100 usa estos pesos para determinar el peso final para cada ventana. Los pesos temporales se pueden representar mediante un vector $\{w_c\}_{c=0\ldots n}$ que es una serie de valores para diferentes recuentos de emparejamientos y w_c es el peso en bruto para un recuento de emparejamientos de c. Dadas las al menos dos estimaciones de Prob (C = c|RGH) en algunas realizaciones, el método 100 determina los pesos temporales para cada estimación de Prob(C = c|RGH) y luego selecciona una serie de pesos temporales para determinar el peso final.

Si alguna optimización falla para un cuantil, se ignora en el paso de decisión, por lo tanto, si los pesos temporales solo pueden estimarse para una opción de cuantil, entonces esos pesos temporales son los pesos finales. Considerar múltiples opciones para el cuantil asegura que no nos falte un buen vector de peso para el peso a la baja simplemente por la elección de un cuantil fijo para todos los individuos. La elección del cuantil para la estimación de Prob(C = c|RGH) se realiza para cada individuo mediante la observación de qué cuantil pondera más a la baja los recuentos.

Para determinar los pesos temporales, el método 100 inicialmente determina los pesos en bruto $\{rj_{c=0\ n}$ con base en las estimaciones de $Prob(\widehat{C = c})$, $\text{Prob}(\widehat{\text{C = c}}|\text{RGH})$, y $\text{Prob}(\widehat{\text{RGH}})$, donde r_c es el peso en bruto para un recuento de emparejamientos de c. En algunas realizaciones, los valores iniciales de los pesos en bruto $\{r\}_{c=0\ldots n}$ se determinan usando la ecuación 2. Posteriormente, el método 100 determina los pesos temporales a partir de los pesos en bruto para que los pesos temporales satisfagan las siguientes tres condiciones:

1. el peso temporal de una ventana con un emparejamiento tiene un peso de uno, es decir, $w_{c=1} = 1$;
2. los valores de los pesos temporales disminuyen monotónicamente al aumentar el recuento de emparejamientos c en una ventana específica, es decir, $w_c > w_{c+1}$ para todos los recuentos de emparejamientos c; y
3. los pesos temporales predeterminados a uno para todas las ventanas, si las estimaciones de Prob(C = c), Prob(C = c|RGH) o Prob(RGH) son pobres.

En algunas realizaciones, las estimaciones de estas probabilidades se consideran pobres si falla el algoritmo de optimización utilizado para determinar las distribuciones beta-binomiales, el número de recuentos de emparejamientos por ventana cae por debajo del valor umbral, por ejemplo, 20, o el número de puntos utilizados para ajustar la distribución beta-binomial está por debajo de un número mínimo.

Las dos primeras condiciones se pueden cumplir haciendo cumplir que el peso disminuya monotónicamente a medida que aumenta el recuento de emparejamientos en una ventana específica. La razón para un peso predeterminado de uno es evitar introducir más en lugar de eliminar el ruido en el cálculo de la puntuación de TIMBER, dado que la estimación de las probabilidades de los pesos en función de los recuentos de segmentos emparejados subrepresentados probablemente introduce ruido en el cálculo. Además, como en este caso solo se miden los recuentos bajos de emparejamientos por ventana, el método 100 tendría en consideración todos los recuentos de segmentos emparejados, que son principalmente recuentos bajos de emparejamientos por ventana, sin ponderar a la baja ningún recuento de segmentos emparejados en particular. Más concretamente, la estimación de las probabilidades usando las distribuciones beta-binomiales es limitada cuando se interpretan niveles bajos de recuento de emparejamientos.

En algunas realizaciones, los pesos temporales $\{w_c\}_{c=0\ldots n}$ están determinados por los siguientes dos pasos que son consistentes con las dos primeras condiciones anteriores, es decir, el primer peso temporal igual a uno y los pesos temporales disminuyendo monotónicamente al aumentar el recuento de emparejamientos. En el primer paso, el método 100 establece los pesos temporales en uno para todas las ventanas con un recuento de emparejamientos menor o igual que *M,* donde *M* es el recuento para el cual la proporción entre las dos distribuciones beta-binomiales estimadas es más alta. En el segundo paso, después del recuento para el que se "estimó" la Prob(RGH), cualquier aumento en el peso con respecto al aumento en el recuento de emparejamientos se cambia para ser un aumento cero. En algunas realizaciones, el método 100 realiza los dos pasos para todas las c aplicando el siguiente algoritmo (excepto cuando se aplica el peso predeterminado debido a estimaciones de probabilidad deficientes):

1.

$$\widetilde{w}_c = 1, \text{Si } c \leq M \qquad (10),$$

2.

$$d_c = \begin{cases} 0, si\ r_c > r_{c-1} \\ r_c - r_{c-1}, además \end{cases} \qquad (11),$$

$$\widetilde{w}_c = 1 + \sum_{i=M+1}^{c} d_i, \qquad \text{si c> } M \quad (12).$$

Las figuras 6 y 7 ilustran un ejemplo de los pesos temporales por ventana $\{w_c\}$ (eje y) en función del posible recuento de emparejamientos por ventana *c* (eje x). En particular, la figura 7 muestra más detalles que la figura 6, ya que el valor máximo del recuento de emparejamientos por ventana *c* mostrado a lo largo del eje x se establece en 40. En este ejemplo, *M* es 4.

II.C.5. Estimación de los pesos finales $\{w_i\}$

El método 100 incluye la determinación de los pesos finales $\{w_i\}_{i = 1...K}$ para ponderar todas las ventanas de segmentos emparejados en función de los pesos temporales $\{\widetilde{W}_{c,V}\}_{c=0...Cmax,V}$. En algunas realizaciones, el peso final W_i es los pesos temporales para una estimación dada de Prob(C = c|RGH,V), es decir, una V dada, que minimiza la suma del recuento ponderado de segmentos por ventana:

$$\{\overline{w}_c\}_{c=0...C_{max}} = \arg\min_{\{w_{c,V}\}_V} \sum_{i=1}^{K} k_i \cdot \widetilde{w}_{c=k_i,V} \qquad (13),$$

$$\{w_i\}_{i=1...K} = \left\{w_i : w_i = \overline{w}_{k_i}, i = 1 \ldots K\right\} \qquad (14).$$

En resumen, dadas las estimaciones de parámetros específicos para Prob(C = c), estimaciones múltiples de Prob(C = c|RGH), Prob(RGH) y algún procesamiento posterior, podemos encontrar el peso para una ventana dada en un individuo dado *i.* La figura 8 ilustra un ejemplo para los pesos (línea continua) dados los recuentos por ventana del ejemplo original en la figura 2 (línea discontinua). Solo para fines comparativos, el peso se vuelve a ampliar para que un peso de 1 tenga un valor de 59 de acuerdo con el eje y. El cálculo del peso genera para cada individuo un valor de peso (dentro del intervalo 0 y 1) para las *K* ventanas de segmento.

II.D. Calcular la suma ponderada de anchuras cM por ventana $\mathrm{cM}_2^{A,B}$

El método 100 incluye calcular 112 una suma ponderada de anchuras cM por ventana para cada segmento emparejado con base en la primera anchura cM y los pesos asociados con las ventanas de segmento del segmento emparejado, de acuerdo con algunas realizaciones. En particular, el método calcula la suma ponderada de las anchuras cM por ventana para un segmento emparejado, entre la persona A y la persona B, dados los pesos estimados de las ventanas específicas de cada persona para la persona A y la persona B. La suma ponderada de las anchuras cM por ventana $\mathrm{cM}_2^{A,B}$ es la suma de las primeras anchuras cM $cM_{1,i}$ para cada ventana que el segmento emparejado abarca ponderada por el producto de los pesos para ambos individuos, A y B, en esas ventanas de segmento:

$$cM_2^{A,B} = \sum_{i=WIN_{inicial}}^{WIN_{fin} \le K} w_i^A \cdot w_i^B \cdot cM_{1,i} \qquad (15),$$

para un segmento entre individuos *A* y *B,* comenzando en la ventana $WIN_{inicial}$ y terminando en la ventana WIN_{final} con w_{win}^A y w_{win}^B siendo los pesos para ambos individuos, respectivamente, en la ventana *i.* Si las ventanas son la ventana inicial o final del segmento emparejado, las anchuras de la ventana se actualizan para ser el primer marcador genético en la ventana o hasta el último marcador genético en la ventana, respectivamente.

