(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 701 872**

(51) Int. CI.:

A01K 67/00 (2006.01)
A01K 67/02 (2006.01)
C12Q 1/68 (2008.01)
G01N 33/48 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **21.12.2007 PCT/AU2007/002006**

(87) Fecha y número de publicación internacional: **26.06.2008 WO08074101**

(96) Fecha de presentación y número de la solicitud europea: **21.12.2007 E 07855371 (6)**

(97) Fecha y número de publicación de la concesión europea: **26.09.2018 EP 2120543**

(54) Título: **Métodos y reactivos de selección artificial**

(30) Prioridad:

21.12.2006 US 876623 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
26.02.2019

(73) Titular/es:

AGRICULTURE VICTORIA SERVICES PTY LIMITED (100.0%)
475 Mickleham Road
Attwood, VIC 3049, AU

(72) Inventor/es:

HAYES, BEN y
GODDARD, MICHAEL

(74) Agente/Representante:

ISERN JARA, Jorge

DESCRIPCIÓN

Métodos y reactivos de selección artificial

Campo de la invención

La presente invención se refiere en general al campo de la selección artificial, que incluye la reproducción de animales y plantas que son importantes desde un punto de vista comercial, y más específicamente a métodos y reactivos para la selección asistida por marcadores en animales y plantas.

Antecedentes de la invención

Los programas de selección artificial se ocupan principalmente de aumentar la ganancia genética en virtud de las aportaciones de más genes procedentes de progenitores "buenos". Los medios tradicionales para determinar la ganancia genética expresan la ganancia como el producto de la intensidad de selección, la precisión y la desviación estándar genética definidas en una sola generación. Woolliams et al, Genetics 153, 1009-1020 (1999) mostraron que el proceso de aportación de genes a una población implica a más de una sola generación y que la ganancia sostenida depende de la variación del muestreo Mendeliano que se introduce en la población en cada generación. En pocas palabras, la ganancia genética de la selección artificial estará relacionada con la aportación genética a largo plazo de un progenitor a la población, así como con el valor reproductivo marginal de un individuo, vinculando así la ganancia genética con el desarrollo de árbol genealógico.

Durante siglos, la selección artificial se ha basado por completo en el fenotipo. Si bien esto ha demostrado ser útil, requiere mucho tiempo y es costoso. En particular, la selección artificial basada en el fenotipo puede usar pruebas de progenitores en las que el valor reproductivo estimado de un individuo se determina realizando múltiples apareamientos del individuo y determinando el rendimiento de la descendencia para un rasgo o carácter fenotípico particular. Por ejemplo, Schaeffer J. Anim. Raza. Genet 123, 218-223 (2006) estimó que el tiempo necesario para acreditar un toro Holstein lleva aproximadamente 64 meses desde la concepción hasta la primera prueba, suponiendo un período de gestación de 9 meses y que los toros jóvenes se aparean al año de vida y las hembras a los 15 meses de vida. En este ejemplo, el coste total para acreditar un toro se estimó en alrededor de 40.000 dólares americanos, incluyendo el coste de alojamiento y alimentación del toro, la extracción y conservación de semen, pruebas de apareamiento y clasificación de las hijas. Sin embargo, el coste para una compañía de inseminación artificial que adquiere becerros jóvenes en grandes cantidades para su uso como sementales sería mucho mayor, aunque compensado por el regreso de cualquier toro joven a la monta.

La genómica ha brindado la posibilidad de una selección artificial basada en el genotipo. Una secuencia hologenómica (de todo el genoma) para una especie permite la construcción de cualquier número de microplacas o micromatrices de ADN de aproximadamente 10.000 o más ácidos nucleicos, cada uno de los cuales comprende un marcador polimórfico. El conocimiento de alelos, genes, polimorfismos, haplotipos o haplogrupos, etc informativos, para un QTL (acrónimo del inglés *quantitative trait locus*, locus de carácter cuantitativo) o rasgo particular, facilita la selección de individuos o germoplasma y realizar las estimaciones de su EBV (acrónimo del inglés e*xpected breeding value*, valor de reproducción estimado). Esto se debe a que la selección genotípica se basa en la capacidad de genotipar a individuos para determinar genes o marcadores específicos que están en equilibrio de ligamiento (marcadores dispersos) o desequilibrio de ligamiento (marcadores densos) con un QTL particular u otro locus de interés de tal manera que el valor de reproducción de un individuo puede estimarse utilizando marcadores de haplotipos asociados con el QTL u otro locus. La selección genotípica es especialmente poderosa cuando la selección es deseable o necesariamente independiente de la expresión, por ejemplo, en el caso de la selección de rasgos de producción de leche en animales macho. La selección genotípica puede no estar basada en un árbol genealógico, cuando las asociaciones genotípicas en las que se basa proceden de una población actual o, en el caso de mapas de marcadores dispersos, cuando las asociaciones genotípicas proceden de datos de familias grandes de medio hermanos o cruces limitados.

La selección genotípica de "mejores" individuos puede basarse en una puntuación asignada a un alelo, gen, polimorfismo, haplotipo o haplogrupo, etc informativo. del individuo solo, o en tándem con un EBV basado en fenotipo o un EBV basado en genotipo. Se prefieren múltiples bases para la selección para minimizar la pérdida en respuesta a poligenes u otros QTL. Walsh Theor. Population Biol 59, 175-184 (2001) también sugirió que el fenotipo debería seguir siendo un componente en la selección, para capturar la variación que surge de nuevas mutaciones y para evitar reducciones drásticas en el tamaño efectivo de la población, la variación mutua acumulada de la deriva genética aleatoria y la tasa de respuesta prolongada a la selección que, de otro modo, surgiría de la selección dirigida a genotipos específicos.

La selección genotípica se facilita mediante medios informáticos, incluidas las estrategias de remuestreo, por ejemplo, ensayos de asignación al azar y remuestreo por reposición (*bootstrapping*), que permiten la construcción de intervalos de confianza y ensayos de significación adecuados, por ejemplo, los mejores indicadores lineales no sesgados (BLUP; Henderson en: " Applications of Linear Models in Animal Breeding", Universidad de Guelph, Guelph, Ontario, Canadá; Lynch y Walsh, En:" Genetics and Analysis of Quantitative Traits", Sunuaer Associates, Sunderland MA, USA, 1998); la estrategia de Monte Carlo de Cadenas de Markov (MCMC) (Geyer et al., Stat. Sci. 7,

73-511, 1992; Tierney et al, Ann. Statist. 22, 1701-1762, 1994; Tanner et al., en: " "Tools for Statistical Analysis", SpringerVerlag", Springer-Verlag, Berlín / Nueva York, 1996); el muestreador de Gibbs (Geman et al, IEEE Trans. Pattern Anal. Mach. Intell 6, 721-741, 1984); Distribución posterior bayesiana (por ej., Smith et al, J. Royal Statist. Soc. Ser. B55, 3-23, 1993). Bajo el análisis bayesiano, las probabilidades semisubjetivas en cuanto a un parámetro poblacional se asignan a incertidumbres y después se analizan y refinan con la experiencia, lo que permite que una creencia previa sobre un parámetro poblacional se actualice a una creencia posterior. Por ejemplo, Sillanpaa y Arjas, Genetics 148, 1373-1388 (1998); Sillanpaa y Arjas, Genetics 151, 1605-1619 (1999); y Stephens y Fisch, Biometrics 54, 1334-1347 (1998), han propuesto métodos bayesianos basados en el remuestreo para el mapeo de QTL múltiples. Meuwissen et al. .Genetics 157, 1819-1829 (2001) simularon un genoma de 1000 cM con marcadores que se supone que están en desequilibrio de ligamiento separados por 1 cM a lo largo del genoma, de modo que los marcadores se combinaron en pares de haplotipos que rodeaban cada región de 1 cM, y compararon los mínimos cuadrados, estrategias BLUP y Bayesianas para estimar los efectos de cada par de haplotipos simultáneamente (50,00 efectos de haplotipos en total), es decir, para toda la población y no específicos de ningún individuo; los autores demostraron que el EBV agregado podría determinarse para la descendencia siempre que esos animales se genotipasen y que los haplotipos marcadores se determinasen con una precisión de 0,75-0,85 para todas las estrategias. En esta simulación, se asumió que el tamaño efectivo de la población era constante.

Los mapas de marcadores dispersos se pueden construir utilizando marcadores en equilibrio de ligamiento y separarse aproximadamente por 20 cM, en función de datos de familias grandes de medios hermanos o cruces limitados. Por ejemplo, Georges et al., Genetics 139, 907-929 (1995) prepararon un mapa genético disperso de marcadores genéticos que resultó en la detección de algunos QTL para la producción de leche, y la inclusión de información de marcadores en los valores de reproducción de BLUP predijo una ganancia de 8-38 % (Meuwissen y Goddard, Genet. Sci. Evol 28, 161-176 (1996). Sin embargo, la utilidad de dicha información está limitada en las poblaciones exogámicas porque la fase de ligamiento entre un marcador y un QTL debe establecerse para cada una y cada familia en la que se va a usar el marcador para la selección. En consecuencia, con las estrategias de mapeo disperso conocidas, hay problemas de implementación significativos.

Los mapas de marcadores densos, generalmente construidos a partir de polimorfismos mononucleotídicos (SNP, forma siglada de *single-mononucleotide polymorphism*) y/o microsatélites, proporcionan el mapeo de locus de rasgos cuantitativos (QTL), estudios de asociación y calcula la relación entre individuos en una muestra de una población. Con los mapas de marcadores densos, es más probable que los marcadores estén en desequilibrio de ligamiento con un QTL y, por lo tanto, se asocien más positivamente con un rasgo cuantitativo de interés que para un mapa disperso, de modo que la selección no requiera que se establezca una fase de vinculación para cada familia. Los marcadores en desequilibrio de ligamiento generalmente están dentro de aproximadamente 1 cM a 5cM de un locus de interés. Además, la identificación de marcadores de desequilibrio de ligamiento requiere genes candidatos (Rothschild y Soller, Probe 8, p13, 1997) o estrategias de mapeo fino (Anderson et al, Nature Reviews Genet. 2, 130-138, 2001). Por lo tanto, para un genoma de aproximadamente 3000 cM (centimorgan), se necesitan aproximadamente 3001 marcadores a intervalos de 1 cM o intervalos mayores.

A pesar de la capacidad teórica de producir mapas hologenómicos de marcadores densos, que en teoría abarcan genomas completos, existen varias limitaciones en la aplicación de dicha tecnología. Debido a que existe un requisito absoluto para que los marcadores en dichos mapas sean informativos, la cantidad real de marcadores requeridos es mucho mayor que un mínimo teórico. Además, existe una necesidad de construir haplotipos heredados del/de los padre(s) para cada par contiguo de marcadores bialélicos, uno de los cuatro haplotipos informativos posibles se vinculará con un solo QTL en promedio, y las frecuencias de cada haplotipo variarán dependiendo de la frecuencia de cada alelo contribuyente, así como de la distancia entre los marcadores. Esto significa que para garantizar que todos los haplotipos estén representados y se determinen sus efectos, se debe genotipar a suficientes animales. El requisito de marcadores densos significa que, dependiendo del tamaño del genoma, también aumentará la cantidad de animales requeridos. Finalmente, no existen mapas de marcadores densos para todas las especies.

El alto coste del genotipado hace que no sea factible implementar todos los marcadores disponibles en los genomas de la mayoría de las especies. Dichos costes surgen de la asociación inicial de los efectos de haplotipos, que se correlaciona con la restricción mencionada en el párrafo anterior, y el coste unitario del genotipado de un individuo para estimar su valor de reproducción. Por ejemplo, en el caso del ganado, Schaeffer J. Anim. Breed. Genet 123, 218-223 (2006) ha estimado que se requeriría un mínimo de aproximadamente 10.000 marcadores en un mapa de marcadores densos de genoma completo, y que el coste unitario aproximado de genotipar a un animal para este número de marcadores de SNP es de aproximadamente 400 dólares americanos. El coste unitario real se compara desfavorablemente con lo que sería aceptable para la industria, es decir, alrededor de 20-200 dólares americanos por animal. Sin embargo, si asumimos que los efectos del haplotipo proceden de 50 familias de padres con 50 hijos cada una, el coste se aproxima a los 1.000.000 dólares americanos. Este coste naturalmente aumentará si se genotipan individuos adicionales, por ejemplo, hijas de los hijos en las pruebas, de acuerdo con la práctica estándar. Por lo tanto, inicializar un esquema de genoma completo utilizando mapas de marcadores densos es costoso de implementar, debido a la gran cantidad de individuos que han de genotiparse para estimar los efectos de haplotipos y debido a los altos costes unitarios. Estos altos costes dificultan la aceptación de la tecnología por parte de la

industria. Para la implementación rentable de la selección hologenómica utilizando mapas de marcadores densos, no se dispone de métodos de manera rutinaria.

Varios autores han propuesto la identificación de subconjuntos informativos mínimos de SNP que permitirían la reconstrucción de los haplotipos inferidos mediante el genotipado de todos los otros SNP conocidos previamente en una población actual, es decir, independientemente del árbol genealógico, especialmente con referencia al genoma humano, es decir, "etiquetar los SNP" (por ejemplo, Avi-Itzhak et al., Proc. Pacific Symposium Biocomputing 8, 466-477, 2003; Hampe et al., Hum. Genet. 114, 36-43, 2003; Ke et al., Bioinformatics 19, 287- 288, 2003; Meng el al., Am. J. Hum. Genet. 73, 115-130, 2003; Sebastiani et al., Proc. Natl Acad. Sci USA 100, 9900-9905, 2003; Stram et al., Hum. Heredity 55, 179-190, 2003 Thompson el a., Hum. Heredity 56, 48-55, 2003; Wang et al., Hum. Mol. Genet. 12, 3145-3149, 2003; Weale et al., Am J. Hum. Genet. 73, 551-565, 2003; Halldórsson et al., Genome Res. 14, 1633-3640, 2006). Dichos métodos requieren la determinación de vecindarios de desequilibrio de ligamiento en el genoma para determinar de ese modo los SNP ("SNP etiquetados") que se pueden usar para inferir entre sí (porque están vinculados). Dichos vecindarios pueden ser bloques de haplotipos para los cuales se considera que dos SNP están correlacionados si se producen en el mismo bloque de haplotipos con poca evidencia de recombinación entre ellos (por ejemplo, Johnson et al, Nature Genetics 29, 233-237, 2001; Zhang et al. , Am. J. Hum. Genet. 73, 63-73, 2003), o una unión de posibles bloques de haplotipos que contienen SNP particulares (p. Ej., Halldórsson et al, Genome Res. 14, 1633-3640, 2006). Como alternativa, se considera que los vecindarios consisten solo en esos SNP a una distancia menor de 1 unidad de LD entre sí, según los mapas métricos de LD (por ejemplo, Maniatis et al, Proc. Natl Acad. Sci USA 99, 2228-2233, 2002). Sin embargo, hasta hace poco no había medios para definir la informatividad de los SNP marcados dentro de los vecindarios del desequilibrio de ligamiento, es decir, determinar cómo de bien cualquier SNP etiquetado caracterizaría la diversidad genética o la variación observada para el vecindario, porque los modelos utilizados asumían que las regiones genómicas tratadas eran pequeñas y no participaron muchos SNP. Zhang et al, Am. J. Hum. Genet.. 73, 63-73 (2003) propusieron un método para tratar grandes conjuntos de datos en donde los cromosomas se dividen en bloques de haplotipos y se selecciona un conjunto de SNP de etiquetado dentro de cada bloque imponiendo un coste por no etiquetar un SNP dado en términos de la pérdida de la diversidad de haplotipos. Halldórsson et al, Genome Res. 14, 1633-3640 (2006) sugirieron un marco algorítmico para definir la informatividad de grandes conjuntos de datos de SNP en el cromosoma humano 22, utilizando un método sin bloque para determinar vecindarios en desequilibrio de ligamiento, que requiere que los datos de fase de haplotipos estén disponibles. Básicamente, la medida de informatividad de Halldörsson et al, se calcula examinando los patrones de haplotipos para un conjunto de vecinos de un SNP diana, determinando aquellos pares de haplotipos que tienen diferentes alelos en el SNP diana, y después determinando la proporción de esos pares de haplotipos que no tienen el mismo conjunto de alelos en todos los SNP en el conjunto de vecinos. A pesar de las ventajas de etiquetar los SNP, tales métodos aún requieren un gran número de SNP para ser genotipados.

En consecuencia, sigue existiendo la necesidad de métodos informativos y rentables para realizar una selección artificial utilizando una estrategia basada en la genómica. La técnica anterior relevante en el campo es, por ejemplo, Hayes, B. et al., Livestock Production Science, 2003, 81, págs.197-211, que se relaciona con la rentabilidad del uso de marcadores de ADN ligados a locus de rasgos cuantitativos (QTL) en una empresa porcina en la selección asistida por marcadores (MAS) y concluye que los marcadores podrían encontrarse en desequilibrio de ligamiento con QTL, de manera que las asociaciones de alelos marcador-QTL persisten en toda la población, la rentabilidad del MAS podría aumentar.

Resumen de la invención

1. Definiciones

El término "alelo" se refiere a una cualquiera de las diferentes formas de un gen o secuencia de ADN en un solo locus, es decir, ubicación cromosómica que incluye una secuencia codificante, una secuencia no codificante o una secuencia reguladora.

La expresión "polimorfismo en la longitud de los fragmentos amplificados" o "AFLP" (siglas del inglés *Amplified fragment length polymorphism)* se refiere a una cualquiera de las diferentes longitudes de los fragmentos de ADN producidos por amplificación cebada al azar de fragmentos de ADN de restricción agrupados o aislados de ADN genómico o ADNc, en donde la longitud del fragmento varía entre individuos en una población.

Por "antepasado" se entiende un individuo que tiene una aportación genética en la población actual. El término "antepasado" es, por lo tanto, una función del árbol genealógico, cuya determinación no requiere el conocimiento previo de un rasgo particular o combinación de rasgos presentes en la población actual y sus progenitores. La información genotípica para un antepasado, a diferencia de un fundador, generalmente es incompleta como consecuencia del mantenimiento deficiente de los registros y la ausencia de material genético, por ejemplo, el semen, del antepasado para permitir el genotipado, por lo que deben deducirse los genotipos ausentes de la población ancestral para completar un análisis de genotipo. Los antepasados en un árbol genealógico pueden estar superpuestos, por ejemplo, un padre y uno de sus hijos, en virtud de aportar material genético común a la población actual, a pesar de cualquier gen aportado independientemente por uno u otro antepasado. Para determinar la

ascendencia, la relación promedio de un progenitor con la población actual se determina excluyendo el doble recuento de las aportaciones ancestrales superpuestas.

Por "selección artificial" se entenderá que es una selección bajo control humano, incluidos los sistemas, procesos, etapas o combinaciones de etapas de un programa de mejoramiento genético para producir ganancia genética, incluido el diseño colectivo y/o la implementación de dicho programa de mejoramiento y las etapas intermedias. por una o más personas. Por tanto, debe entenderse que la selección artificial requiere una determinación por parte del hombre, basada en un criterio de selección definido o criterios de selección definidos, de uno o más individuos en una población que serán padres y, en última instancia, antepasados, produciendo así una ganancia genética tal como se define en el presente documento. Esto es distinto de la mera observación de la genética de poblaciones, por ejemplo, para determinar un parámetro genético, tal como heredabilidad, diversidad, endogamia, etc. Los sistemas de selección artificial incluyen procesos de selección fenotípica y de selección genotípica. Las etapas de selección artificial incluyen, por ejemplo, la determinación de uno o más de los siguientes parámetros: criterios de selección y/o unos o más objetivos de reproducción; uno o más índices de selección; una o más dianas de selección; la intensidad de selección; una o ambas parejas sexuales para un solo apareamiento o para apareamientos múltiples, incluidas referencias y/o reemplazos; el número de apareamientos que uno o más individuos aportará a un programa de reproducción y el tiempo que un individuo permanecerá en una población de reproducción; intervalo de generación; valor reproductivo o ganancia genética. Las etapas de selección artificial también pueden incluir, por ejemplo, realizar una o más etapas de reproducción basadas en la determinación de uno o más parámetros anteriores y/o seleccionar la descendencia.

"Objetivo de reproducción" se refiere a un objetivo de un programa de selección artificial, por ejemplo, un germoplasma mejorado. El objetivo de reproducción se puede determinar mediante una combinación ponderada de rasgos que definen el valor de reproducción agregado de un animal.

"Valor de reproducción" significa el valor genético de un individuo como padre en un programa de reproducción y, más particularmente, el efecto de uno o más genes o marcadores genéticos del individuo, cuando se considera de forma aislada o combinada ("valor de reproducción agregado") sobre el rendimiento frente a un criterio de selección o criterios de selección.

A lo largo de esta memoria descriptiva y de las reivindicaciones que siguen, a menos que el contexto requiera lo contrario, se entenderá que la palabra "comprende", o variaciones tales como "comprende" o "que comprende", implican la inclusión de una etapa o elemento o número entero o grupo de etapas o elementos o números enteros, pero no la exclusión de cualquier otra etapa o elemento o número entero o grupo de elementos o números enteros.

Por "población actual" se entiende una población que es candidata para la selección. Normalmente, la población actual incluye individuos, por ejemplo, animales que están en o cerca de un punto final en un árbol genealógico.

Como se usa en el presente documento, la expresión "procede de" se tomará para indicar que un número entero específico puede obtenerse de una fuente particular, aunque no necesariamente de esa fuente.

La expresión "tamaño eficaz de la población" o "N_e "se refiere al número de individuos en una población que aporta gametos a la siguiente generación y preferentemente también a las generaciones futuras. El tamaño eficaz de la población generalmente se calcula como el número de individuos reproductores en una población idealizada que mostraría la misma cantidad de dispersión de frecuencias alélicas bajo deriva genética aleatoria o la misma cantidad de endogamia que una población considerada. Por ejemplo, en una población de apareamiento aleatorio que consiste en 1000 individuos de los cuales 500 son hombres y 500 son mujeres con generaciones separadas, la fracción esperada de los genes transportados por cualquier generación futura aportada por un animal cualquiera en la generación actual es 0,1 % y el tamaño eficaz de la población es el mismo que el tamaño absoluto o real de la población (N), es decir, 1000. Sin embargo, como la mayoría de las poblaciones son endogámicas, los individuos no seleccionan parejas al azar, las generaciones pueden superponerse, y se reproducen menos machos que las hembras, el tamaño eficaz de la población generalmente tiene un valor menor que el tamaño absoluto o real de la población.

El "valor de reproducción estimado" o "EBV, *estímate breeding value*", se refiere a un valor de reproducción previsto de la descendencia de un suceso de apareamiento, determinado multiplicando la ploidía del organismo en cuestión por la diferencia de descendencia, es decir, la diferencia entre los rendimientos promedio de la descendencia de un individuo y los rendimientos promedio de toda la descendencia en una población asumiendo apareamiento aleatorio. Para un organismo diploide, la diferencia de descendencia se duplica, porque el valor de reproducción es una medida de todos los genes para el organismo, mientras que la diferencia de descendencia se basa en la aportación de un solo genoma haploide de un progenitor. Las diferencias de descendencia se basan en el rendimiento promedio previsto de la descendencia porque cada padre contribuye con el mismo número de genes a cada descendencia en la población.

Por "fundador" en un árbol genealógico se entiende que es un individuo cuyos padres se desconocen. En lugar de fundadores en el método del presente documento descrito, pueden utilizarse antepasados cuando el árbol

genealógico conocido está incompleto y/o los genotipos de los antepasados no se conocen o no es posible saber su procedencia. La presente invención tiene utilidad cuando se utilizan genotipos de una población fundadora para inferir los genotipos de la población actual, sin embargo, esto es menos preferido que utilizar genotipos de los antepasados porque se espera que haya menos antepasados clave que fundadores. Como la invención tiene un alto nivel de precisión cuando se utilizan genotipos de fundadores, la población fundadora también puede servir como un modelo adecuado para una población de antepasados.

"Intervalo de generación" significa la cantidad de tiempo necesaria para reemplazar una generación con la siguiente y, en una población cerrada que está sujeta a selección artificial, la edad promedio de los padres cuando nace la descendencia seleccionada.

Tal como se usa en este documento, la expresión "ganancia genética" debe entenderse como el cambio promedio en un rasgo hereditario o combinación de rasgos hereditarios de una generación a la siguiente generación, incluida una ganancia genética prevista y/o una ganancia genética real. Más particularmente, el cambio promedio es en la dirección de una o más dianas de selección, o al menos impedirá una ganancia genética negativa significativa, es decir, un efecto no deseado para los criterios de selección. La ganancia genética puede surgir de la selección artificial.

"Selección genotípica" significa una selección artificial basada en la presencia y/o ausencia de uno o más genes o marcadores genéticos de un individuo asociado a un gen particular, combinación de genes, rasgo de un solo gen, rasgo cuantitativo o combinación de rasgos. La selección genotípica incluye una amplia gama de métodos de selección asistida por marcadores que comprenden el uso de marcadores genéticos, por ejemplo, alelos, haplotipos, haplogrupos, loci, loci de rasgos cuantitativos o polimorfismos de ADN [polimorfismos de longitud de fragmentos de restricción (RFLP), polimorfismos de longitud de fragmentos amplificados (AFLP), polimorfismos mononucleotídicos (SNP), inserciones y deleciones (indel), repeticiones cortas en tándem (STR, *Short Tadem Repeat*), microsatélites y minisatélites], en donde el marcador o los marcadores son determinantes del valor de reproducción estimado del individuo.

Un "haplogrupo" es un grupo de haplotipos similares, por ejemplo, haplogrupos del cromosoma Y humano definidos basándose en sucesos de mutación únicos en repeticiones cortas en tándem en el cromosoma Y (Y-STR).

El término "haplotipo" se refiere a una combinación de alelos, loci o polimorfismos de ADN que se ligan para cosegregarse en una proporción significativa de gametos durante la meiosis. Los alelos de un haplotipo pueden estar en desequilibrio de ligamiento (LD).

El término "indel" se refiere a una cualquiera de las diferentes inserciones o deleciones de ADN en un alelo o locus en particular que están presentes en diferentes individuos en una población, por ejemplo, polimorfismos AIu del cromosoma Y (YAP).

Como se usa en este documento, el término "inferir" o términos equivalentes como "infiriendo" o "inferido", por ejemplo, en el contexto de un genotipo, haplotipo, QTL, marcador, etc., se tomará para dar a entender que un genotipo se deduce de la información disponible, y más particularmente que la información ausente, tal como un genotipo ausente con respecto a uno cualquiera o más marcadores, por ejemplo, en una ubicación específica en el genoma de un individuo, se deduce. Por ejemplo, un genotipo ausente para un antepasado (y/o fundador) se "infiere" utilizando los datos del genotipo de un individuo en la población actual relacionada por el árbol genealógico con el antepasado (y/o fundador), al realizar la presente invención como se describe de acuerdo con a una o más realizaciones del presente documento. Como alternativa, o además, se "infiere" un genotipo ausente para un individuo de una población actual utilizando datos genotípicos de un antepasado (y/o fundador) relacionados por el árbol genealógico con ese individuo, por ejemplo, empleando uno o más medios estadísticos, tales como, entre otros, el modelado de MCMV. Mediante dichas inferencias, los datos genotípicos tanto de los antepasados (o fundadores) como los de la población actual, se hacen más completos de lo que sería el caso.

El término "desequilibrio de ligamiento " o "LD" se refiere a los alelos o locus o polimorfismos de ADN que se asocian a una frecuencia mayor que la esperada para los alelos o marcadores independientes, de manera que aparecen como un haplotipo. Por ejemplo, cuando las variantes de dos locus genéticos se encuentran en fuerte desequilibrio de ligamiento, la variante en un locus es predictiva de la variante en el otro locus en un cromosoma individual.

En el presente contexto, el término "apareamiento" o término similar, tal como "emparejamiento" se interpretará sin referencia a un Reino o Filo para que signifique cualquier reproducción sexual en la que un genoma haploide se transfiere de un individuo de una población a otro individuo de una población , incluido el apareamiento de un animal o célula (p. ej., célula de levadura) con otro(a) por medios naturales o asistidos, p. ej., inseminación artificial (IA); y la autopolinización de una planta o polinización cruzada entre plantas.

"Variación de muestreo mendeliano" significa la variación en la desviación del valor de reproducción de un individuo de los valores de reproducción medios de sus progenitores.

El término "minisatélite" se refiere a una repetición en tándem de número variable (VNTR, *variable number tandem repeat*) que comprende más de aproximadamente 5 repeticiones y de 6 a aproximadamente 60 pares de bases por unidad de repetición, en donde el número de unidades de repetición varía entre los individuos en una población. Al igual que con los microsatélites, pueden producirse cambios y el número de repeticiones puede aumentar o disminuir.

"Selección fenotípica" significa una selección artificial basada en uno, y posiblemente más, fenotipos de un individuo. La selección fenotípica generalmente comprende pruebas de descendencia en las que el valor de reproducción estimado de un individuo se determina realizando múltiples apareamientos del individuo y determinando el rendimiento de la descendencia.

En el presente contexto, el término "población" significa un grupo de individuos que posiblemente se reproducen entre sí de tal manera que contribuyen genéticamente a la próxima generación, incluidos, entre otros, aquellos individuos en un programa de reproducción. El grupo puede ser de cualquier tamaño, por ejemplo, una especie, raza, línea, variedad de cultivo, rebaño o manada, etc.

La expresión "rasgo cuantitativo" se refiere a un rasgo que se determina mediante la expresión de más de un gen.

La expresión "locus de rasgo cuantitativo" o "QTL" se refiere a una región de ADN que está asociada a un rasgo cuantitativo particular, en donde la variación en el QTL está asociada a la variación en el rasgo cuantitativo según lo determina el mapeo genético o la selección asistida por marcadores.

"Referencia" significa un progenitor o antepasado (y/o fundador) que proporciona una aportación genética a varios grupos de individuos, lo que permite comparar los rendimientos de la descendencia dentro y entre los grupos en relación con el rendimiento de la descendencia de otros progenitores o antepasados (y/o fundadores). Las referencias permiten seleccionar los mejores antepasados (y/o fundadores) y utilizarlos en la selección artificial.

"Reemplazo" significa un individuo que se convertirá en progenitor por primera vez en un programa de selección artificial.

La expresión "polimorfismo de longitud de fragmento de restricción" o "RFLP" se refiere a una cualquiera de las diferentes longitudes de fragmentos de ADN producidas por digestión con enzimas de restricción de ADN genómico o ADNc con una o más enzimas endonucleasas, en donde la longitud del fragmento varía entre individuos en una población.

Tal como se usa en el presente documento, el término "selección" se tomará para referirse a uno o más sistemas, procesos, etapas o combinaciones de etapas que determinan uno o más individuos en una población que han de contribuir a la próxima generación, incluida la selección natural y la selección artificial.