Los pesos están en el intervalo entre 0 y 1 y pueden considerarse intuitivamente como una probabilidad de que esta ventana contribuya a la nueva "anchura". La consideración del producto de los pesos para los individuos *A* y B garantiza que la ventana sea válida en ambos individuos para poder contribuir a la suma ponderada de las anchuras cM por ventana. La primera suma ponderada de anchuras cM por ventana sería idéntica, si todos los pesos son iguales a uno. La nueva "anchura" o suma ponderada de las anchuras cM por ventana generalmente es más pequeña que la anchura cM en bruto o primera, y, por lo tanto, pondera a la baja esos emparejamientos en ventanas donde hay una gran cantidad de emparejamientos, ya sea en forma individual *A* o individual *B* para una gran población. Así, la ponderación a la baja da como resultado emparejamientos de peso a la baja, que son menos propensos a ser del historial genealógico reciente de los individuos. La figura 9 ilustra un ejemplo de recuentos de emparejamientos por ventana en un fragmento del genoma para un individuo tanto pre-TIMBER (línea discontinua) como post-TIMBER (línea continua).

III. PREDICCIÓN DE RELACIÓN ANCESTRAL

El método 100 incluye estimar 114 un grado de relación ancestral entre dos individuos con base en la suma ponderada de las anchuras cM por ventana de cada segmento emparejado entre los dos individuos, de acuerdo con algunas realizaciones. Si un par de individuos tiene una suma total de las primeras anchuras cM de menos de 60 cM, que elimina eficazmente los parientes cercanos de la evaluación del parentesco, la distancia de relación se predice en función de la suma de todas las sumas ponderadas de las anchuras cM por ventana de los segmentos compartidos. De otra manera, se utiliza la suma de las primeras anchuras cM. Esto da como resultado una predicción de relación que es más precisa para relaciones más distantes, siendo a la vez tan precisa para las relaciones cercanas. Esto es especialmente verdadero para ciertos grupos étnicos (por ejemplo, personas judías) y si están asignados a estar relacionados de forma lejana o no.

El método 100 usando, por ejemplo, el grado de parentesco ancestral, permite a las personas encontrar a sus familiares recientes y proporcionarles nueva información sobre su genealogía dentro de una red de familiares (con genealogía conocida). En algunas realizaciones, el grado de parentesco entre dos individuos representa una probabilidad de que los dos individuos estén relacionados ancestralmente y es igual a la suma ponderada de las anchuras cM por ventana de los segmentos emparejados de los individuos. En algunas realizaciones, el grado de parentesco entre dos individuos es una respuesta binaria sí o no si los dos individuos están relacionados ancestralmente en función de la suma ponderada de las anchuras cM por ventana. Por ejemplo, si la suma ponderada de las anchuras cM por ventana excede un valor umbral de parentesco, se dice que los dos individuos están relacionados ancestralmente. En algunas realizaciones, el valor umbral de parentesco es 20, 25, 30, 35, 40, 45, 50 cM o cualquier valor mayor de 20 cM. En algunas realizaciones, el valor umbral de parentesco es 30, 40 o 50 cM.

Previamente, la predicción de la distancia de relación entre dos individuos (por ejemplo, primos, primos terceros, etc.) se basó exclusivamente en la anchura total de todos los segmentos de IBD, donde la anchura del segmento de IBD se determinó por su anchura en la distancia de recombinación (en cM). En este método 100, se calcula una puntuación total con base en todos los segmentos de IBD entre dos individuos para proporcionar una distancia de relación. Sin embargo, el método 100 usa una suma de la suma ponderada reponderada de las anchuras cM por ventana en este cálculo para predecir el parentesco, si la suma de las primeras anchuras cM de menos de 60 cM.

En un ejemplo, los perfiles de peso se generan a partir de emparejamientos con un conjunto de referencia estático de poco más de 300K muestras. Esos pesos se utilizan para repuntuar, es decir, reponderar, todos los emparejamientos entre cualquier par de individuos en la base de datos. Los pesos utilizados por el método 100 se almacenan en su propia base de datos. En una realización de ejemplo, con base en el análisis de conjuntos de pruebas y datos reales, una anchura cM umbral de 5 cM fue la anchura mínima, en la que se incluye un segmento emparejado en la estimación de peso 110 y la suma ponderada del cálculo de anchuras cM por ventana 112.

IV. Ejemplo

En un ejemplo, el comportamiento de TIMBER se analizó con un conjunto de pruebas simuladas. El conjunto de pruebas consistió en exactamente 3703 pares de genotipos (7406 muestras individuales) que representan las relaciones de padre/hijo (1 meiosis) hasta los primos quintos (12 meiosis) y todos los intermedios. Cada relación se creó independientemente de todas las demás mediante simulaciones de meiosis para crear genotipos que representan a individuos en la parte relevante del pedigrí. Los "fundadores" de los que provienen los genotipos no simulados fueron un conjunto de aproximadamente 24.000 genotipos de la base de datos que no tienen relaciones cercanas estimadas entre ellos en función del análisis de las primeras anchuras cM no ponderadas. Los "fundadores" utilizados para crear una relación determinada se descartaron y no se utilizaron para simular ninguna de las otras relaciones. El hecho de que los fundadores iniciales tuvieran muy pocas relaciones genuinas y no fueran reutilizados ayudó a minimizar la probabilidad de relaciones entre genotipos sintéticos que los inventores no pudieron documentar, pero eso todavía era posible. El conjunto de prueba podría no ser ideal para un escenario del mundo real, ya que los individuos fueron emparejados al azar para ser parientes. Sin embargo, dado que en la simulación se conocían los segmentos reales de RGH, la simulación proporcionó una forma de verificar qué tan bien TIMBER ayudó a refinar el análisis de los segmentos emparejados. En la prueba, un poco más de 300 pares de cada nivel de meiosis estuvieron representados entre las relaciones padre-hijo (1 meiosis) y los primos quintos (12 meiosis).

TABLA 1: Resultados de TIMBER para diferentes cortes mínimos (= anchura cM umbral)

	Porcentaje de segmentos emparejados mantenidos		
Segmento	Min. de 5 cM	Min. de 6 cM	Min. de 7 cM
Verdadero	92	88	84
Falso	3	1	1

La Tabla 1 ilustra que TIMBER mantuvo la gran mayoría (alrededor del 90 %) de los segmentos de IBD descubiertos

inicialmente que se superponen con segmentos de IBD reales. Los resultados no variaron en gran medida con respecto a las diferentes primeras anchuras cM de 5, 6 y 7 y se usaron para determinar si un segmento de IBD descubierto se retuvo o no. TIMBER solo mantuvo como máximo el 3 % de los segmentos de IBD descubiertos inicialmente que son falsos positivos. Así, TIMBER presenta un filtro muy útil para mantener las señales reales mientras elimina las señales falsas positivas de los segmentos de IBD entre pares de individuos debido al historial genealógico reciente de los individuos. No se utilizó el filtrado "ibs" para filtrar los resultados en bruto y sin ponderar de los segmentos de IBD emparejados.

La tabla 2 muestra que TIMBER fue un poco más preciso para los parientes más cercanos, pero en general funcionó bien en todo el espectro del historial genealógico reciente, a partir de las relaciones entre padres e hijos (1 meiosis) hasta los primos quintos (12 meiosis). La figura 10 ilustra los resultados de TIMBER usando diferentes anchuras cM en bruto, es decir, primeros filtros de anchura cM, incluyendo los porcentajes emparejados de segmentos guardados para las meiosis conocidas y desconocidas.

TABLA 2: Resultados de TIMBER para segmentos verdaderos por meiosis y diferentes límites mínimos

	Porcentaje de segmentos emparejados mantenidos		
Meiosis	Min. de 5 cM	Min. de 6 cM	Min. de 7 cM
1	98	97	95
2	95	91	89
3	92	87	83
4	88	82	76
5	87	80	73
6	85	76	69
7	82	72	64
8	81	74	65
9	84	73	66
10	80	69	58
11	75	68	57
12	88	71	59

V. CONSIDERACIONES ADICIONALES

El sistema informático 120 se implementa usando uno o más ordenadores que tienen uno o más procesadores que ejecutan un código de aplicación para realizar los pasos descritos en el presente documento, y los datos pueden almacenarse en cualquier medio de almacenamiento no transitorio convencional y, donde corresponda, incluir una implementación de servidor de base de datos convencional. Por motivos de claridad y porque son bien conocidos por los expertos en la materia, varios componentes de un sistema informático, por ejemplo, procesadores, memoria, dispositivos de entrada, dispositivos de red y similares no se muestran en la figura 1B. En algunas realizaciones, se utiliza una arquitectura informática distribuida para implementar las características descritas. Un ejemplo de dicha plataforma informática distribuida es el proyecto Apache HADOOP® disponible de Apache Software Foundation.