"Criterio de selección" se refiere a un fenotipo o genotipo que establece una decisión de selección, incluida la presencia o ausencia de uno o más genes, o de uno o más marcadores genéticos asociados a un gen particular, combinación de genes, rasgo o combinación de rasgos .

"Índice de selección" significa una clasificación de un criterio de selección o criterios de selección según una ponderación o calificación, utilizada para estimar el valor de reproducción.

"Intensidad de selección" se refiere a la medida en que un criador se adhiere a una decisión sobre la selección de un individuo o grupo de individuos en particular para el apareamiento. Estadísticamente, la intensidad de selección se determina como la diferencia entre el criterio de selección promedio de los individuos seleccionados para contribuir a la próxima generación y el criterio de selección promedio de todos los posibles progenitores, expresada en unidades de desviación típica.

"Diana de selección" se refiere a un valor de reproducción óptimo deseado.

La expresión "repetición corta en tándem" o "STR" se refiere a una repetición en tándem de número variable (VNTR) que comprende de 2 a aproximadamente 5 o 6 pares de bases por unidad de repetición, en donde el número de unidades de repetición varía entre los individuos en una población. Los microsatélites son un ejemplo de una STR que generalmente es muy polimórfica y que se distribuye aleatoriamente en el genoma y que puede contener variabilidad en la secuencia y/o para el cual el número de unidades repetidas puede aumentar o disminuir.

La expresión "rasgo de un solo gen" se refiere a un rasgo que se determina mediante la expresión de un gen.

La expresión "polimorfismo mononucleotídico" o "SNP" se refiere a uno cualquiera de los diferentes mononucleótidos en un alelo o locus en particular que varía entre los individuos en una población. Muchos SNP son bialélicos.

2. Argumentario

La selección utilizando datos de marcadores, por ejemplo, procedentes de marcadores de ADN, requiere que los genotipos de los candidatos a la selección se conozcan en aquellos loci que tienen un efecto sobre los rasgos dentro del objetivo de reproducción. Es probable que este sea un gran número de marcadores y la lista de dichos marcadores se ampliará a medida que la investigación proporcione datos de ligamiento adicionales.

En el trabajo que condujo a la presente invención, los inventores razonaron que los costes para seleccionar individuos de una población podían reducirse si los candidatos a la selección podrían genotiparse para un número relativamente pequeño de marcadores y preferentemente un conjunto constante de marcadores. Los inventores razonaron que dicha reducción de costes se realizaría genotipando a antepasados clave de los candidatos de selección para marcadores útiles, y preferentemente para todos los marcadores útiles, y los candidatos de selección se genotiparían solo para un subconjunto de esos marcadores, y esto se puede lograr rastreando uno o más segmentos cromosómicos que llevan los marcadores útiles en un candidato de selección con el correspondiente segmento o segmentos cromosómicos en un antepasado clave del cual procede el candidato de selección.

Aunque es deseable que los antepasados clave se hayan genotipado para todos los marcadores útiles, esto no siempre es posible. Por ejemplo, cuando no es posible disponer de ninguna fuente de ADN de un antepasado (y/o fundador). En dichas circunstancias, los inventores han razonado que el/los genotipo(s) del antepasado (y/o fundador) clave para los marcadores puede(n) inferirse a partir del de parientes adecuados que hayan sido genotipado para esos marcadores, por ejemplo, utilizando una estrategia algorítmica que complete los valores ausentes, tales como el modelado de Monte Carlo de Cadenas de Markov (MCMV).

Aunque para esta finalidad es deseable poder disponer del(los) árbol(es) genealógico(s) del(los) candidato(s) de selección, incluyendo las relaciones con uno o más antepasados [y/o fundador(es)] clave, esto no siempre es posible. A menudo, dichos árboles genealógicos tampoco están disponibles, porque los datos del árbol genealógico están incompletos. En dichas circunstancias, los inventores razonaron que la(s) relación(es) del(los) candidato(s) de selección con la del antepasado(s) [y/o fundadores] clave, pueden inferirse utilizando marcadores genéticos que se hayan genotipado tanto en los candidatos de selección como en los antepasados clave. Como alternativa, o además, en el análisis con los antepasados clave, pueden incluirse animales fundadores dentro del árbol genealógico conocido.

La presente invención se basa en una comprensión por parte del inventor de que, para especies que tienen un tamaño de población eficaz pequeño, el número de antepasados (y con frecuencia el número de fundadores) clave, es pequeño en relación con el número de candidatos seleccionados. Por lo tanto, hay un ahorro en el coste si los candidatos de selección se genotipan solo para un subconjunto de los marcadores cuyos genotipos se conocen o se pueden inferir sobre los antepasados (y/o fundadores) clave. Es posible inferir los genotipos ausentes de los candidatos de selección porque la relación entre los candidatos de selección y los antepasados (y/o fundadores) clave, se conoce a partir del árbol genealógico o se infiere de otros marcadores genéticos. Métodos para inferir genotipos ausentes que no aprovechan la relación entre los candidatos seleccionados y los antepasados (y/o fundadores) clave, serían mucho menos eficaces y, por lo tanto, el ahorro de costes sería mucho menor.

Por otra parte, el inventor ha llegado a la conclusión de que basándose en la informatividad de marcadores etiquetados sobre un árbol genealógico, pueden generarse ahorros de costes adicionales para el genotipado de individuos en una población actual. Más particularmente, el inventor ha llegado a la conclusión de que para una especie que tiene un tamaño de población eficaz pequeño, la diversidad de la población se explica sustancialmente por la suma de aquellos antepasados (y/o fundadores) clave que contribuyen a la población a largo plazo, y que la diversidad se hereda como vecindarios de segmentos cromosómicos que comprenden marcadores antepasados, que pueden estar en desequilibrio de ligamiento (LD). Procediendo sobre esta base, el inventor ha llegado a la conclusión de que la cantidad de marcadores informativos a genotipar en un individuo en una población actual se reduce al inferir genotipos ausentes de un segmento cromosómico de un antepasado (y/o fundador) que contribuyen a que el segmento cromosómico sea el mismo que el del individuo de la población actual a genotipar un marcador informativo dentro del segmento cromosómico. Esto difiere de los métodos de SNP de etiquetado conocidos que son independientes del árbol genealógico y que generalmente requieren el genotipado de un gran número de marcadores porque se basan en bloques de haplotipos o en una unión de bloques de haplotipos, o requieren mapas métricos detallados de LD.

En consecuencia, la presente invención se refiere a un método de selección artificial que comprende:

(i) genotipar a un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en una pluralidad de segmentos cromosómicos, en donde la población es una población de plantas o de animales con una población de 1000 individuos o de menos individuos,
(ii) rastrear los linajes de la pluralidad de segmentos cromosómicos hasta uno o más antepasados y/o fundadores de los que proceden,
(iii) inferir que los genotipos de cada segmento cromosómico son los mismos que para uno o más antepasados y/o fundadores de los que proceden los segmentos cromosómicos,

(iv) estimar el valor de reproducción del individuo basándose en los genotipos inferidos, en donde cada antepasado o fundador es un antepasado o fundador que proporciona a la población actual, al menos el 0,1 % de la variación genética total y donde se conocen los genotipos de uno o más antepasados y/o fundadores de uno o más marcadores informativos,
y
(v) seleccionar un individuo que tenga un valor de reproducción estimado elevado, y en el que la población o el individuo no es un ser humano.

Además, la invención también se refiere al uso del método de acuerdo con lo anterior para seleccionar un individuo o material reproductivo o regenerativo del individuo para uso en reproducción, inseminación artificial, fertilización *in vitro*, implantación de embriones, o transgénicos, en donde el individuo no es un ser humano.

La divulgación también se refiere a un procedimiento para producir ganancia genética en una población que comprende realizar el método de acuerdo con lo anterior y seleccionar un individuo que tenga un valor de reproducción estimado elevado de una población en el que el individuo no es un ser humano.

Finalmente, la invención se refiere a un método de selección artificial en ganado bovino que comprende:

(i) identificar el conjunto mínimo de antepasados clave que representan la mayoría de los segmentos cromosómicos en una población actual que tiene un tamaño de población de 1000 individuos o de menos individuos;
(ii) genotipar a los antepasados clave para un conjunto de marcadores densos;
(iii) genotipar a uno o más individuos de una población actual para determinar marcadores suficientes para permitir así que los segmentos cromosómicos coincidan con los segmentos llevados por antepasados clave;
(iv) rastrear los segmentos cromosómicos de uno o más individuos de la población actual hasta un antepasado clave;
(v) inferir los genotipos de los marcadores dentro de uno o más segmentos cromosómicos de uno o más individuos en la población actual para que sean los mismos que los del antepasado clave;
(vi) utilizar el genotipo inferido de uno o más individuos en la población actual para estimar el valor de reproducción de dicho uno o más individuos; y
(vii) seleccionar un individuo que tenga un valor de reproducción estimado elevado

y en el que la población o el individuo no es un ser humano.

3. Realizaciones específicas

La presente divulgación proporciona un método de selección artificial para un solo gen o locus, que incluye un locus de un solo gen o un QTL, comprendiendo dicho método el genotipado de un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en un segmento cromosómico que comprende un gen o locus de interés, inferir que el genotipo en el locus o QTL es el mismo que el de un antepasado (y/o fundador) del cual procede el segmento cromosómico, y estimar el valor de reproducción del individuo basado en el genotipo inferido, en el que el antepasado (y/o fundador) es un antepasado (y/o fundador) que proporciona a la población actual una aportación genética significativa prolongada, y en el que se conoce el genotipo del antepasado (y/o fundador) para el uno o más marcadores informativos y para el locus o QTL.

Se entenderá que este método es más generalmente aplicable para obtener el genotipo de un individuo para cualquier número de loci o QTL, en cualquier número de ubicaciones de cromosomas. De acuerdo con este ejemplo, la presente invención proporciona un método de selección artificial para uno o más loci o QTL, comprendiendo dicho método el genotipado de un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en uno o más segmentos cromosómicos conteniendo cada uno de ellos uno o más loci o QTL de interés, inferir genotipos en el uno o más locus o QTL para ser el mismo que para un antepasado (y/o fundador) a partir del cual procede un segmento cromosómico, y estimar el valor de reproducción del individuo basándose en los genotipos inferidos, en donde uno o más antepasados (y/o fundadores) es un antepasado (y/o fundador) que proporciona a la población actual una aportación genética significativa prolongada, y en el que se conocen los genotipos del uno o más antepasados (y/o fundadores) para el uno o más marcadores informativos y para el locus o QTL.

En otro ejemplo, el método puede utilizarse para obtener el genotipo de un individuo, por ejemplo, el genotipo hologenómico (de todo el genoma). De acuerdo con este ejemplo, la presente divulgación proporciona un método de selección artificial que comprende el genotipado de un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en una pluralidad de segmentos cromosómicos, infiriendo genotipos de cada segmento cromosómico para que sea igual en el individuo que en uno o más antecesores (y/o fundadores) de los que proceden los segmentos cromosómicos, y estimando el valor de reproducción del individuo basándose en los genotipos inferidos, en donde cada antepasado (y/ o fundador) es un antepasado (y/o fundador) que proporciona una aportación genética significativa prolongada a la población actual y en donde los genotipos de uno o más antecesores (y/o fundadores) para uno o más marcadores informativos son

sustancialmente conocidos. Preferentemente, se conocen el genotipo (o genotipos) de uno o más antepasados (y/o fundadores) para cada marcador informativo. Para obtener el genotipo hologenómico, se prefiere que los segmentos cromosómicos abarquen el genoma.

Para el genotipado aplicado a un solo locus o QTL, múltiples locus o QTL, o contextos de exploración hologenómica, se prefiere que el genotipado de un individuo comprenda la detección de uno o más marcadores informativos en un sistema de alto rendimiento que comprende un soporte sólido que consiste esencialmente en, o que tiene, ácidos nucleicos de diferente secuencia, unidos directa o indirectamente a los mismos, en donde cada ácido nucleico de diferente secuencia comprende un marcador genético polimórfico procedente de un antepasado (y/o fundador) que es representativo de la población actual. Preferentemente, el sistema de alto rendimiento comprende marcadores suficientes para ser representativos del genoma de la población actual, es decir, abarcan todo el genoma y comprenden marcadores polimórficos suficientes para ser útiles para explorar todo el genoma. Los marcadores pueden agruparse en grupos de ligamiento, opcionalmente de acuerdo con un segmento cromosómico con el que se encuentra en desequilibrio de ligamiento. La información del marcador contenida en el sistema de alto rendimiento puede obtenerse mediante una etapa intermedia en un método de la presente invención.

Como se usa en el presente documento, el término "genotipar a un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos" simplemente significa determinar la presencia o la ausencia del marcador (o marcadores). El experto en la materia sabrá que si un marcador se selecciona a favor o en contra dependerá de la asociación del marcador para un genotipo deseado. El experto en la materia también será consciente de que, en vista del objetivo de seleccionar la reserva de reproducción o el germoplasma para mejorar la ganancia en las generaciones futuras, la generalidad de la invención no pretende limitarse a determinar la presencia o ausencia de un genotipo, haplotipo o haplogrupo específico, como para un locus particular o QTL.

Debe entenderse que la aplicación de la presente invención no está limitada a ninguna especie particular, sino que está determinada por el tamaño eficaz de la población de la especie. Sin embargo, la invención no se refiere a seres humanos. Po consiguiente, la presente invención es aplicable a la selección artificial en plantas y animales que tienen tamaños de población eficaces y pequeños. También debe entenderse que, debido a que la selección de antepasados (y/o fundadores) es una función del árbol genealógico, la presente invención también es aplicable a la selección de especies predominantemente exogámicas y/o predominantemente endogámicas. Los ejemplos de poblaciones a las que se aplica fácilmente la presente invención incluyen ganado (por ejemplo, ganado vacuno y lechero como Holstein, Frisona, Holstein-Frisona, Braunvieh, Brown Swiss, Jersey, Danish Red, Aberdeen Angus), ovejas (por ejemplo, cruces de Meatlinc, Dorset x Rambouillet x. Finnsheep), cerdos (por ejemplo, cruce de Large White x Landrace, Large White, Duroc, Yorkshire, Landrace), aves de corral (por ejemplo, ponedoras), pescado (por ejemplo, salmón del Atlántico), crustáceos, raigrás, etc.

De acuerdo con los ejemplos anteriores de la presente divulgación, la población actual será una población de individuos que tiene un tamaño eficaz de la población pequeño. Esto significa que el tamaño eficaz de la población debe ser menor que el número de individuos en la población actual requerida que necesitaría genotiparse para estimar todos los efectos de haplotipos, y preferentemente menos de la mitad o menos de un tercio o menos de una cuarta parte o aproximadamente una décima parte de la cantidad de individuos en la población actual que necesitaría genotiparse para estimar los efectos de los haplotipos. En términos de los números reales de antepasados (y/o fundadores) que necesitarían genotiparse al realizar la presente divulgación, esto variará dependiendo de la población en cuestión y del nivel de selección artificial que se haya aplicado a la población en generaciones anteriores. Por ejemplo, preferentemente esto significa menos de aproximadamente 1000 individuos, más preferentemente menos de aproximadamente 350 individuos, aún más preferentemente menos de aproximadamente 250 individuos, aún más preferentemente menos de aproximadamente 200 individuos y aún más preferentemente menos de aproximadamente 150 o menos de aproximadamente 100 individuos. Como alternativa, el tamaño eficaz de la población está en el intervalo de aproximadamente 30-350 o de aproximadamente 30-200 o de aproximadamente 30-100 individuos. Para poblaciones más grandes que estas estimaciones, el coste del beneficio de realizar una selección genómica basada en el linaje ancestral de los segmentos cromosómicos disminuye.