Además de las realizaciones específicamente descritas anteriormente, los expertos en la materia apreciarán que la invención se puede practicar adicionalmente en otras realizaciones. Dentro de esta descripción escrita, el nombre particular de los componentes, la capitalización de términos, los atributos, las estructuras de datos o cualquier otro aspecto de programación o estructural no es obligatorio o significativo a menos que se indique lo contrario, y los mecanismos que implementan la invención descrita o sus características pueden tener diferentes nombres, formatos o protocolos. Además, el sistema puede implementarse mediante una combinación de hardware y software, como se describe, o completamente en elementos de hardware. También, la división particular de funcionalidad entre los diversos componentes del sistema descritos en el presente documento no es obligatoria; las funciones realizadas por un solo módulo o componente del sistema pueden ser realizadas por múltiples componentes, y las funciones realizadas por múltiples componentes pueden ser realizadas por un solo componente. Del mismo modo, el orden en que se realizan los pasos del método no es obligatorio a menos que se indique lo contrario o se requiera lógicamente. Cabe señalar que los pasos del proceso y las instrucciones de la presente invención podrían incorporarse en software, firmware o hardware, y cuando se incorporan en software, podrían descargarse para residir y ser puestos en funcionamiento desde diferentes plataformas utilizadas por los sistemas operativos de red en tiempo real.

Se entiende que las descripciones y representaciones algorítmicas incluidas en esta descripción son implementadas por programas informáticos. Además, también ha demostrado ser conveniente a veces, para referirse a estas disposiciones de operaciones como módulos o dispositivos de código, sin pérdida de generalidad.

A menos que se indique lo contrario, los debates que utilizan términos como "seleccionar" o "computar" o "determinar" o similares se refieren a la acción y los procesos de un sistema informático o dispositivo informático electrónico similar, que manipula y transforma datos representados como cantidades físicas (electrónicas) dentro de las memorias o registros del sistema informático u otro tipo de dispositivos de almacenamiento, transmisión o visualización de información.

La presente invención también se refiere a un aparato para realizar las operaciones en este documento. Este aparato puede construirse especialmente para los fines requeridos, o puede comprender un ordenador de propósito general activado selectivamente o reconfigurado por un programa informático almacenado en el ordenador. Dicho programa informático puede almacenarse en un medio de almacenamiento legible por ordenador no transitorio, tal como, aunque sin limitación, cualquier tipo de disco, incluidos los disquetes, discos ópticos, DVD, CD-ROM, discos magnéticos-ópticos, memorias de solo lectura (ROM), memorias de acceso aleatorio (RAM), EPROM, EEPROM, tarjetas magnéticas u ópticas, circuitos integrados de aplicación específica (ASIC), o cualquier tipo de medio adecuado para almacenar instrucciones electrónicas, y cada uno acoplado a un bus de sistema informático.

Además, los ordenadores a los que se hace referencia en la memoria desciptiva pueden incluir un único procesador o pueden ser arquitecturas que emplean diseños de múltiples procesadores para una mayor capacidad informática.

Los algoritmos y pantallas presentados no están inherentemente relacionados con ningún ordenador en particular u otro aparato. También se pueden usar varios sistemas de propósito general con programas de acuerdo con las enseñanzas anteriores, o puede resultar conveniente construir aparatos más especializados para realizar los pasos del método requeridos. La estructura requerida para una variedad de estos sistemas aparecerá a partir de la descripción anterior. Así mismo, se puede utilizar una variedad de lenguajes de programación para implementar las enseñanzas anteriores.

Finalmente, cabe señalar que el lenguaje utilizado en la memoria descriptiva se ha seleccionado principalmente para fines de legibilidad y de instrucción, y puede que no se haya seleccionado para delimitar o circunscribir la materia objeto inventiva. Por consiguiente, la divulgación de la presente invención pretende ser ilustrativa, aunque sin limitación, del alcance de la invención.

REIVINDICACIONES

1. Un producto de programa informático para estimar un grado de parentesco ancestral entre dos individuos, el producto de programa informático almacenado en un medio legible por ordenador no transitorio y que incluye instrucciones configuradas para hacer que un procesador ejecute pasos que comprenden:

recibir datos de haplotipos para una población de individuos, incluyendo los datos de haplotipos una pluralidad de marcadores genéticos compartidos entre los individuos;
dividir los datos de haplotipos en ventanas de segmento con base en los marcadores genéticos; para cada individuo en la población:

con base en los marcadores genéticos, emparejar segmentos de los datos de haplotipos que son idénticos entre el individuo y cualquier otro individuo en la población, cada segmento emparejado tiene una primera anchura cM que excede una anchura cM umbral y es parte de una o más de las ventanas de segmento;
contar los segmentos emparejados en cada ventana de segmento;
estimar un peso asociado con cada ventana de segmento con base en el recuento de segmentos emparejados en la ventana de segmento asociada;
calcular una suma ponderada de anchuras cM por ventana para cada segmento emparejado en función de la primera anchura cM y los pesos asociados con las ventanas de segmento del segmento emparejado; y

estimar un grado de parentesco ancestral entre dos individuos en función de la suma ponderada de las anchuras cM por ventana de cada segmento emparejado entre los dos individuos.

2. El producto de programa informático de la reivindicación 1, en donde la anchura cM umbral es 5 cM, 6 cm, 7 cM, 8 cM, 9 cM, 10 cM, o cualquier número real dentro del intervalo de 5 cM a 10 cM.

3. El producto de programa informático de la reivindicación 1, en donde el peso asociado con una ventana de segmento para el individuo A se aproxima como:

$$w_i^A = \frac{\text{Prob}\left(\widehat{C = k_i}\middle|\text{RGH}\right)\text{Prob}(\text{RGH})}{\text{Prob}\widehat{(C = k_i)}},$$

en donde $\text{Prob}(\widehat{\text{C} = \text{c}}|\text{RGH})$, $\text{Prob}\widehat{(\text{RGH})}$, $\text{Prob}\widehat{(\text{C} = \text{c})}$ son las estimaciones de una probabilidad de un segmento de RGH dado el recuento medido c de segmentos emparejados en una ventana, una probabilidad de un segmento de RGH en una ventana, y la probabilidad de medir un recuento c de segmentos emparejados en una ventana, respectivamente.

4. El producto de programa informático de la reivindicación 3, en donde $\text{Prob}\widehat{(\text{RGH})}$ se aproxima por el máximo de:

$$\frac{1}{\frac{\text{Prob}\left(\widehat{C = c}\middle|\text{RGH}\right)}{\text{Prob}\widehat{(C = c)}}}.$$

5. El producto de programa informático de la reivindicación 1, en donde la suma ponderada de anchuras cM por ventana para un segmento entre dos individuos *A* y *B* se aproxima como:

$$cM\binom{A,B}{2} = \sum_{i=WIN_{inicial}}^{WIN_{fin} \leq K} w_i^A \cdot w_i^B \cdot cM_{1,i},$$

el segmento entre los individuos *A* y *B* comienza en la ventana $WIN_{inicial}$ y termina en la ventana WIN_{final} con w_{win}^A y w_{win}^B siendo los pesos asociados con la ventana de segmento *i* para los individuos A y B, respectivamente.

6. El producto de programa informático de la reivindicación 1, en donde la estimación de un peso comprende calcular pesos temporales $\widetilde{w}_c$ para el recuento *c* de segmentos emparejados en la ventana del segmento asociado se aproxima como:

$$\widetilde{w}_c = 1, \text{ Si } c \leq M$$

$$d_c = \begin{cases} 0, si\ r_c > r_{c-1} \\ r_c - r_{c-1}, además \end{cases},$$

$$\widetilde{w}_c = 1 + \sum_{i=M+1}^{c} d_i, \qquad \text{si c> } M,$$

donde r_c es el peso con base en el recuento de segmentos emparejados c y aproximado como:

$$r_c = \frac{\text{Prob}\left(\widehat{C = c}\middle|\text{RGH}\right)\text{Prob}(\text{RGH})}{\text{Prob}\widehat{(C = c)}} .$$

7. El producto de programa informático de la reivindicación 1, en donde los pesos asociados con cada ventana de segmento disminuyen si aumenta el recuento de segmentos emparejados en la ventana de segmento asociada.

8. El producto de programa informático de la reivindicación 3 o 6, en donde $\text{Prob}\widehat{(C = c)}$ se aproxima como:

$$\text{Prob}\widehat{(C = c|c > 0)} = \binom{n}{c}\frac{\text{B}(c+\alpha, n-c+\beta)}{\text{B}(\alpha,\beta)},$$

donde *n* es el tamaño de la población y α, β son parámetros de la función Beta B.