Un antepasado (y/o fundador) que proporcione a la población actual una aportación genética significativa prolongada proporcionará preferentemente al menos aproximadamente el 0,1 % de la variación total a la población actual y, más comúnmente, al menos aproximadamente el 0,5 % o el 1 % de la variación total. Los antepasados (y/o los fundadores) particularmente importantes o "clave", generalmente proporcionan al menos aproximadamente del 2 al 10 % de la variación total a la población actual, por ejemplo, 2 % o 3 % o 4 % o 5 % o 6 % o 7 % u 8 % o 9 % o 10%, sin embargo, no se excluyen aportaciones de antepasados más grandes.

Los marcadores pueden ser cualquier marcador genético, por ejemplo, uno o más alelos, haplotipos, haplogrupos, locus, locus de rasgos cuantitativos o polimorfismos de ADN [polimorfismos de longitud de fragmentos de restricción (RFLP), polimorfismos de longitud de fragmentos amplificados (AFLP), polimorfismos mononucleotídicos (SNP), indeles, repeticiones cortas en tándem (STR), microsatélites y minisatélites]. Convenientemente, los marcadores son

SNP o STR, tales como microsatélites, y más preferentemente SNP. Preferentemente, los marcadores, dentro de cada segmento cromosómico, están en desequilibrio de ligamiento.

La presente invención incluye claramente la realización de etapas adicionales donde no se conocen datos informativos sobre los antepasados (y/o fundadores), incluida la identificación y/o caracterización de los antepasados (y/o fundadores) y/o el establecimiento del linaje de uno o más segmentos cromosómicos. Por ejemplo, los antepasados (y/o fundadores) pueden caracterizarse por obtener y/o proporcionar sus genotipos, por ejemplo, para marcadores útiles, un gran número de marcadores útiles o la mayoría de los marcadores que utilizan procedimientos estándar para hacerlo, en donde dichos genotipos también pueden inferirse de datos de sus familiares, por ejemplo, utilizando medios estadísticos tales como el modelado de MCMV para predecir valores ausentes. En un ejemplo, los antepasados (y/o fundadores) se caracterizan por proporcionar y/u obtener genotipos conocidos y/o inferir sus genotipos.

Por consiguiente, en un ejemplo adicional, el método desvelado comprende rastrear el linaje del uno o más segmentos cromosómicos hasta uno o más antecesores (y/o fundadores) de los cuales proceden. De acuerdo con este ejemplo, la presente divulgación proporciona un método de selección artificial para un solo gen o locus, que incluye un locus de un solo gen o un QTL, comprendiendo dicho método el genotipado de un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en un segmento cromosómico que comprende un gen o locus de interés, rastrear el linaje del segmento cromosómico en el individuo hasta un antepasado (y/o fundador) del cual procede, inferir un genotipo en el locus o QTL para que sea el mismo que en un antepasado (y/o fundador) del cual procede el segmento cromosómico, y estimar el valor de reproducción del individuo basándose en el genotipo inferido, en donde el antepasado (y/o fundador) es un antepasado (y/o fundador) que proporciona una aportación genética significativa prolongada a la población actual y en donde se conoce el genotipo del antepasado (y/o fundador) para el uno o más marcadores informativos y para el locus o QTL. Para múltiples locus o QTL en cualquier número de ubicaciones de cromosomas, la invención proporciona un método de selección artificial para uno o más locus o QTL, comprendiendo dicho método el genotipado de un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en uno o más segmentos cromosómicos, conteniendo cada uno de ellos uno o más locus o QTL de interés, rastreando el linaje del uno o más segmentos cromosómicos hasta uno o más antepasados (y/o fundadores) de los que proceden, infiriendo genotipos en el uno o más locus o QTL para que sean iguales que los de un antepasado (y/o fundador) a partir del cual procede un segmento cromosómico, y estimar el valor de reproducción del individuo basándose en los genotipos inferidos, donde uno o más antepasados (y/o fundadores) es un antepasado (y/o fundador) que proporciona una aportación genética significativa prolongada a la población actual, y en donde se conocen los genotipos del uno o más antepasados (y/o fundador) para el uno o más marcadores informativos y para los locus o QTL. Para la selección hologenómica, la presente divulgación proporciona un método de selección artificial que comprende el genotipado de un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en una pluralidad de segmentos cromosómicos, rastreando los linajes de la pluralidad de segmentos cromosómicos hasta uno o más antepasados (y/o fundadores) de los que proceden, infiriendo genotipos de cada segmento cromosómico en el individuo para que sea el mismo que en uno o más antepasados (y/o fundadores) de los que proceden los segmentos cromosómicos y estimar el valor de reproducción del individuo basándose en los genotipos inferidos, donde cada antepasado (y/o fundador) es un antepasado (y/o fundador) que proporciona una aportación genética significativa prolongada a la población actual y en donde los genotipos de uno o más antepasados (y/o fundadores) para uno o más marcadores informativos es sustancialmente conocido. Preferentemente, el genotipo (o genotipos) de uno o más antepasados (y/o fundadores) para cada marcador informativo es conocido. Para obtener el genotipo hologenómico, se prefiere que los segmentos cromosómicos abarquen el genoma.

En otro ejemplo más, el método de la divulgación comprende la caracterización de los antepasados (y/o fundadores), por ejemplo, genotipando uno o más antepasados (y/o fundadores) para determinar marcadores conocidos. De acuerdo con este ejemplo, la presente divulgación proporciona un método de selección artificial para un solo gen o locus, incluyendo un locus de un solo gen o un QTL, comprendiendo dicho método el genotipado de un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en un segmento cromosómico que comprende un gen o locus de interés, rastreando el linaje del segmento cromosómico en el individuo hasta un antepasado (y/o fundador) del cual procede, el genotipado del antepasado (y/o fundador) para determinar marcadores conocidos, infiriendo un genotipo en el locus o QTL para que sea el mismo que el de un antepasado (y/o fundador) del cual procede el segmento cromosómico, y estimar el valor de reproducción del individuo basándose en el genotipo inferido, en donde el antepasado (y/o fundador) es un antepasado (y/o fundador) que proporciona una aportación genética significativa a largo plazo a la población actual, y en donde se conoce el genotipo del antepasado (y/o fundador) para el locus o QTL. Para múltiples locus o QTL en cualquier número de ubicaciones de cromosomas, la invención proporciona un método de selección artificial para uno o más locus o QTL, comprendiendo dicho método el genotipado de un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en uno o más segmentos cromosómicos, conteniendo cada uno de ellos uno o más locus o QTL de interés, rastreando el linaje del uno o más segmentos cromosómicos hasta uno o más antepasados (y/o fundadores) de los que proceden, genotipando el uno o más antepasados (y/o fundadores) para detectar marcadores conocidos, infiriendo genotipos en uno o más locus o QTL para que sea el mismo que el del antepasado (o fundador) del cual procede un segmento cromosómico, y

estimar el valor de reproducción del individuo basándose en los genotipos inferidos, en donde uno o más antepasados (y/o fundadores) es un antepasado es un antepasado (y/o fundador) que proporciona una aportación genética significativa largo plazo a la población actual, y en donde se conocen los genotipos del uno o más antepasados (y/o fundadores) para los locus o QTL. Para la selección hologenómica, la presente divulgación proporciona un método de selección artificial que comprende el genotipado de un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en una pluralidad de segmentos cromosómicos, rastreando los linajes de la pluralidad de segmentos cromosómicos hasta uno o más antepasados (y/o fundadores) de los que proceden, genotipando uno o más antepasados (y/o fundadores) para determinar marcadores conocidos, inferir genotipos de cada segmento cromosómico en el individuo para que sean los mismos que en uno o más antepasados (y/o fundadores) de los que proceden los segmentos cromosómicos, y estimar el valor reproductor del individuo basándose en los genotipos inferidos, en donde cada antepasado (y/o fundador) es un antepasado (y/o fundador) que proporciona una aportación genética significativa, a largo plazo, a la población actual. Para obtener el genotipo hologenómico, se prefiere que los segmentos cromosómicos abarquen el genoma.

En otro ejemplo más, el método de divulgación comprende la identificación de los antepasados (y/o fundadores), por ejemplo, determinando un conjunto mínimo de antepasados (y/o fundadores) representativos de la población actual. De acuerdo con este ejemplo, la presente divulgación proporciona un método de selección artificial para un solo gen o locus, que incluye un locus de un solo gen o un QTL, comprendiendo dicho método el genotipado de un individuo en una población actual para detectar la presencia o la ausencia de uno o más marcadores informativos en un segmento cromosómico que comprende un gen o locus de interés, determinando un conjunto mínimo de antepasados (y/o fundadores) representativos de la población actual, rastreando el linaje del segmento cromosómico en el individuo hasta un antepasado (y/o fundador) del cual procede, infiriendo un genotipo en el locus o QTL para que sea e mismo que en un antepasado (y/o fundador) del cual procede el segmento cromosómico, y estimar el valor de reproducción del individuo basado en el genotipo inferido, en donde se conoce el genotipo del antepasado (y/o fundador) para el uno o más marcadores informativos y para el locus o QTL. Para múltiples locus o QTL en cualquier número de ubicaciones de cromosomas, la divulgación proporciona un método de selección artificial para uno o más locus o QTL comprendiendo dicho método el genotipado de un individuo en una población actual para detectar la presencia o la ausencia de uno o más marcadores informativos en uno o más segmentos cromosómicos, conteniendo cada uno de ellos uno o más locus o QTL de interés, determinar un conjunto mínimo de antepasados (y/o fundadores) representativos de la población actual, rastreando el linaje de uno o más segmentos cromosómicos hasta uno o más antepasados (y/o fundadores) de los que proceden, inferir genotipos en el uno o más locus o QTL para que sean iguales en un antepasado (y/o fundador) a partir del cual procede un segmento cromosómico y estimar el valor de reproducción del individuo basándose en los genotipos inferidos, en donde se conocen los genotipos del uno o más antepasados (y/o fundadores) para el uno o más marcadores informativos y para los locus o QTL. Para la selección hologenómica, la presente divulgación proporciona un método de selección artificial que comprende genotipar a un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en una pluralidad de segmentos cromosómicos, determinando un conjunto mínimo de antepasados (y/o fundadores) representativos de la población actual, rastreando los linajes de la pluralidad de segmentos cromosómicos hasta uno o más antepasados (y/o fundadores) de los que proceden, infiriendo genotipos de cada segmento cromosómico en el individuo para ser igual que en uno o más antepasados (y/o fundadores) de los que proceden los segmentos cromosómicos, y estimando el valor reproductor del individuo basándose en los genotipos inferidos, en donde se conocen los genotipos del uno o más antepasados (y/o fundadores) para el uno o más marcadores informativos. Para obtener el genotipo hologenómico, se prefiere que los segmentos cromosómicos abarquen el genoma.

En otro ejemplo más, los segmentos cromosómicos individuales en los candidatos de selección (es decir, miembros de la población actual) se remontan hasta los antepasados (y/o fundadores) clave mediante un proceso que comprende el rastreo de segmentos cromosómicos en candidatos de selección a uno o más antepasados (y/o fundadores) inmediatos, utilizando un pequeño número de marcadores y rastreando los segmentos cromosómicos en los antepasados (y/o fundadores) inmediatos a los segmentos cromosómicos correspondientes en uno o más antepasados (y/o fundadores) clave. Preferentemente, los segmentos cromosómicos en los antepasados (y/o fundadores) inmediatos se remontan a los segmentos cromosómicos en los antepasados (y/o fundadores) clave utilizando un mayor número de marcadores. El rastreo de segmentos cromosómicos a antepasados (y/o fundadores) inmediatos puede minimizar costes. Por ejemplo, los antepasados (y/o fundadores) inmediatos pueden ser todos los animales macho utilizados en la manada o el rebaño en las últimas generaciones. Dado que en la mayoría de las especies se utilizan pocos machos, el número de antepasados (y/o fundadores) inmediato es pequeño en comparación con el número de candidatos de selección, de tal manera que el coste del genotipado de estos marcadores, que es suficiente para rastrear segmentos cromosómicos hasta antepasados (y/o fundadores) clave, también se reduce o minimiza.

En otro ejemplo más, los árboles genealógicos de los animales no se conocen, pero se infieren de los marcadores de ADN que se utilizan para rastrear segmentos cromosómicos. Por ejemplo, la raza del animal puede ser desconocida, pero puede deducirse de los marcadores de ADN.

En otro ejemplo más, la secuencia genómica de los antepasados (y/o fundadores) clave se conoce y es preferentemente completa, lo que permite inferir la secuencia del genoma casi completa o completa de todos los

animales actuales, tal como, por ejemplo, rastreando sus segmentos cromosómicos hasta los antepasados (y/o fundadores) clave. Dichos datos de la secuencia del genoma son útiles para la selección.

La divulgación permite además un método para determinar un conjunto de antepasados (y/o fundadores) que es representativo de una población actual que tiene un tamaño de población eficaz pequeño, comprendiendo dicho método la determinación de aportaciones a largo plazo de antepasados (y/o fundadores) a la población con referencia a los árboles genealógicos de individuos de la población actual y la selección de aquellos individuos que proporcionan las mayores aportaciones a largo plazo a la población actual, de tal manera que se seleccione el número más pequeño de antepasados (y/o fundadores) para describir sustancialmente la variación en la población actual.

Como se usa en este documento, la expresión "conjunto de antepasados (y/o fundadores) que son representativos de una población actual" significa que el conjunto de antepasados (y/o fundadores) representa la mayor parte de la variación en la población actual, es decir, la suma de todos los antepasados (y/o fundadores) en el conjunto describe sustancialmente la variación en la población actual. La expresión "describe sustancialmente la variación en la población actual" significa al menos aproximadamente el 70 %, preferentemente al menos aproximadamente el 80 % y aún más preferentemente al menos aproximadamente el 90 % de la variación total en la población actual.

En cada uno de los métodos y usos descritos en el presente documento, se prefiere que los árboles genealógicos de los individuos en la población actual esté completo o casi completo, es decir, que comprenda al menos aproximadamente el 80 % de los antepasados (y/o fundadores), o el 85 % de los antepasados (y/o fundadores) o el 90 % de los antepasados (y/o fundadores) o el 95 % de los antepasados (y/o fundadores) o el 99 % de los antepasados (y/o fundadores) o el 100 % de los antepasados (y/o fundadores). En dichas circunstancias, la proporción acumulada de genes aportados por los antepasados (y/o fundadores) a una población actual será de al menos aproximadamente 80 % y preferentemente de al menos aproximadamente 90 % o 95 % o 99 % o 100 %. En casos en los que los datos del árbol genealógico estén incompletos, la presente invención claramente abarca el uso de uno o más marcadores para inferir el árbol genealógico de uno o más animales de una población actual que tiene un árbol genealógico incompleto.

Las realizaciones anteriores describen el uso de genotipos de antepasados y/o fundadores para inferir los genotipos de candidatos de selección en programas de reproducción. Sin embargo, debe entenderse que, a pesar de la aplicabilidad general de la presente invención al uso de genotipos de antepasados y/o fundadores para este propósito, se prefiere el uso de genotipos de antepasados porque los conjuntos de datos son generalmente más pequeños que para las poblaciones fundadoras y por lo tanto, proporcionan una mayor ventaja en términos de reducción de costes que los genotipos basados en genotipos de fundadores.