9. El producto de programa informático de la reivindicación 3 o 6, en donde Prob(C = c|RGH, V) se aproxima como:

$$\text{Prob}(C = c|\text{RGH}, V, c > 0) = \binom{n}{c}\frac{\text{B}(c+\alpha, n-c+\beta)}{\text{B}(\alpha,\beta)} \Big/ \text{Prob}(0 < C \leq V|\alpha,\beta),$$

donde *n* es el tamaño de la población, α, β son parámetros de la función Beta B.

10. El producto de programa informático de la reivindicación 8, en donde α y β se estiman usando una estimación de máxima verosimilitud de una distribución conjunta aproximada como:

$$f(\{k_i\}|n,\alpha,\beta,k_i > 0) = \prod_{i=1}^{K}\binom{n}{k_i - 1}\frac{\text{B}(k_i - 1 + \alpha, n - k_i + 1 + \beta)}{\text{B}(\alpha,\beta)}.$$

11. El producto de programa informático de la reivindicación 9, en donde α y β se estiman usando una estimación de máxima verosimilitud de una distribución conjunta aproximada como:

$$f(\{m_i\}|n,\alpha,\beta,V,m_i > 0) = \frac{\prod_{i=1}^{M}\binom{n}{m_i - 1}\frac{\text{B}(m_i - 1 + \alpha, n - m_i + 1 + \beta)}{\text{B}(\alpha,\beta)}}{\text{Prob}(0 < C \leq V|\alpha,\beta)} .$$

12. El producto de programa informático de la reivindicación 1, en donde un tamaño de las ventanas de segmento comprende 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150 marcadores genéticos o cualquier número que se encuentre dentro del intervalo de 50 a 500 marcadores genéticos.

13. El producto de programa informático de la reivindicación 3, en donde $\text{Prob}\left(\widehat{C = c}\middle|\text{RGH}\right)$ está determinado por:

generar una primera estimación de Prob *(C* = c|RGH) con base en un primer valor de cuantil;
generar una segunda estimación de Prob (C = c|RGH) con base en un segundo valor de cuantil; y
seleccionar una de la primera estimación o la segunda estimación de Prob (C = c|RGH), en donde los primero y segundo

valores de cuantil son 75 % y 90 %, respectivamente, o el primer valor de cuantil es uno del 50 %, 55 %, 60 %, 65 % o
70 % y el segundo valor de cuantil es uno del 65 %, 70 %, 75 %, 80 % u 85 %.

14. El producto de programa informático de la reivindicación 1, que comprende además instrucciones configuradas para hacer que un procesador ejecute pasos que comprenden:
eliminar parientes cercanos de la población de individuos, en donde opcionalmente los parientes cercanos comprenden dos individuos que tienen una primera anchura cM total mayor que 60 cM.

15. El producto de programa informático de la reivindicación 1, en donde el grado de parentesco entre dos individuos comprende uno de probabilidad de que los dos individuos estén relacionados ancestralmente o una respuesta binaria sí o no si los dos individuos están relacionados ancestralmente.

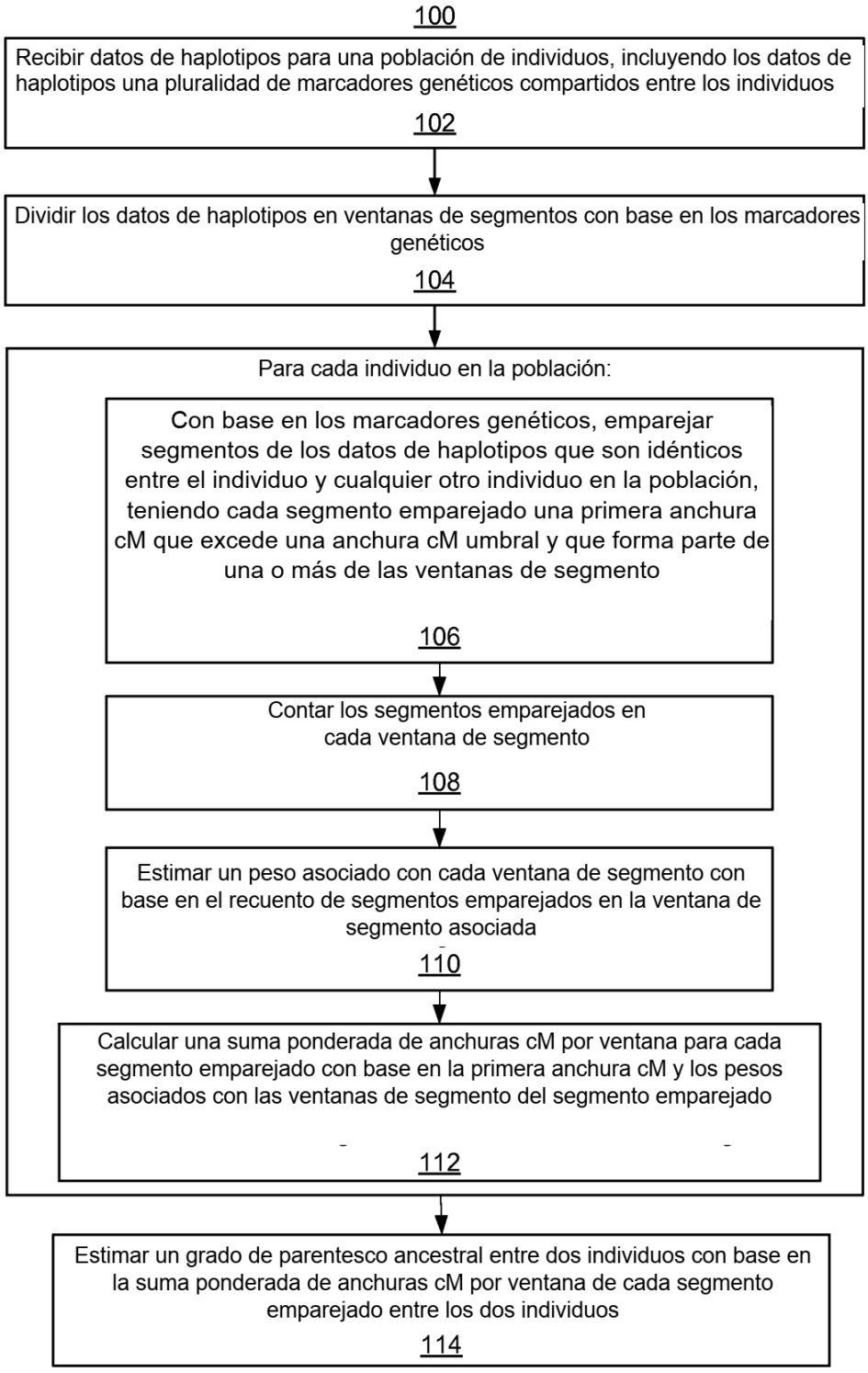
100
Recibir datos de haplotipos para una población de individuos, incluyendo los datos de haplotipos una pluralidad de marcadores genéticos compartidos entre los individuos
102
Dividir los datos de haplotipos en ventanas de segmentos con base en los marcadores genéticos
104
Para cada individuo en la población:
Con base en los marcadores genéticos, emparejar segmentos de los datos de haplotipos que son idénticos entre el individuo y cualquier otro individuo en la población, teniendo cada segmento emparejado una primera anchura cM que excede una anchura cM umbral y que forma parte de una o más de las ventanas de segmento
106
Contar los segmentos emparejados en cada ventana de segmento
108
Estimar un peso asociado con cada ventana de segmento con base en el recuento de segmentos emparejados en la ventana de segmento asociada
110
Calcular una suma ponderada de anchuras cM por ventana para cada segmento emparejado con base en la primera anchura cM y los pesos asociados con las ventanas de segmento del segmento emparejado
112
Estimar un grado de parentesco ancestral entre dos individuos con base en la suma ponderada de anchuras cM por ventana de cada segmento emparejado entre los dos individuos
114