La presente invención se realiza sin excesiva experimentación utilizando, a menos que se indique lo contrario, técnicas convencionales de biología molecular, microbiología, virología, tecnología de ADN recombinante, síntesis de péptidos en solución, síntesis de péptidos en fase sólida e inmunología.

Las siguientes referencias reflejan técnica pertinente:

1. Sambrook, Fritsch y Maniatis, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratories, Nueva York, segunda edición (1989), con todas los Vols I, II y III;
2. DNA Cloning: A Practical Approach, Vols. I y II (D. N. Glover, ed., 1985), IRL Press, Oxford, texto completo;
3. Oligonucleotide Synthesis: A Practical Approach (M. J. Gait, ed., 1984) IRL Press, Oxford, texto completo y particularmente documentos en su interior de Gait, pp1-22; Atkinson et al., pp35-81; Sproat et al., pp 83-115; y Wu et al., pp 135-151;
4. Nucleic Acid Hybridization: A Practical Approach (B. D. Hames y S. J. Higgins, eds., 1985) IRL Press, Oxford, texto completo;
5. Perbal, B., A Practical Guide to Molecular Cloning (1984);
6. Bulmer, M.G., The mathematical theory of quantitative genetics. Clarendon Press, Oxford, (1980);
7. Falconer D.S., Introduction to Quantitative Genetics. Oliver & Boyd, London (1960);
8. Falconer D.S., Introduction to Quantitative Genetics. Segunda edición, Longmann, Londres (1981);
9. Falconer D.S., Introduction to Quantitative Genetics. Tercera edición, Longmann, Londres (1989);
10. Falconer D.S., Mackay T.F.C., Introduction to Quantitative Genetics. Cuarta edición, Longmann y Co, Londres (1996); y
11. Kearsey M., Pooni HS., 1996. The Genetical Analysis of Quantitative traits. Chapman y Hall, Londres (1996)

Descripción detallada de las realizaciones preferidas

Especies que tienen tamaños de población eficaces pequeños (Ne)

La presente invención se aplica fácilmente a cualquier contexto de selección artificial o de reproducción que involucre a individuos de poblaciones eficaces pequeñas, especialmente de poblaciones a las que se les ha reducido

el tamaño eficaz de su población, por ejemplo, mediante reproducción selectiva. Los métodos estándar conocidos por los expertos en la técnica se utilizan para determinar el tamaño eficaz de la población.
Por ejemplo, el tamaño eficaz de la población Ne se calcula como:

$$Ne = 1 / (2\Delta F)$$

en donde F es el coeficiente de endogamia, una medida de la cantidad de diversidad genética que se ha perdido tal como mediante endogamia. El término ΔF puede estimarse por regresión de coeficientes de endogamia individuales en el número de generación. El cambio en la endogamia por generación puede utilizarse después para estimar el número eficaz de animales reproductores (Ne). El propósito del tamaño eficaz de la población es estimar el número de animales que producirían una tasa observada de endogamia si se criaran en condiciones ideales de apareamiento aleatorio en la generación actual (Lacy, Zoo Biol. 14, 565-578, 1995).

Los ejemplos de métodos para determinar el tamaño eficaz de la población se describen en la referencia enumerada en la Tabla 1. Las poblaciones preferidas que tienen un tamaño de población eficaz pequeño se habrán producido más o menos recientemente, por ejemplo, de más de 4 a 10 generaciones, en virtud de un cuello de botella poblacional, o como alternativa, durante un período de tiempo durante el cual se dispone de los datos de árbol genealógico sobre antepasados significativos. Esto es para permitir que la cobertura suficiente del genoma de la población actual sea inferida por los haplotipos de los antepasados altamente significativos que aportan la mayor parte de la variación genética a la población actual. Ejemplos de poblaciones a las que se aplica fácilmente la presente invención incluyen ganado (p. ej., vacuno y lechero como Holstein, Frisona, Holstein-Frisona, Braunvieh, Brown Swiss, Jersey, Danish Red, Aberdeen Angus), ovejas (por ejemplo, cruces de Meatlinc, Dorset x Rambouillet x. Finnsheep), cerdos (por ejemplo, cruce de Large White x Landrace, Large White, Duroc, Yorkshire, Landrace), aves de corral (por ejemplo, ponedoras), pescado (por ejemplo, salmón del Atlántico), crustáceos, raigrás, etc.. El tamaño estimado de la población eficaz (Ne) de algunas de estas poblaciones de animales se muestran en la Tabla 1 de este documento.

Tabla 1

Especie	Raza	Ne estimado	Referencia
Bovina			
	Holstein-Frisona	50	Boichard INRA Prod. Anim. 9, 323-335 (1996)
	Holstein-Frisona	100	Young et al., J. Dairy Sci. 79, 502-505 (1996)
	Braunvieh	114	Hagger, J. Anim. Breed. and Genet. 22, 405 (2005)
	Brown Swiss	46	Hagger, J. Anim. Breed. and Genet. 22, 405 (2005)
	Holstein	49	Sorensen et al., J Dairy Sci. 88, 1865-1872 (2005)
	Jersey	53	Sorensen et al., J Dairy Sci. 88, 1865-1872 (2005)
	Danish Red	47	Sorensen et al., J Dairy Sci. 88, 1865-1872 (2005)
Ovina			
	Cruce Dorset-Rambouliet-Finnsheep	35	Mackinnon et al., J. Anim Sci 81 (Supp.1), p267 (2003)
Porcina			
	Entrecruzamiento Large White/Landrace	<200*	Harmegnies et al., Anim Genet. 37, 225-231 (2006)
	Large White	200	Nsengimana et a/., Genetics 166, 1395-1404 (2004)
	Duroc/Large White	85	Nsengimana et a/., Genetics 166, 1395-1404 (2004)

Especie	Raza	Ne estimado	Referencia
	Yorkshire/Large White	60	Nsengimana et a/., Genetics 166, 1395-1404 (2004)
	Large White	300	Nsengimana et a/., Genetics 166, 1395-1404 (2004)
	Landrace	190	Nsengimana et a/., Genetics 166, 1395-1404 (2004)
Gallinas			
	Ponedoras	91-123	Hagger et al., J. Anim Breed Genet. 122 (Suppl 1), 15-21 (2005)
Salmón del Atlántico			
	Programa de reproducción de la población	50-200	Mork et al., Norges Offentlige Utredninger 9, 181-200 (1999)

Definición de antepasados y fundadores

Para determinar las aportaciones de antepasados/fundadores a una población actual, preferentemente en poblaciones para las cuales los datos de árbol genealógico están completos o casi completos, por ejemplo, al menos aproximadamente 80-90 % completos o al menos aproximadamente 85- 95 % completos, y más preferentemente al menos aproximadamente 90 % o 95 % o 96 % o 97 % o 98 % o 99 % completos, se utilizan métodos convencionales.

Por ejemplo, el cálculo obtenido por Boichard et al. (1997) puede utilizarse para identificar a los antepasados y/o fundadores más influyentes en un árbol genealógico:

$$fa = \sum_{i=1}^{m} ai^2$$

en donde ai es la aportación marginal de cada antepasado/fundador (es decir, cualquier animal en el árbol genealógico, excepto los animales de la generación actual), comparado con cada fundador, a la generación actual, y m es el número total de antepasados contributivos. La aportación marginal de todos los antepasados/fundadores debe sumar uno, y el número eficaz de antepasados siempre es menor o igual al número eficaz de fundadores. Las aportaciones individuales al número eficaz de antepasados/fundadores pueden utilizarse para encontrar a los antepasados y/o fundadores más influyentes. Los números obtenidos a partir del cálculo de fa tienen en cuenta una disminución en la variación genética en poblaciones que han pasado por un cuello de botella. Es el individuo que transfiere la mayoría de los genes a una población actual que hace la mayor aportación. Aunque un antepasado influyente (por ejemplo, el hijo de un padre) transfiere la mayoría de sus genes a través de muchos descendientes, solo tiene la mitad de los genes de su padre fundador. Los animales en la población actual en estudio reciben un valor de uno y las aportaciones marginales se obtienen procesando el árbol genealógico del más joven al más mayor. Cuando se identifica un antepasado / fundador importante (un animal con la mayoría de las relaciones con la población actual), la información de su padre y madre se elimina del árbol genealógico, por lo que las aportaciones a la población actual no se cuentan dos veces. Un algoritmo vuelve a ejecutar los cálculos cada vez que se elimina un antepasado, de modo que las aportaciones marginales que no se deben al antepasado ya seleccionado son las únicas medidas (Boichard et al, Genet. Set Evol. 29, 5-23,1997).

Del ejemplo del párrafo anterior, es evidente que, si un hijo fue seleccionado como un antepasado importante, su padre no obtendría acreditación por sus aportaciones a través de su hijo influyente en la próxima iteración.

Existe una discrepancia que puede ocurrir cuando se estima fa a través de la ecuación anterior. Dado que los animales se seleccionan en función de su aportación marginal, si varios animales tienen la misma aportación marginal dentro de una iteración, entonces el número de antepasados eficaces puede cambiar dependiendo de cuál se elija. Las poblaciones grandes no se ven muy afectadas, pero en las poblaciones pequeñas podría haber un mayor efecto en el fa, ya que las aportaciones marginales tienen el potencial de ser más grandes. La variable fa explica cuellos de botella en el árbol genealógico, pero la deriva genética. El cálculo es útil para identificar a los antepasados más influyentes, que pueden ser importantes en poblaciones seleccionadas.

Como alternativa, los antepasados y/o fundadores significativos pueden determinarse considerando el efecto de diferentes cohortes de antepasados/fundadores en la ganancia genética, según se determina estudiando la relación entre las aportaciones genéticas a largo plazo de los antepasados/fundadores y las puntuaciones de índice, esencialmente como describen Avendano et al, J. Anim. ScL 81, 2964-2975 (2003). En este método, la aportación a largo plazo (r) se calcula siguiendo la estrategia de Woolliams et al., Anim. Sci. 61, 177-187 (1995) en donde para calcular r, una generación de antepasados/fundadores y una generación de descendientes se definen de acuerdo con los intervalos de generación promedio calculados previamente de tal manera que las generaciones ancestrales y descendientes se definan utilizando el intervalo de generación (L) . Esta definición garantiza que r sumado sobre todos los antepasados durante un período de L y r es igual a la unidad (Bijma et al., Genetics 151, 1197-1210, 1990). Se asume la convergencia de aportaciones si la variación de las aportaciones de los antepasados/fundadores entre los descendientes es inferior a $1,0 \times 10^{-4}$. La regresión de la aportación genética a largo plazo de los antepasados / fundadores en sus puntuaciones de índice se calcula para cada cohorte de antepasados.

En un ejemplo particularmente preferido, los antepasados se definen por un método de la divulgación que comprende determinar las aportaciones a largo plazo de los antepasados y/o fundadores a la población con referencia al árbol genealógico de los individuos de la población actual y seleccionar a aquellos individuos que proporcionan las mayores aportaciones a largo plazo a la población actual, de manera que se seleccione el número más pequeño de antepasados y/o fundadores para describir sustancialmente la variación en la población actual.

Por ejemplo, dejando que se apliquen las siguientes condiciones o suposiciones a la relación de cualquier antepasado o fundador o grupo de antepasados o fundadores a una población actual:

(i) n representa el número de posibles antepasados y/o fundadores;
(ii) A es una matriz de relación aditiva nxn entre n posibles antepasados y/o fundadores;
(iii) c es un vector nx1 con los n posibles antepasados y/o fundadores ordenados de la misma manera que en la matriz de relación aditiva A;
(iv) c_i, es la relación promedio de antepasado/fundador i con una población actual, es decir, la fracción de genes en la población actual que procede directa o indirectamente del antepasado/fundador *i*; y
(v) $A_m$es una submatriz de A que describe la relación entre *m* de los antepasados y/o fundadores;
(vi) c_m es un subvector de c que describe la relación entre *m* de los antepasados y/o fundadores y la población actual;
(vii) p es un vector que tiene un elemento *i* igual a la proporción de genes en una población que procede únicamente del antepasado/fundador *i*; y
(viii) p'1 es la proporción de genes en la población que procede de m antepasados y/o fundadores determinada como un total de los elementos de *p*.

Por lo tanto,

$$p = A_m^{-1} c_m$$

Esto significa que pueden seleccionarse los antepasados y/o fundadores clave determinando un subconjunto de antepasados que maximicen p'1. Por ejemplo, los antepasados y/o fundadores más significativos para una población pueden seleccionarse gradualmente, a través de: (i) la selección de un antepasado o fundador que aporte la mayor proporción de genes a la población actual; (ii) la selección de un antepasado o fundador que proporcione la mayor aportación marginal de los genes en comparación con el antepasado en (i); y (iii) la realización de iteraciones suficientes de (ii) para describir sustancialmente la variación en la población actual.

Por ejemplo, como se muestra en la Tabla 2, utilizando esta estrategia, se ha determinado un conjunto de aproximadamente 25 antepasados significativos en la población Holstein Frisona australiana. Los nombres completos de estos antepasados también se enumeran en la Tabla 3 de este documento que proporciona una lista más completa de los antepasados clave y la disponibilidad de datos de semen o genotipo en esos antepasados clave. En las Tablas 2 y 3 se indica un antepasado por línea. En el ejemplo proporcionado en la Tabla 2, la proporción acumulada de genes aportados a la población actual es de aproximadamente 33 %, posiblemente debido a que los árboles genealógicos de los animales en la población actual están incompletos. Como puede observarse en los datos de la Tabla 3, de 100 antepasados clave, solo se ha genotipado aproximadamente la mitad y hay reservas limitadas de semen disponibles para el genotipado, por lo que es necesario inferir los genotipos ausentes.

Es preferible que la proporción acumulativa de genes aportados a la población actual sea más de aproximadamente 80 % o 90 % o 95 % o 99 % o 100 %. En circunstancias en las que este no sea el caso, se prefiere utilizar una población que tenga un mejor registro de árbol genealógico y/o utilizar los propios marcadores para inferir el árbol genealógico de animales con árboles genealógicos incompletos.