FIG. 1A

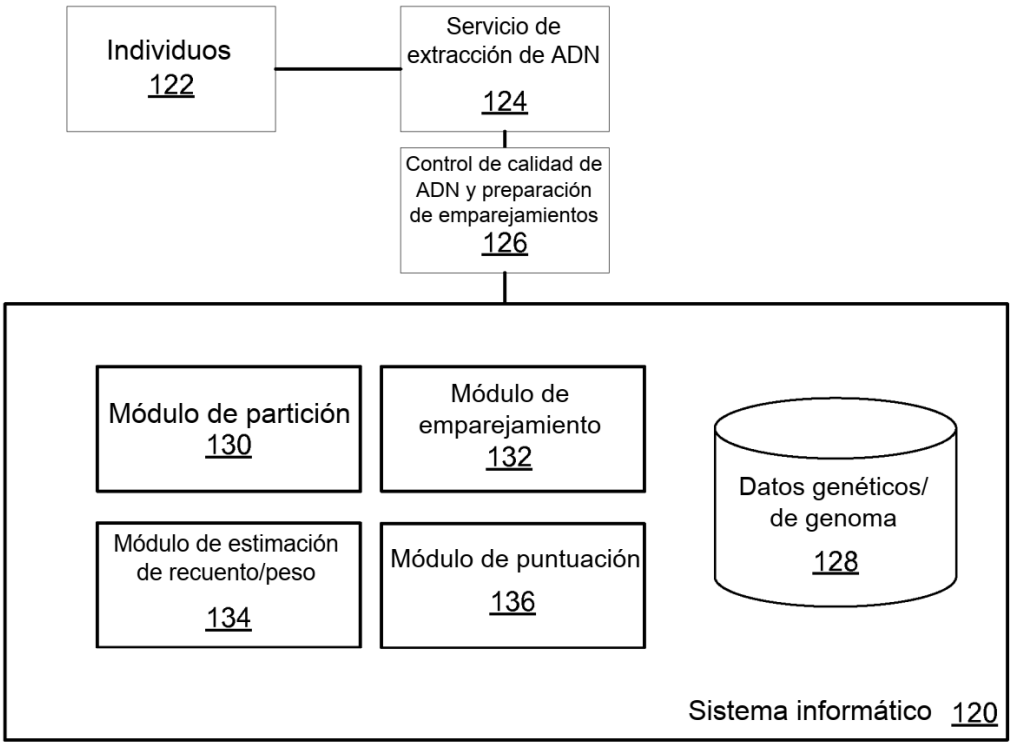
Individuos
122
Servicio de extracción de ADN
124
Control de calidad de ADN y preparación de emparejamientos
126
Módulo de partición
130
Módulo de emparejamiento
132
Módulo de estimación de recuento/peso
134
Módulo de puntuación
136
Datos genéticos/ de genoma
128
Sistema informático 120

FIG. 1B

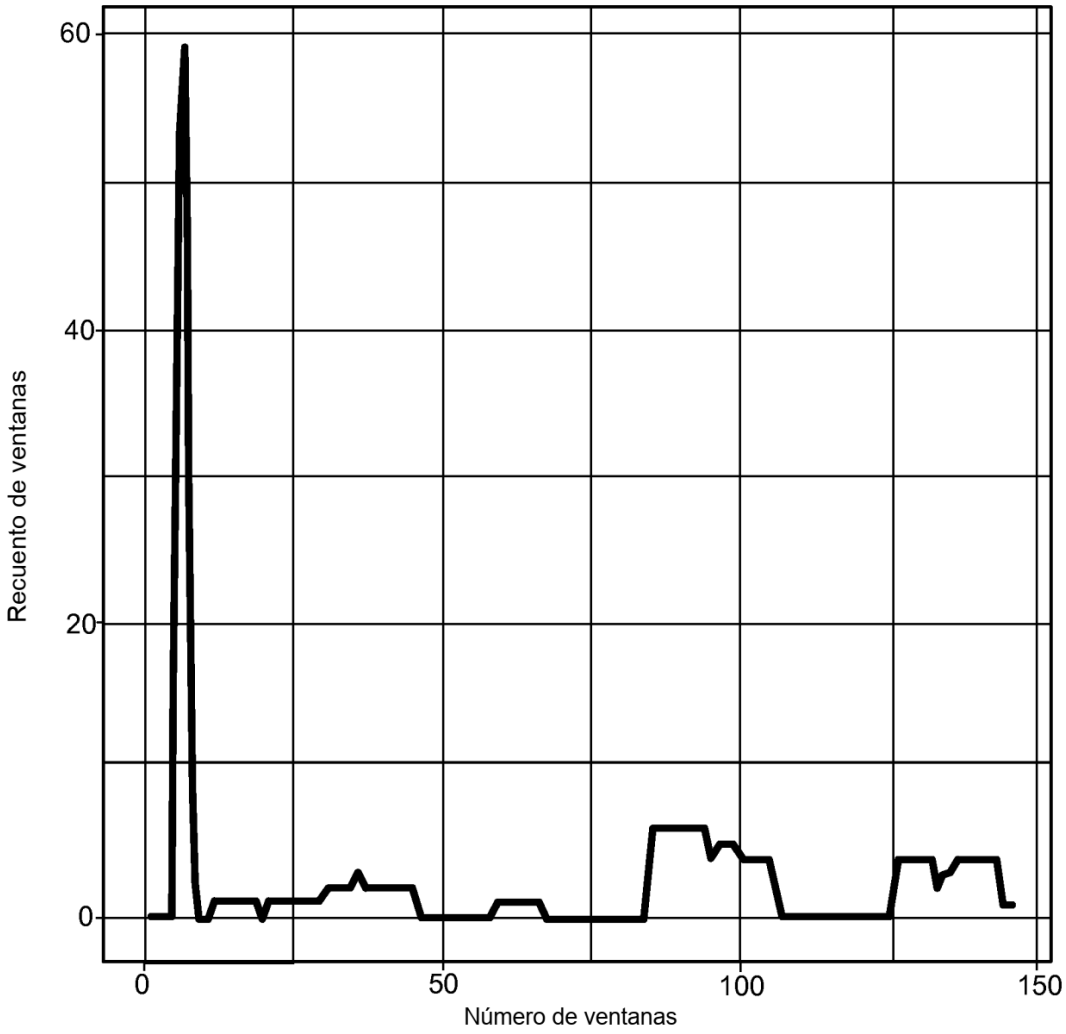
60
40
20
0
Recuento de ventanas
0
50
100
150
Número de ventanas

FIG. 2

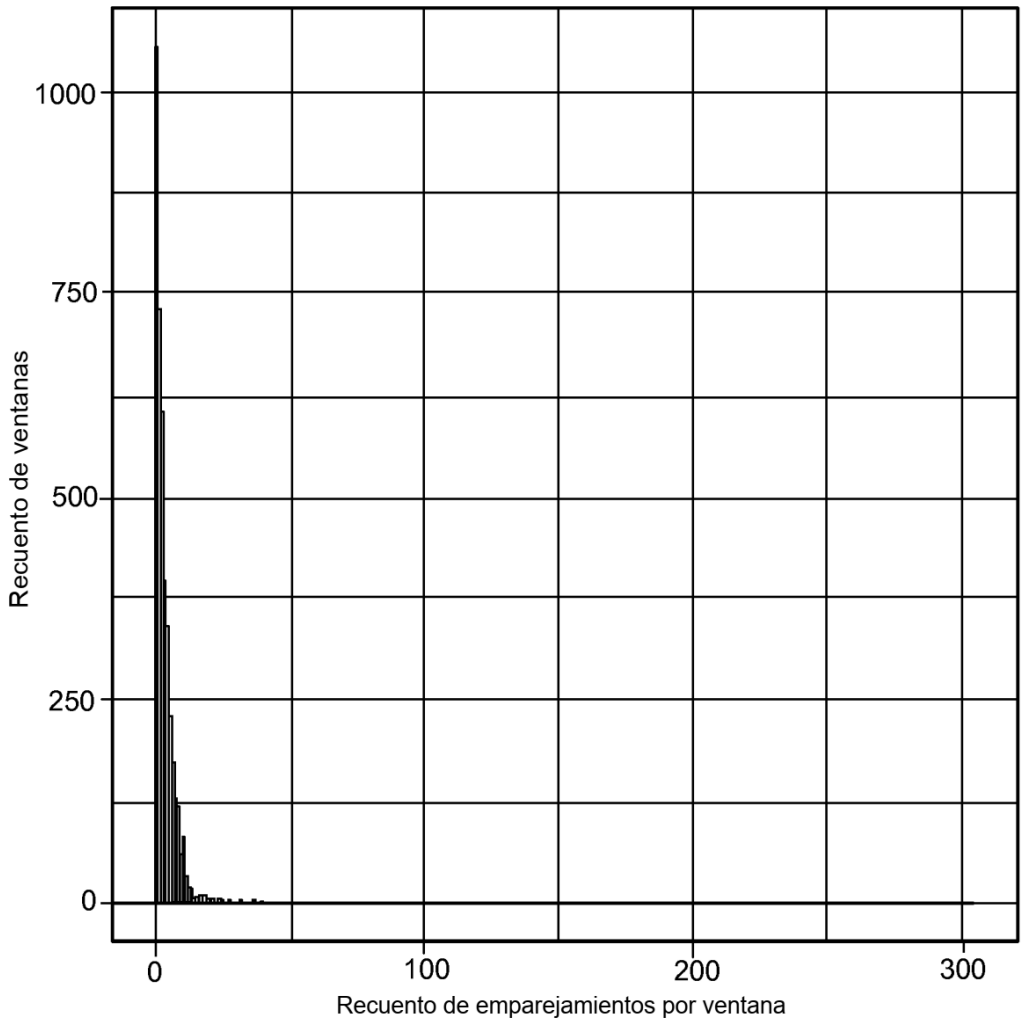
1000
750
500
250
0
Recuento de ventanas
0
100
200
300
Recuento de emparejamientos por ventana