Tabla 2

Ejemplos de antepasados clave en la población de ganado Holstein-Frisona australiana	
Nombre del antepasado	Proporción acumulativa de genes aportados a la población actual
VALIANT	0,05371
OAK RAG APPLE ELEVATION	0,09220
IVANHOE BELL	0,12534
STARBUCK	0,15426
MASCOT	0,18180
BLACKSTAR	0,20459
ENHANCER	0,22700
LINMACK KRISS KING	0,24425
ROTATE	0,25660
TRADITION CLEITUS	0,26798
ROYBROOK TELSTAR	0,27699
PACLAMAR ASTRONAUT	0,28531
FOND MATT	0,29287
WHITTIER-FARMS NED BOY	0,29872
ROYBROOK STARLITE	0,30399
WAPA ARLINDA CONDUCTOR	0,30851
ROSAFE CITATION	0,31247
CAM VIEW SOVEREIGN	0,31638
KIRK JUPITER	0,32027
TRAILYND ROYAL BEAU	0,32374
AGRO ACRES MARQUIS NED	0,32717
RONNYBROOK PRELUDE	0,33046
SUNNY BOY	0,33379
HILL NSPIRATION	0,33658
VIC KAI	0,33915

Tabla 3

Ejemplos de antepasados clave en la población de ganado Holstein-Frisona Australiana

AUS_ID	International_ID	Nombre del antepasado	Reservas de semen[1]	Registros de genotipos[1]
A00000013	HOAUS000A00000013	WARRAWEE ADEMA AMBASSADOR	0	0
A00000096	HOAUS000A00000096	KOTAHA KIM	0	0
A00000103	HOAUS000A00000103	VICTORIA OLSON	0	0
A00000193	HOAUS000A00000193	CAM VIEW SOVEREIGN	0	0
A00000207	HOAUS000A00000207	CLINELL RAELENE MAGIC	0	0
A00000253	HOAUS000A00000253	GLENJOY GRIFFLAND RANDY	0	0
A00001001	HOAUS000A00001001	SNIDERS FOND HOPE KING	0	0
A00001037	HOAUS000A00001037	FRASEA LORD JEWEL	0	0
A00001555	HOAUS000A00001555	HADSPEN BUTTERMAN 01 F B BUTTE	0	0

AUS_ID	International_ID	Nombre del antepasado	Reservas de semen[1]	Registros de genotipos[1]
A00001643	HOAUS000A00001643	CLARIS VALE MASTER	0	0
A00001742	HONZL000000062011	PUKERORO ISAR IMPERIAL	0	0
A00001744	HONZL000000004209	GAYTON INGA VIC	0	0
A00001746	HONZL000000005211	WINDERMERE PERFECT MAX	0	0
A00001753	HONZL000000007213	LYNCREST S Q VICTOR	0	0
A00001756	HONZL000000062147	PUKERORO NORBERT LOCK	0	0
A00001786	HONZL000000062387	ATHOL SOVEREIGN FAME	0	0
A00001911	HOGBR000000265781	SUTTONHOO IDENA DIVIDEND PI	0	0
A00001931	HOCAN000000292057	FREELEA INKA JERRY	0	0
A00001938	HOGBR000000303735	MMB OAKRIDGES REFLECTION PI	0	0
A00001941	HOCAN000000294213	LINMACK KRISS KING	0	0
A00001957	HOCAN000000313602	AGRO ACRES REVENUE	0	0
A00001975	HOGBR000000360323	LOCUSLANE SUPREME	0	0
A00002138	HOCAN000000280596	EDGEWARE WAYNE ACHILLES	0	0
A00002142	HOCAN000000289318	TAYSIDE PABST ROCKMAN	0	0
A00002144	HOCAN000000290516	AGRO ACRES MARQUIS NED	0	0
A00002145	HOCAN000000293299	WAY BROOK SIR WINSTON	0	0
A00002148	HOCAN000000302981	MOOREVILLE ROCKET KEMP	0	0
A00002169	HOCAN000000320891	QUALITY ULTIMATE	0	0
A00002296	HONZL000000062112	PITCAIRNS T B TOPPER	0	0
A00002798	HOCAN000000276333	BOND HAVEN SOVEREIGN	0	0
A00002935	HONZL000000000161	KITEROA MUTUAL MIKE	0	0
A00002944	HONZL000000027893	OTAKI H C T GRAHAM	0	0
A00003116	HOCAN000000364963	ALBRECHT CASCADE	0	0
A00005113	HOCAN000000299855	FAIRLEA ROYAL MARK	0	0
A00005114	HOCAN000000288790	ROYBROOK TELSTAR	0	0
A00005147	HOUSA000001392858	NO-NA-ME FOND MATT	0	0
A00005158	HOCAN000000267150	ROSAFE CITATION R	0	0
A00005482	HOUSA000001721509	BROWNCROFT JETSON	0	0
A00006409	HOUSA000001617348	HILLIANA VALEDICTORIAN	0	0
A00006411	HOUSA000001685359	DONACRES DYNAMO-TWIN	0	0
A00006862	HOUSA000001516360	HEINDEL K C KIRK JUPITER	0	0
A00008144	HONLD000311651443	SKALSUMER SUNNY BOY	0	0
A00011920	HONZL000000093290	ATHOL MURRAYS EMINENCE	0	0
A00012752	HONZL000000096329	SRB COLLINS ROYAL HUGO	0	0
A00014530	HOCAN000000259668	GLENHOLM ALERT DEAN PABST	0	0
A00014543	HODNK000000010763	VARARLI	0	0
A00014643	HOUSA000001282262	ELLBANK ADMIRAL ORMSBY PRIDE	0	0
A00014647	HOUSA000001242221	POLYTECHNIC IMPERIAL KNIGHT	0	0
A00014648	HOUSA000001199324	SKOKIE BENEFACTOR	0	0
A00014679	HOUSA000001531866	PACLAMAR COMBINATION	0	0
A00014692	HOUSA000001648394	ACK-LEE CHIEF MONEY MAKER	0	0
A00017159	HOUSA000001352979	SKOKIE NED BOY	0	0
A00000378	HOAUS000A00000378	ONKAVALE GRIFFLAND MIDAS	1	1
A00001061	HOAUS000A00001061	TRAILYND ROYAL BEAU	1	1
A00001978	HOGBR000000370161	DALESEND CASCADE PI	1	1

AUS_ID	International_ID	Nombre del antepasado	Reservas de semen[1]	Registros de genotipos[1]
A00002502	HOCAN000000340909	CAL-CLARK CUTLASS	1	1
A00004350	HOCAN000000371440	HANOVERHILL SABASTIAN ET	1	1
A00006720	HOCAN000000402729	MEADOW BRIDGE MANHATTAN	1	1
A00006889	HOAUS000A00006889	SHOREMAR PERFECT STAR (ET)	1	1
A00011268	HONLD000829877874	HOLIM BOUDEWIJN	1	1
A00006484	HOUSA000001747862	COR-VEL ENCHANTMENT	1	2
A00006968	HOUSA000001772090	CRESCENTMEAD CHIEF STEWART	1	2
A00014532	HOCAN000000260599	ROSAFE SHAMROCK PERSEUS	1	2
A00014669	HOUSA000001563453	WILLOW-FARM ROCKMAN IVANHOE	1	2
A00015051	HOUSA000001483844	HARBORCREST HAPPY CRUSADER	1	2
A00004325	HOUSA000001781631	ROBE-JAN SKYLER CHIEF	1	1
A00004805	HOCAN000000363162	HANOVER-HILL INSPIRATION	1	1
A00005146	HOUSA000001626813	MARSHFIELD ELEVATION TONY	1	1
A00005148	HOUSA000001458744	PACLAMAR ASTRONAUT	1	1
A00005149	HOUSA000001450228	PACLAMAR BOOTMAKER	1	1
A00005151	HOUSA000001491007	ROUND OAK RAG APPLE ELEVATION	1	1
A00005152	HOUSA000001650414	S-W-D VALIANT	1	1
A00005154	HOUSA000001583197	WAPA ARLINDA CONDUCTOR	1	1
A00005156	HOUSA000001381027	IDEAL FURY REFLECTOR	1	1
A00005424	HOUSA000001806201	WHITTIER-FARMS NED BOY	1	1
A00005425	HOUSA000001667366	CARLIN-M IVANHOE BELL	1	1
A00005426	HOUSA000001682485	SWEET-HAVEN TRADITION	1	1
A00005569	HOUSA000001697572	ARLINDA ROTATE	1	1
A00005707	HOUSA000001879085	BIS-MAY TRADITION CLEITUS	1	1
A00006187	HOUSA000001929410	TO-MAR BLACKSTAR-ET	1	1
A00006410	HOUSA000001512026	HARRISBURG GAY IDEAL	1	1
A00007236	HOUSA000001930394	HICKS-HOLLOW PROMPT	1	1
A00007435	HOCAN000000392457	A RONNYBROOK PRELUDE ET	1	1
A00007990	HOUSA000001874634	HOW-EL-ACRES K BELLMAN-ET	1	1
A00014631	HOUSA000001399824	HILLTOP APOLLO IVANHOE	1	1
A00014632	HOUSA000001393997	PROVIN MTN IVANHOE JEWEL	1	1
A00014636	HOUSA000001428104	SUNNYSIDE STANDOUT-TWIN	1	1
A00014670	HOUSA000001560362	C ROMANDALE SHALIMAR MAGNET	1	1
A00014702	HOUSA000001608425	ARLINDA CINNAMON	1	1
A00014705	HOUSA000001674245	I-O-STATE CHIEF FORD	1	1
A00002151	HOCAN000000308691	ROYBROOK STARLITE	2	1
A00002677	HOCAN000000343514	GLENAFTON ENHANCER	2	1
A00003460	HOCAN000000352790	HANOVERHILL STARBUCK	2	1
A00005339	HOUSA000001856904	THONYMA SECRET	2	1
A00006485	HOUSA000001964484	SOUTHWIND BELL OF BAR-LEE	2	1
A00006577	HOCAN000000383622	MADAWASKA AEROSTAR	2	1
A00007094	HOUSA000002005253	PICKARD-ACRES VIC KAI	2	1
A00007170	HOUSA000002020049	SINGING-BROOK N-B MASCOT-ET	2	1

AUS_ID	International_ID	Nombre del antepasado	Reservas de semen[1]	Registros de genotipos[1]
A00008149	HOUSA000002070579	BIS-MAY S-E-L MOUNTAIN-ET	2	1
A00010003	HONLD000775328514	EASTLAND CASH	2	1

1, los números indican totales acumulados para Australia y Estados Unidos

Métodos de genotipado

El genotipado generalmente implica detectar uno o más marcadores de interés, por ejemplo, SNP (polimorfismos mononucleotídicos) en una muestra de un individuo que se somete a ensayo, y analizar los resultados obtenidos para determinar el haplotipo del sujeto. Como será evidente a partir de la divulgación de este documento, se prefiere particularmente detectar el uno o más marcadores de interés utilizando un sistema de alto rendimiento que comprende un soporte sólido que consiste esencialmente en, o que tiene, ácidos nucleicos de diferente secuencia unidos directa o indirectamente al mismo, en donde cada ácido nucleico de diferente secuencia comprende un marcador genético polimórfico procedente de un antepasado o fundador que es representativo de la población actual y, más preferentemente, en donde dicho sistema de alto rendimiento comprende marcadores suficientes para ser representativos del genoma de la población actual.

Muestras adecuadas para el genotipado

Las muestras preferidas comprenden ácido nucleico, por ejemplo, ARN o ADN genómico y preferentemente ADN genómico.

Por ejemplo, los ensayos genéticos de plantas pueden implicar ensayos de cualquier parte de la planta, por ejemplo, hojas, órganos florales, semillas, etc.

El ensayo genético de animales puede realizarse utilizando un folículo piloso, por ejemplo, aislado de la cola de un animal que va a someterse a ensayo. Otros ejemplos de muestras fácilmente accesibles incluyen, por ejemplo, piel o un fluido corporal o un extracto del mismo o una fracción del mismo. Por ejemplo, un fluido corporal fácilmente accesible incluye, por ejemplo, sangre entera, saliva, semen u orina. Los ejemplos de fracciones de sangre entera se seleccionan del grupo que consiste en una fracción de la capa leucocitaria, una fracción IH-III obtenible por fraccionamiento con etanol de Cohn (EJ Cohn et al., J. Am. Chem. Soc, 68, 459 (1946), una fracción II obtenible por fraccionamiento con etanol de Cohn (EJ Cohn et al., J. Am. Chem. Soc., 68, 459 (1946), una fracción de albúmina, una fracción que contiene inmunoglobulina y mezclas de los mismos. Preferentemente, una muestra de un animal se ha aislado u obtenido previamente de un sujeto animal, por ejemplo, mediante cirugía, o utilizando una jeringa o un hisopo.

En otra realización, una muestra puede comprender una célula o extracto celular o una mezcla de los mismos procedente de un tejido u órgano tal como se ha descrito anteriormente en el presente documento. También es particularmente útil la preparación de ácido nucleico procedente de órganos, tejidos o células.

La muestra puede prepararse en una matriz sólida para realizar análisis histológicos, o como alternativa, en una solución adecuada tal como, por ejemplo, un tampón de extracción o un tampón de suspensión, y la presente invención se extiende claramente al análisis de soluciones biológicas preparadas de este modo. Sin embargo, en una realización preferida, el sistema de alto rendimiento de la presente invención se emplea utilizando muestras en solución.

Diseño de sonda/cebador

El experto en la materia sabe que una sonda o un cebador adecuado, es decir, uno capaz de detectar específicamente un marcador, se hibridará específicamente con una región del genoma en el ADN genómico del individuo que se está sometiendo a ensayo y que comprende el marcador. Como se usa en este documento, "hibrida selectivamente" significa que el polinucleótido utilizado como sonda se utiliza en condiciones en las que se encuentra que un polinucleótido diana se hibrida con la sonda a un nivel significativamente por encima del fondo. La hibridación de fondo puede ocurrir debido a la presencia de otros polinucleótidos, por ejemplo, ADN genómico a explorar. En este caso, el fondo implica un nivel de señal generado por la interacción entre la sonda y el ADN no específico que es menor de 10 veces, preferentemente menor de 100 veces tan intenso con la interacción específica observada con el ADN diana. La intensidad de la interacción se mide, por ejemplo, marcando la sonda con átomos radioactivos (radiomarcado), por ejemplo, con ^{32}P.

Como sabrá el experto en la materia, una sonda o un cebador comprende ácido nucleico y puede consistir en oligonucleótidos sintéticos con una longitud de hasta aproximadamente 100-300 nucleótidos y más preferentemente una longitud de aproximadamente 50-100 nucleótidos y aún más preferentemente una longitud de al menos aproximadamente 8-100 u 8-50 nucleótidos. Por ejemplo, para la detección de uno o más SNP, las sondas u

oligobalizas de ácido nucleico bloqueado (LNA, *locked nucleic acid*) o de proteína-ácido nucleico (PNA, *protein-nucleic acid*) tienen generalmente una longitud de al menos aproximadamente 8 a 12 nucleótidos. También pueden utilizarse fragmentos de ácido nucleico más largos de hasta varias kilobases de longitud, por ejemplo, obtenidos de ADN genómico que se ha cizallado o digerido con una o más endonucleasas de restricción. Como alternativa, las sondas/cebadores pueden comprender ARN.

Las sondas o cebadores preferidos para su uso en la presente invención serán compatibles con el sistema de alto rendimiento descrito en el presente documento. Los ejemplos de sondas y cebadores comprenderán sondas u oligobalizas de ácido nucleico bloqueado (LNA) o de proteína-ácido nucleico (PNA), preferentemente unidas a una fase sólida. Por ejemplo, se utilizan sondas de LNA o PNA unidas a un soporte sólido, en donde cada una de ellas comprende un SNP y se unen suficientes sondas al soporte sólido para abarcar el genoma de la especie a la que pertenece un individuo que se está sometiendo a ensayo.

El número de sondas o cebadores variará dependiendo del número de locus o QTL que se esté explorando y, en el caso de exploraciones hologenómicas, del tamaño del genoma que se esté explorando. La determinación de dichos parámetros la determina con facilidad un experto en la materia sin excesiva experimentación.

La especificidad de las sondas o cebadores también puede depender del formato de hibridación o de la reacción de amplificación empleada para el genotipado.

La secuencia (o secuencias) de cualquier sonda(s) o cebador(es) particular(es) que se utilice(n) en el método de la presente invención, dependerá del locus o QTL o combinación de los mismos, que se esté explorando. A este respecto, la presente invención puede aplicarse en general al genotipado de cualquier locus o QTL o al genotipado simultáneo o secuencial de cualquier número de QTL o locus, incluido el genotipado hologenómico. Esta generalidad no debe quitarse de, ni leerse en, un locus o QTL específico, o combinación de los mismos. La determinación de las secuencias de la sonda/el cebador la determina con facilidad un experto en la técnica sin excesiva experimentación.