FIG. 3

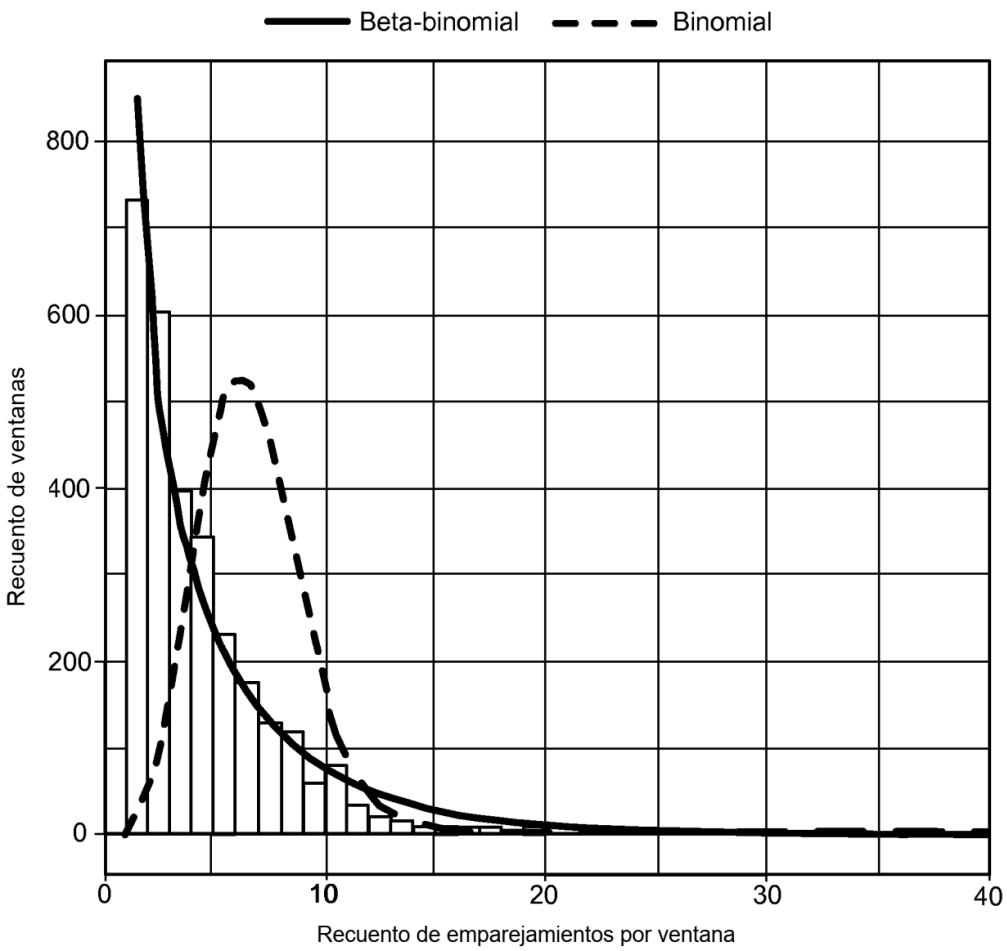
Beta-binomial
Binomial
800
600
400
200
0
Recuento de ventanas
0
10
20
30
40
Recuento de emparejamientos por ventana

FIG. 4

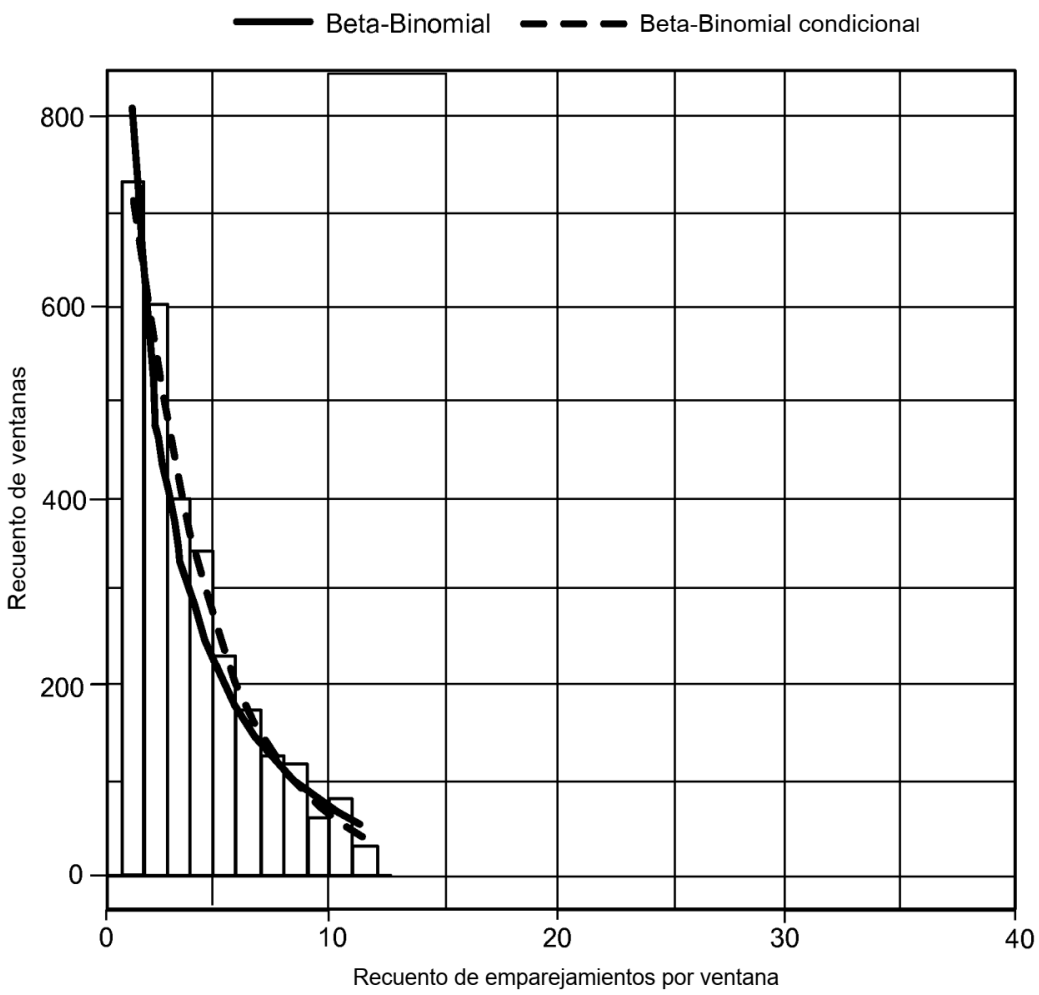
Beta-Binomial
Beta-Binomial condicional
800
600
400
200
0
Recuento de ventanas
0
10
20
30
40
Recuento de emparejamientos por ventana