Para diseñar sondas y/o cebadores se emplean métodos convencionales, por ejemplo, como los descritos por Dveksler (Eds) (En: PCR Primer: A Laboratory Manual, Cold Spring Harbor Laboratories, NY, 1995). También se dispone de paquetes de programas informáticos para diseñar sondas y/o cebadores óptimos para una variedad de ensayos, por ejemplo, Primer 3 disponible en el Centro de Investigación del Genoma, Cambridge, MA, EE. UU. Las sondas y/o cebadores se evalúan preferentemente para determinar cuáles son los que no forman horquillas, se autoceban o forman dímeros de cebadores (por ejemplo, con otra sonda o cebador utilizado en un ensayo de detección). Además, una sonda o un cebador (o su secuencia) se evalúa preferentemente para determinar la temperatura a la cual se desnaturaliza de un ácido nucleico diana (es decir, la temperatura de fusión de la sonda o del cebador, o Tf). En la técnica se conocen métodos para determinar la Tf y se describen, por ejemplo, en Santa Lucía, Proc. Natl Acad Sci. USA, 95: 1460-1465, 1995 o en Bresslauer et al, Proc. Natl Acad Sci. USA, 83: 3746-3750, 1986.

Para las sondas u oligobalizas de LNA o PNA, se prefiere particularmente que la sonda u oligobaliza tenga una longitud de al menos aproximadamente 8 a 12 nucleótidos y, más preferentemente, que el SNP, se coloque aproximadamente el centro de la sonda, facilitando así una hibridación selectiva y una detección exacta.

Para detectar uno o más SNP utilizando un ensayo de PCR específico de alelo o un ensayo de reacción en cadena de la ligasa, la sonda/cebador generalmente se diseña de manera que el nucleótido terminal en posición 3’ se hibride en el sitio del SNP. El nucleótido terminal en posición 3’ puede ser complementario a cualquiera de los nucleótidos que se sabe que están presentes en el sitio del SNP. Cuando se producen nucleótidos complementarios tanto en la sonda/cebador como en el sitio del polimorfismo, el extremo 3' de la sonda o cebador se hibrida completamente con el marcador de interés y facilita, por ejemplo, la amplificación por PCR o el ligamiento con otro ácido nucleico. Por consiguiente, una sonda o un cebador que se hibrida completamente con el ácido nucleico diana produce un resultado positivo en un ensayo.

Para las reacciones de extensión con cebador, generalmente, la sonda/el cebador, se diseña de modo que se hibride específicamente con una región adyacente a un nucleótido específico de interés, por ejemplo, un SNP. Si bien la hibridación específica de una sonda o un cebador puede estimarse determinando el grado de homología de la sonda o del cebador con cualquier ácido nucleico utilizando un programa informático, tal como, por ejemplo, BLAST, la especificidad de una sonda o de un cebador generalmente se determina empíricamente utilizando métodos conocidos en la técnica.

En la materia se conocen métodos para producir/sintetizar sondas y/o cebadores útiles en la presente invención. Por ejemplo, la síntesis de oligonucleótidos se describe en Gait (Ed) (En: Oligonucleotide Synthesis: A Practical Approach, IRL Press, Oxford, 1984); La síntesis de LNA se describe, por ejemplo, en Nielsen et al, J. Chem. Soc. Perkin Trans., 1: 3423, 1997; Singh y Wengel, Chem. Comun. 1247, 1998; y la síntesis de PNA se describe, por ejemplo, en Egholm et al., Am. Chem. Soc, 114: 1895, 1992; Egholm et al, Nature, 365: 566, 1993; y Orum et al., Nucl. Acids Res., 21: 5332, 1993.

Métodos de detección de marcadores

En la técnica se conocen numerosos métodos para determinar la aparición de un marcador particular en una muestra.

En una realización preferida, un marcador se detecta utilizado una sonda o un cebador que se hibrida selectivamente con dicho marcador en una muestra de un individuo en condiciones de rigurosidad moderada y, preferentemente, en condiciones de rigurosidad elevada. Si la sonda o el cebador está marcado de manera detectable con una molécula indicadora adecuada, por ejemplo, un marcador quimioluminiscente, un marcador fluorescente, un marcador radioactivo, una enzima, un hapteno, una secuencia de oligonucleótidos única, etc., la hibridación puede detectarse directamente determinando la unión de la molécula indicadora. Como alternativa, la sonda o el cebador hibridado puede detectarse realizando una reacción de amplificación tal como la reacción en cadena de la polimerasa (PCR) o un formato similar, y detectando el ácido nucleico amplificado. Preferentemente, la sonda o el cebador está unido a un soporte sólido, por ejemplo, en el sistema de alto rendimiento de la presente invención.

Para definir el nivel de rigurosidad que se utilizará en la hibridación, en el presente documento una rigurosidad baja se define como una etapa de hibridación y/o una o más etapas de lavado realizadas en tampón SSC 2-6 x, SDS al 0,1 % (p/v) a 28 °C, o condiciones equivalentes. En el presente documento una rigurosidad moderada se define como una etapa de hibridación y/o una o más etapas de lavado realizadas en tampón SSC 0,2-2 x, SDS al 0,1 % (p/v) a una temperatura en el intervalo de 45 C a 65 °C, o condiciones equivalentes. En el presente documento una rigurosidad elevada se define como una etapa de hibridación y/o una o más etapas de lavado realizadas en tampón SSC 0,1 x, SDS al 0,1 % (p/v), o una menor concentración salina, y a una temperatura de al menos 65 °C, o condiciones equivalentes. La referencia en este documento a un nivel de rigurosidad particular incluye condiciones equivalentes que utilizan soluciones de lavado/hibridación distintas a las de SSC (*Saline-Sodium Citrate*) conocidas por los expertos en la materia.

En general, la rigurosidad se incrementa al reducir la concentración de tampón SSC, y/o al aumentar la concentración de SDS y/o al aumentar la temperatura de la hibridación y/o lavado. Los expertos en la materia sabrán que las condiciones para la hibridación y/o lavado pueden variar dependiendo de la naturaleza de la matriz de hibridación utilizada para dar soporte a la muestra de ADN, o del tipo de sonda de hibridación utilizada.

También pueden emplearse condiciones de rigurosidad progresivamente más elevadas, en las que la rigurosidad se incrementa gradualmente desde condiciones de rigurosidad más bajas a más elevadas. Son ejemplos de condiciones de rigurosidad progresiva los siguientes: 2 x SSC/SDS al 0,1 % aproximadamente a temperatura ambiente (condiciones de hibridación); 0,2 x SSC/SDS al 0,1 % aproximadamente a temperatura ambiente (condiciones de rigurosidad baja); 0,2 x SSC/SDS al 0,1 % a aproximadamente 42 ºC (condiciones de rigurosidad moderada); y 0,1 x SSC a aproximadamente 68 °C (condiciones de rigurosidad elevada). El lavado puede realizarse utilizando solo una de estas condiciones, por ejemplo, condiciones de rigurosidad elevada, o puede utilizarse cada una de las condiciones, por ejemplo, durante 10-15 minutos cada una, en el orden indicado anteriormente, repitiendo cualquiera o todas las etapas indicadas. Sin embargo, como se mencionó anteriormente, las condiciones óptimas variarán, dependiendo de la reacción de hibridación particular involucrada, y pueden determinarse empíricamente.

Por ejemplo, utilizando un método, tal como reacción en cadena de la polimerasa. (PCR), amplificación por desplazamiento de cadenas, reacción en cadena de la ligasa, tecnología de ciclado de sonda o una placa de micromatriz de ADN, entre otros, se detecta un cambio en la secuencia de una región del genoma o de un producto de expresión de la misma, tal como, por ejemplo, una inserción, una eliminación, una transversión, una transición.

Los métodos de PCR son conocidos en la técnica y se describen, por ejemplo, en Dieffenbach (ed) y Dveksler (ed) (En: PCR Primer: A Laboratory Manual, Cold Spring Harbor Laboratories, NY, 1995). En general, para la PCR, dos moléculas cebadoras de ácido nucleico no complementarias, que comprenden al menos aproximadamente 15 nucleótidos, más preferentemente al menos 20 nucleótidos de longitud, se hibridan con diferentes cadenas de una molécula de ácido nucleico molde, y por vía enzimática, se amplifican copias específicas de la molécula de ácido nucleico molde. Los productos de la PCR pueden detectarse utilizando electroforesis y detección con un marcador detectable que se une a ácidos nucleicos. Como alternativa, uno o más de los oligonucleótidos se marcan con un marcador detectable (por ejemplo, un fluoróforo) y el producto de amplificación se detecta utilizando, por ejemplo, un ciclador de luz (Perkin Elmer, Wellesley, MA, EE. UU.). Claramente, la presente invención también incluye formas cuantitativas de PCR, tales como, por ejemplo, ensayos de Taqman.

La amplificación por desplazamiento de cadenas (SDA, *strand displacement amplification*) utiliza oligonucleótidos, una ADN polimerasa y una endonucleasa de restricción, para amplificar una secuencia diana. Los oligonucleótidos se hibridan con un ácido nucleico diana y la polimerasa se utiliza para producir una copia de esta región. Después, los dúplex de ácido nucleico copiado y ácido nucleico diana se cortan con una endonucleasa que reconoce específicamente una secuencia al comienzo del ácido nucleico copiado. La ADN polimerasa reconoce el ADN cortado y produce otra copia de la región diana al mismo tiempo que desplaza el ácido nucleico generado

previamente. La ventaja de la SDA es que se produce en un formato isotérmico, lo que facilita el análisis automatizado de alto rendimiento.

La reacción en cadena de la ligasa (descrita, por ejemplo, en los documentos EP 320.308 y US 4.883.750) utiliza al menos dos oligonucleótidos que se unen a un ácido nucleico diana de tal manera que son adyacentes. Después, se utiliza una enzima ligasa para ligar los oligonucleótidos. Después, utilizando termociclado, los oligonucleótidos ligados se convierten en una diana para otros oligonucleótidos. Los fragmentos ligados se detectan, por ejemplo, utilizando electroforesis o MALDI-TOF. Como alternativa, o además, una o más de las sondas se marcan con un marcador detectable, lo que facilita la detección rápida.

La tecnología de ciclado de sonda utiliza una sonda sintética quimérica que comprende ADN-ARN-ADN que es capaz de hibridarse con una secuencia diana. Tras la hibridación con una secuencia diana, el dúplex de ARN-ADN formado es una diana para la ARNasa H, por lo que la sonda se escinde. La sonda escindida se detecta después utilizando, por ejemplo, electroforesis o MALDI-TOF.

En la técnica se conocen métodos para detectar SNP y dichos métodos se revisan, por ejemplo, en Landegren et al, Genome Research 8: 769-776, 1998).

Por ejemplo, un SNP que introduce o altera una secuencia que es una secuencia de reconocimiento para una endonucleasa de restricción, se detecta al digerir el ADN con la endonucleasa y el fragmento de interés se detecta utilizando, por ejemplo, transferencia Southern (descrita por Ausubel et al (En: Current Protocols in Molecular Biology. Wiley Interscience, ISBN 047 150338, 1987) y por Sambrook et al (En: Molecular Cloning: Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratories, Nueva York, tercera edición 2001)). Como alternativa, se utiliza un método de amplificación de ácido nucleico descrito anteriormente, para amplificar la región que rodea al SNP. El producto de amplificación se incuba después con la endonucleasa y los fragmentos resultantes se detectan, por ejemplo, mediante electroforesis, MALDI-TOF o PCR.

El análisis directo de la secuencia de polimorfismos de la presente invención puede realizarse utilizando el método didesoxi por terminación de cadena o el método de Maxam-Gilbert (véase Sambrook et al, Molecular Cloning, A Laboratory Manual (2ª ed., CSHP, Nueva York 1989); Zyskind et al, Recombinant DNA Laboratory Manual, (Acad. Press, 1988)). Por ejemplo, una región de ADN genómico que comprende uno o más marcadores se amplifica utilizando una reacción de amplificación, por ejemplo, PCR, y después de la purificación del producto de amplificación, el ácido nucleico amplificado se utiliza en una reacción de secuenciación para determinar la secuencia de uno o ambos alelos en el sitio de un SNP de interés.

Como alternativa, uno o más SNP se detectan utilizando polimorfismo conformacional monocatenario (SSCP, *single stranded conformational polymorphism*). El SSCP se basa en la formación de estructuras secundarias en los ácidos nucleicos y en la naturaleza dependiente de la secuencia de estas estructuras secundarias. En una forma de este análisis, se utiliza un método de amplificación, tal como, por ejemplo, un método descrito anteriormente, para amplificar un ácido nucleico que comprende un SNP. Después, los ácidos nucleicos amplificados se desnaturalizan, se enfrían y se analizan utilizando, por ejemplo, electroforesis en gel de poliacrilamida no desnaturalizante, espectrometría de masas o cromatografía líquida (por ejemplo, HPLC o dHPLC). Las regiones que comprenden diferentes secuencias forman diferentes estructuras secundarias y, como consecuencia, migran a diferentes velocidades a través de, por ejemplo, un gel y/o un campo cargado. Claramente, para facilitar la detección rápida del marcador, en el análisis de SSCP puede incorporarse un marcador detectable en una sonda/cebador útil.

Como alternativa, cualquier cambio de nucleótido puede detectarse utilizando, por ejemplo, espectrometría de masas o electroforesis capilar. Por ejemplo, los productos amplificados de una región de ADN que comprende un SNP de una muestra de ensayo se mezclan con productos amplificados de un individuo que tiene un genotipo conocido en el sitio del SNP. Los productos se desnaturalizan y se les permite volver a emparejarse. Aquellas muestras que comprenden un nucleótido diferente en la posición del SNP no se emparejarán completamente con una molécula de ácido nucleico de la muestra de control, cambiando así la carga y/o la conformación del ácido nucleico, en comparación con un ácido nucleico completamente emparejado. Dicha formación de pares de bases incorrecta es detectable utilizando, por ejemplo, espectrometría de masas.

La PCR específica de alelo (como se describe, por ejemplo, en Liu et al, Genome Research, 7: 389-398, 1997) también es útil para determinar la presencia de uno u otro alelo de un SNP. Se diseña un oligonucleótido, en el cual la base más en el extremo 3' del oligonucleótido, se hibrida con una forma específica de un SNP de interés (es decir, alelo). Durante una reacción de PCR, el extremo 3' del oligonucleótido no se hibrida con una secuencia diana que no comprende la forma particular del SNP detectado. Por consiguiente, se produce poco o ningún producto de PCR, lo que indica que, en la muestra, en el sitio de SNP, está presente una base distinta de la presente en el oligonucleótido. Los productos de la PCR se detectan utilizando, por ejemplo, electroforesis en gel o capilar o espectrometría de masas.

Los métodos de extensión de cebadores (descritos, por ejemplo, en Dieffenbach (ed) y Dveksler (ed) (En: PCR Primer: A Laboratory Manual, Cold Spring Harbour Laboratories, NY, 1995)) también son útiles para la detección de

un SNP. Se utiliza un oligonucleótido que se hibrida con la región de un ácido nucleico adyacente al SNP. Este oligonucleótido se utiliza en un protocolo de extensión con cebador con una polimerasa y un nucleótido difosfato libre que se corresponde con cualquiera o cualquiera de las bases posibles que aparecen en el sitio del SNP. Preferentemente, el nucleótido difosfato está marcado con un marcador detectable (por ejemplo, un fluoróforo). Después de la extensión con cebador, se eliminan los nucleótidos difosfato marcados no unidos, por ejemplo, mediante cromatografía de exclusión por tamaño o electroforesis, o se hidrolizan, utilizando, por ejemplo, fosfatasa alcalina. y se detecta la incorporación del nucleótido marcado en el oligonucleótido, lo que indica que la base está presente en el sitio del SNP. Como alternativa, o además, como se ilustra en el presente documento, los productos de extensión de cebadores se detectan utilizando espectrometría de masas (por ejemplo, MALDI-TOF).