FIG. 5

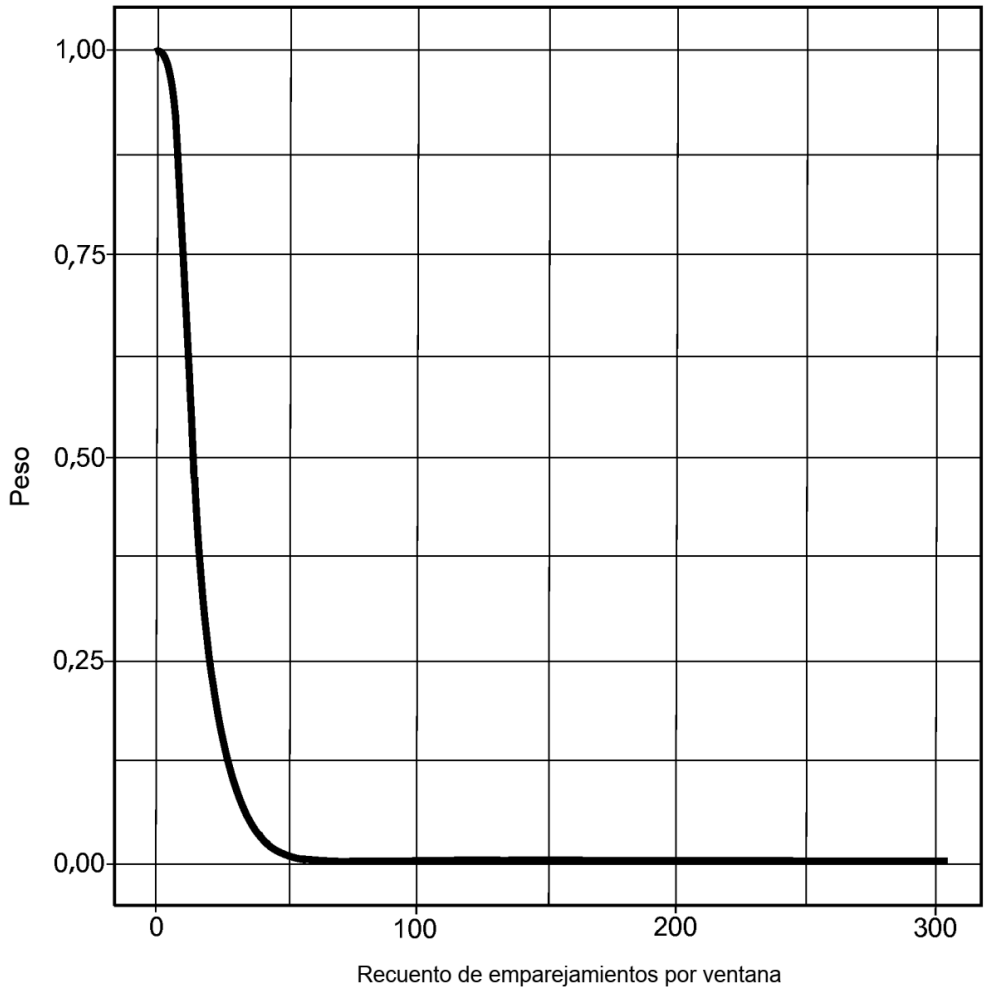
1,00
0,75
0,50
0,25
0,00
Peso
0
100
200
300
Recuento de emparejamientos por ventana

FIG. 6

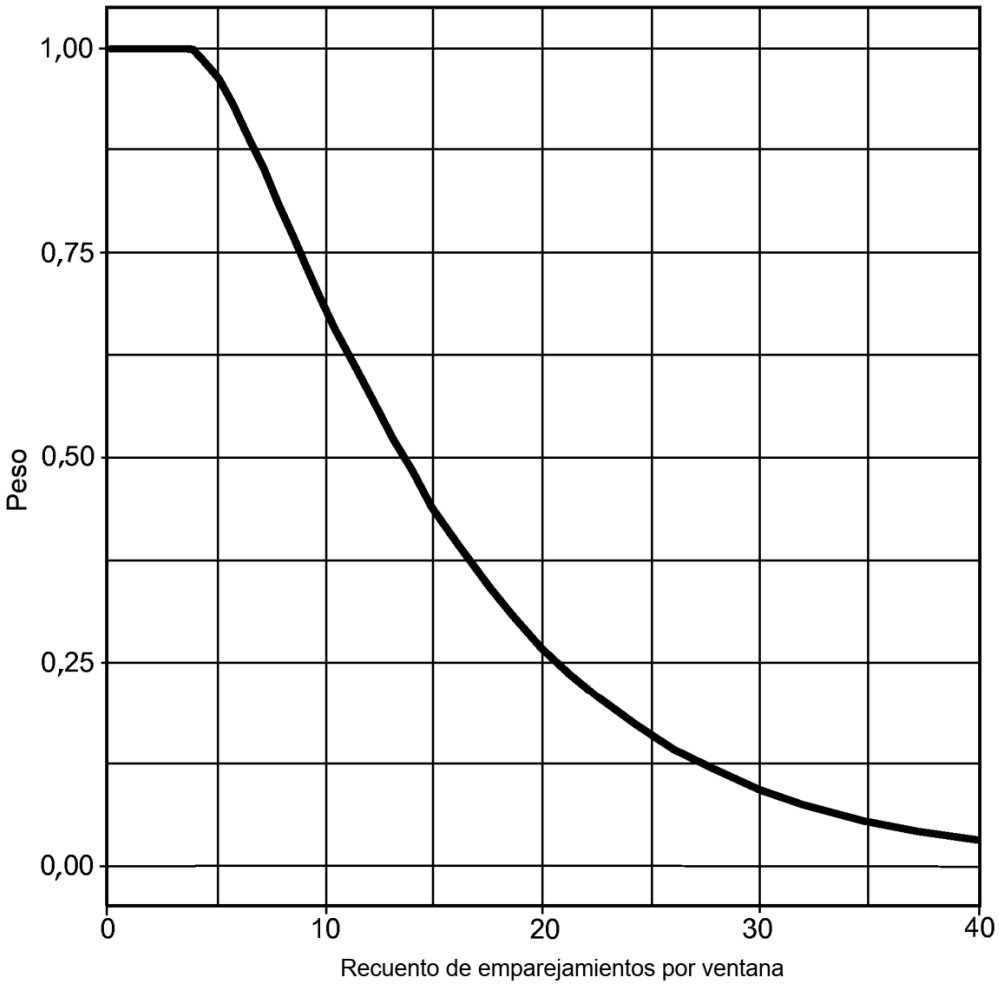
1,00
0,75
0,50
0,25
0,00
Peso
0
10
20
30
40
Recuento de emparejamientos por ventana

FIG. 7

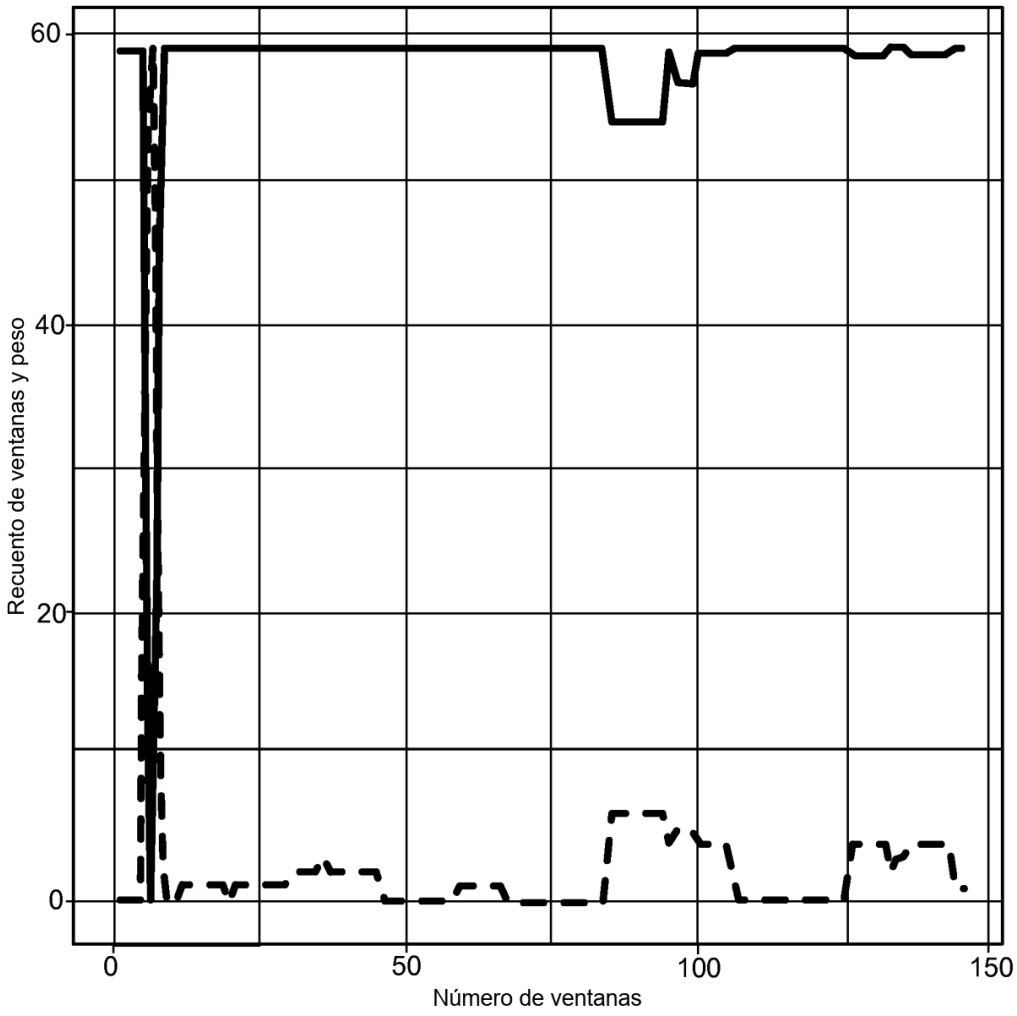
60
40
20
0
Recuento de ventanas y peso
0
50
100
150
Número de ventanas

FIG. 8

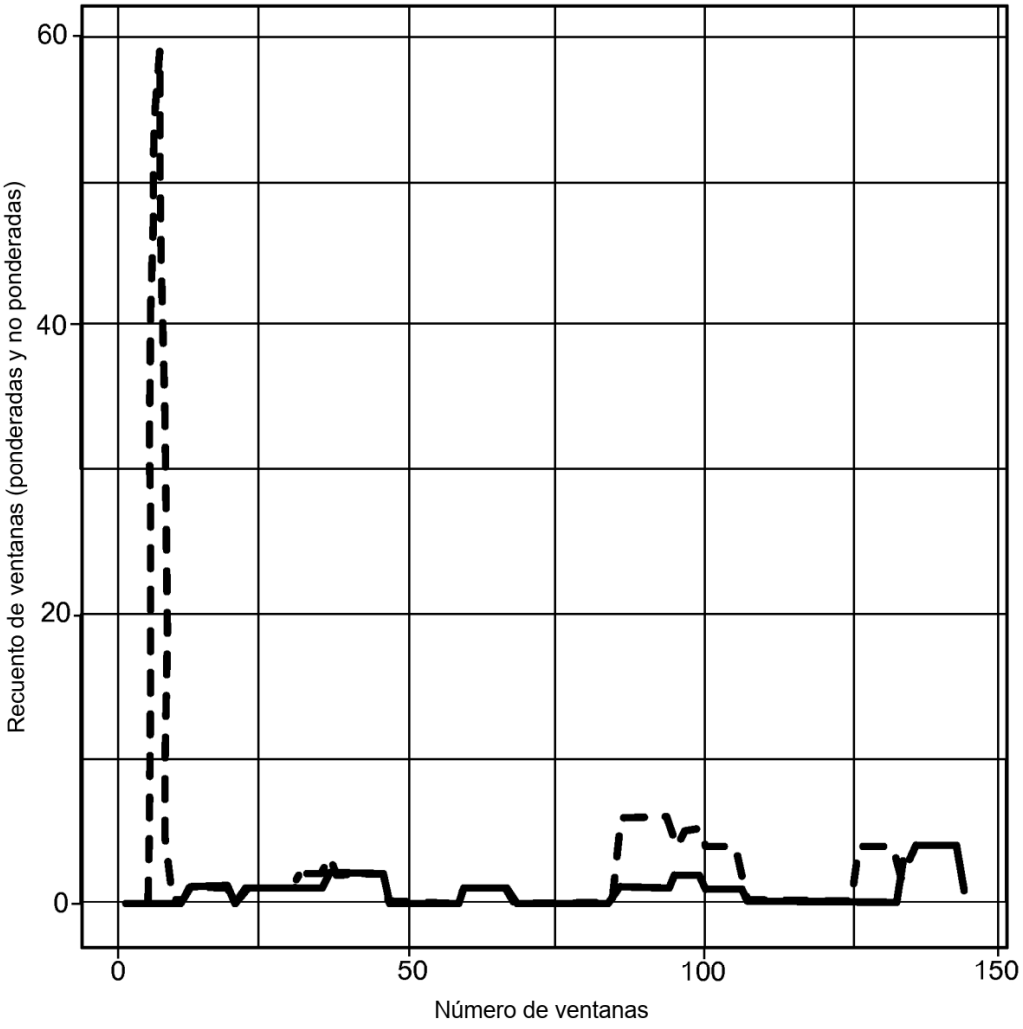
60
40
20
0
Recuento de ventanas (ponderadas y no ponderadas)
0
50
100
150
Número de ventanas

FIG. 9

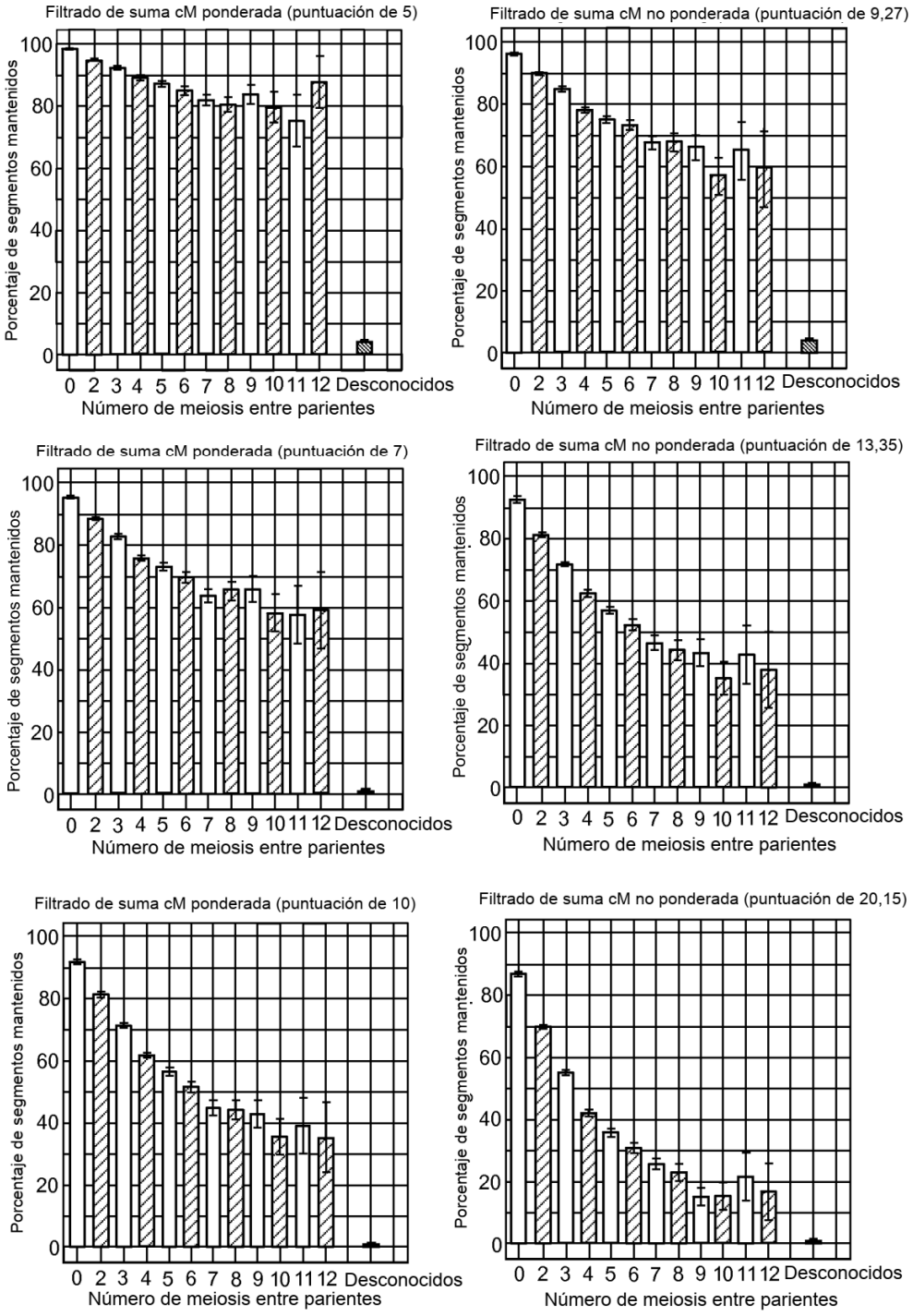
Filtrado de suma cM ponderada (puntuación de 5)
Filtrado de suma cM no ponderada (puntuación de 9,27)
Filtrado de suma cM ponderada (puntuación de 7)
Filtrado de suma cM no ponderada (puntuación de 13,35)
Filtrado de suma cM ponderada (puntuación de 10)
Filtrado de suma cM no ponderada (puntuación de 20,15)
Porcentaje de segmentos mantenidos
0 20 40 60 80 100
0 2 3 4 5 6 7 8 9 10 11 12 Desconocidos
Número de meiosis entre parientes

FIG. 10