La presente invención se extiende a formas de alto rendimiento de análisis de extensión de cebadores, tales como, por ejemplo, minisecuenciación (Sy Vämen et al., Genomics 9: 341-342, 1995) en donde una sonda o un cebador o múltiples sondas o cebadores se inmovilizan en un soporte sólido (por ejemplo, un portaobjetos de vidrio), una muestra que comprende ácido nucleico se pone en contacto con la(s) sonda(s) o cebador(es), se lleva a cabo una reacción de extensión del cebador en donde cada una de las bases de nucleótidos libres A, C, G , T está marcada con un marcador detectable diferente y la presencia o ausencia de uno o más SNP se determina determinando el marcador detectable unido a cada sonda y/o cebador.

Las moléculas de ácido nucleico bloqueado (LNA) marcadas con fluorescencia o las moléculas de ácido nucleico – proteína (PNA) marcadas con fluorescencia, son útiles para la detección de los SNP (como se describe en Simeonov y Nikiforov, Nucleic Acids Research, 30 (17): 1 -5, 2002) . Las moléculas de LNA y PNA se unen, con alta afinidad, al ácido nucleico, en particular, al ADN. Los fluoróforos (en particular, la rodomina o la hexaclorofluoresceína) conjugados con la sonda de LNA o PNA emiten fluorescencia a un nivel significativamente mayor después de la hibridación de la sonda con el ácido nucleico diana en comparación con una sonda que no se ha hibridado con un ácido nucleico diana. Sin embargo, el nivel de aumento de fluorescencia no se incrementa al mismo nivel incluso cuando se produce una falta de coincidencia de un nucleótido. Por consiguiente, el grado de fluorescencia detectado en una muestra es indicativo de la presencia de una falta de coincidencia entre la sonda de LNA o PNA y el ácido nucleico diana, tal como, en presencia de un SNP. Preferentemente, la tecnología de LNA o PNA marcada con fluorescencia se utiliza para detectar un cambio de una sola base en un ácido nucleico que se ha amplificado previamente utilizando, por ejemplo, un método de amplificación descrito anteriormente.

Como será obvio para el experto en la materia, la tecnología de detección de LNA o PNA es susceptible de una detección de alto rendimiento de uno o más marcadores que inmovilizan una sonda de LNA o PNA a un soporte sólido, como se describe en Oram et al, Clin. Chem. 45: 1898-1905, 1999.

De manera similar, las oligobalizas, son útiles para detectar los SNP directamente en una muestra o en un producto amplificado (véase, por ejemplo, Mhlang y Malmberg, Methods 25: 463-471, 2001). Las oligobalizas son moléculas de ácido nucleico monocatenario con una estructura en forma de tallo y bucle. La estructura en forma de bucle es complementaria a la región que rodea el SNP de interés. La estructura en forma de tallo se forma emparejando dos "brazos" complementarios entre sí a cada lado de la sonda (bucle). Un resto fluorescente está unido a un brazo y a un resto de extinción que suprime cualquier fluorescencia detectable cuando la oligobaliza no está unida a una secuencia diana unida al otro brazo. Tras la unión de la región en forma de bucle con su ácido nucleico diana, los brazos se separan y la fluorescencia es detectable. Sin embargo, incluso un solo emparejamiento de bases erróneo altera significativamente el nivel de fluorescencia detectado en una muestra. Por consiguiente, la presencia o ausencia de una base particular en el sitio de un SNP se determina por el nivel de fluorescencia detectado.

La presente invención incluye otros métodos de detección de un SNP, tales como, por ejemplo, micromatrices de SNP (disponibles en Affymetrix, o descritas, por ejemplo, en los documentos US 6.468.743 o en Hacia et al, Nature Genetics, 14: 441, 1996), ensayos de Taqman (como los descritos en Livak et al, Nature Genetics, 9: 341-342, 1995), minisecuenciación en fase sólida (como los descritos en Syvamen et al, Genomics, 13: 1008-1017, 1992), minisecuenciación con FRET (como los descritos en Chen y Kwok, Nucleic Acids Res. 25: 347-353, 1997) o pirominisecuenciación (según revisión de Landegren et al., Genome Res., 8 (8): 769-776, 1998).

En aquellos casos en los que el polimorfismo o marcador aparece en una región de ácido nucleico que codifica ARN, dicho polimorfismo o marcador se detecta utilizando un método tal como, por ejemplo, RT-PCR, NASBA o TMA.

Los métodos de RT-PCR son conocidos en la técnica y se describen, por ejemplo, en Dieffenbach (ed) y Dveksler (ed) (En: PCR Primer: A Laboratory Manual, Cold Spring Harbor Laboratories, NY, 1995).

Los métodos de TMA o replicación de secuencia autosostenida (3SR) utilizan dos o más oligonucleótidos que flanquean una secuencia diana, una ARN polimerasa, RNasa H y una transcriptasa inversa. Un oligonucleótido (que también comprende un sitio de unión a la ARN polimerasa) se hibrida a una molécula de ARN que comprende la secuencia diana y la transcriptasa inversa produce una copia de ADNc de esta región. La RNasa H se utiliza para digerir el ARN en el complejo ARN-ADN, y el segundo oligonucleótido se utiliza para producir una copia del ADNc. La ARN polimerasa se utiliza después para producir una copia de ARN del ADNc, y el proceso se repite.

Los sistemas NASBA se basan en la actividad simultánea de tres enzimas (una transcriptasa inversa, una RNasa H y una ARN polimerasa) para amplificar selectivamente secuencias de ARNm diana. El molde de ARNm se transcribe a ADNc mediante transcripción inversa utilizando un oligonucleótido que se hibrida con la secuencia diana y comprende un sitio de unión a la ARN polimerasa en su extremo 5'. El molde de ARN se digiere con RNasa H y se sintetiza ADN monocatenario. Después, la ARN polimerasa produce múltiples copias de ARN del ADNc y el proceso se repite.

La hibridación y/o amplificación de un marcador es detectable utilizando, por ejemplo, electroforesis y/o espectrometría de masas. A este respecto, una o más de las sondas/cebadores y/o uno o más de los nucleótidos utilizados en las reacciones de amplificación pueden marcarse con un marcador detectable para facilitar la detección rápida de un marcador, por ejemplo, una etiqueta fluorescente (por ejemplo, Cy5 o Cy3) o un radioisótopo (por ejemplo, ^{32}P).

Como alternativa, la amplificación de un ácido nucleico puede verificarse ininterrumpidamente utilizando un método de análisis de curva de fusión, tal como el que se describe, por ejemplo, en el documento US 6.174.670. Tales métodos son adecuados para determinar el nivel de una forma de corte y empalme alternativa en una muestra biológica.

Los métodos de la invención permiten identificar la aparición de nucleótidos en los SNP utilizando métodos de secuenciación o "microsecuenciación" hologenómica. La secuenciación hologenómica de individuos identifica todos los genotipos de SNP en un solo análisis. Los métodos de microsecuenciación determinan la identidad de un solo nucleótido en un sitio "predeterminado".

Dichos métodos tienen una utilidad particular en la determinación de la presencia e identidad de polimorfismos en un polinucleótido diana. Dichos métodos de microsecuenciación, así como otros métodos para determinar la aparición de nucleótidos en locus de SNP, se analizan en Boyce-Jacino, et al., Patente estadounidense Nº 6.294.336.

Los métodos de microsecuenciación incluyen el método de análisis Genetics Bit descrito por Goelet, P. et al. (WO 92/15712). También se han descrito procedimientos adicionales de incorporación de nucleótidos guiados por cebador para analizar los sitios polimórficos en el ADN (Komher et al, Nucl. Acids. Res. 17, 7779-7784, 1989; Sokolov, Nucl. Acids Res. 18, 3671 (1990); Syvanen et al., Genomics 8, 684-692, 1990; Kuppuswamy et al., Proc. Natl. Acad. Sci. (USA) 88, 1143-1147, 1991; Prezant et al.,Hum. Mutat. 1, 159- 164, 1992; Ugozzoli et al, GATA 9, 107-112, 1992; Nyren et al, Anal Biochem. 208, 171-175, 1993; Wallace, WO89 / 10414; Mundy, Patente estadounidense No. 4.656.127; Cohen et al. , Patente francesa n. ° 2.650.840; documento WO91 / 02087). En respuesta a las dificultades encontradas en el empleo de electroforesis en gel para analizar secuencias, se han desarrollado métodos alternativos para la microsecuenciación, por ejemplo, Macevicz, Patente estadounidense No 5.002.867. Boyce-Jacino et al., patente estadounidense No. 6.294.336, proporcionan un método de secuenciación en fase sólida para determinar la secuencia de moléculas de ácido nucleico (ya sea ADN o ARN) utilizando un cebador que se une selectivamente a un polinucleótido diana en un sitio donde el SNP es el más nucleótido más en el extremo 3' unido selectivamente a la diana. Oliphant et al, Suppl Biotechniques, junio de 2002, describen el uso de la tecnología BeadArray™ para determinar la aparición de nucleótidos de un SNP. Como alternativa, las apariciones de nucleótidos para los SNP pueden determinarse utilizando un sistema DNAMassARRAY (Sequenom, San Diego, California), que combina SpectroChips™, microfluidos, nanodispensación, bioquímica y MALDI-TOF MS (espectrometría de masas con desorción e ionización por láser asistida por matriz-tiempo de vuelo).

Los métodos particularmente útiles incluyen los que pueden adaptarse fácilmente a un formato de alto rendimiento, a un formato multiplex, o a ambos. Los sistemas de alto rendimiento para analizar marcadores, especialmente los SNP, pueden incluir, por ejemplo, una plataforma, tal como la plataforma SNP-IT™ de UHT (Orchid Biosciences, Princeton, NJ, Estados Unidos), el sistema MassArray™ (Sequenom, San Diego, California, USA), el sistema integrado de genotipado de SNP (Illumina, San Diego, Calif, EE. UU.), TaqMan™ (ABI, Foster City, Calif, EE. UU.), amplificación de círculo rodante, polarización fluorescente, entre otros, descritos anteriormente. En general, la plataforma SNP-IT™ es una reacción de extensión de cebador de 3 etapas. En la primera etapa, se aísla un polinucleótido diana de una muestra mediante hibridación con un cebador de captura, que proporciona un primer nivel de especificidad. En una segunda etapa, el cebador de captura se extiende desde un nucleótido trisfosfato de terminación en el sitio SNP diana, lo que proporciona un segundo nivel de especificidad. En una tercera etapa, el nucleótido trisfosfato extendido puede detectarse utilizando una variedad de formatos conocidos, que incluyen: fluorescencia directa, fluorescencia indirecta, un ensayo colorimétrico indirecto, espectrometría de masas, polarización de fluorescencia, etc. Las reacciones pueden procesarse en formato de 384 pocillos en un formato automatizado utilizando un instrumento SNPstream ™ (Orchid BioSciences, Inc., Princeton, NJ).

Sistema de alto rendimiento para selección genotípica

Son ejemplos de sistemas de alto rendimiento los medios de hibridación, por ejemplo, un dispositivo microfluídico o un medio de ensayo homogéneo. Se conocen numerosos dispositivos microfluídicos que incluyen soportes sólidos con microcanales (véanse, por ejemplo, las patentes estadounidenses Nos 5.304.487, 5.110.745, 5.681.484 y 5.593.838). En una realización particularmente preferida, el sistema de alto rendimiento comprende una microplaca

de SNP que comprende 10.000-100.000 oligonucleótidos, cada uno de los cuales consiste en una secuencia que comprende un SNP. Cada uno de estos medios de hibridación es adecuado para determinar la presencia o la ausencia de un marcador asociado con un rasgo.

Los ácidos nucleicos son típicamente oligonucleótidos, unidos directa o indirectamente al soporte sólido. Por consiguiente, los oligonucleótidos se utilizan para determinar la aparición de nucleótidos de un marcador asociado con un rasgo, en virtud de la hibridación del ácido nucleico, del sujeto que se está analizando, con un oligonucleótido de una serie de oligonucleótidos unidos al soporte sólido afectado por la aparición del nucleótido del marcador en cuestión, por ejemplo, por la presencia o ausencia de un SNP en el ácido nucleico del sujeto. Por consiguiente, pueden seleccionarse oligonucleótidos que se unen en o cerca de una ubicación genómica de cada marcador. Dichos oligonucleótidos pueden incluir oligonucleótidos directos e inversos que pueden dar soporte a la amplificación de un marcador polimórfico particular presente en el ácido nucleico molde obtenido del sujeto que se está analizando. Como alternativa, o además, los oligonucleótidos pueden incluir secuencias de cebadores de extensión que se hibridan cerca de un marcador para así dar soporte a la extensión del marcador con fines de identificación. Un método de detección adecuado detectará la unión o el etiquetado de los oligonucleótidos, por ejemplo, en un método de genotipado descrito en el presente documento.

En la materia se han descrito técnicas para producir matrices inmovilizadas de moléculas de ADN. En general, la mayoría de los métodos describen cómo sintetizar matrices de moléculas de ácido nucleico monocatenario, utilizando, por ejemplo, técnicas de enmascaramiento para construir varias permutaciones de secuencias en las diversas posiciones distintas sobre el sustrato sólido. La Patente estadounidense Nº 5.837.832, describe un método mejorado para producir matrices de ADN inmovilizadas en sustratos de silicio basadas en tecnología de integración a gran escala. En particular, la Patente estadounidense Nº 5.837.832 describe una estrategia denominada "mosaico" para sintetizar conjuntos específicos de sondas en ubicaciones definidas espacialmente sobre un sustrato que se utilizan para producir la matriz de ADN inmovilizada. La patente estadounidense Nº 5.837, también proporciona referencias de técnicas anteriores que también pueden utilizarse.

El ADN se puede sintetizar *in situ* sobre la superficie del sustrato. Sin embargo, el ADN también puede imprimirse directamente sobre el sustrato utilizando, por ejemplo, dispositivos robóticos equipados con pines o dispositivos piezoeléctricos. Las micromatrices generalmente se producen por etapas, mediante la síntesis *in situ* de la diana directamente sobre el soporte, o como alternativa, por deposición exógena de dianas previamente preparadas. La fotolitografía, el micromanchado (*microspotting*) mecánico y la tecnología de inyección de tinta, se emplean generalmente para producir micromatrices.

En la fotolitografía, una oblea de vidrio, modificada con grupos protectores fotolábiles, se activa selectivamente, por ejemplo, para la síntesis de ADN, mediante centelleo lumínico a través de una fotomáscara. Los repetidos ciclos de desprotección y acoplamiento permiten la preparación de micromatrices de oligonucleótidos de alta densidad (véase, por ejemplo, la patente estadounidense n.º 5.744.305, publicada el 28 de abril de 1998).

El micromanchado incluye tecnologías de deposición que permiten la producción automatizada de micromatrices, a través de la impresión de pequeñas cantidades de sustancias diana prefabricadas en superficies sólidas. La impresión se realiza mediante el contacto directo de la superficie entre el sustrato de impresión y un mecanismo de suministro, tal como un alfiler o un capilar. Los sistemas de control robótico y los cabezales de impresión multiplexados permiten la fabricación automatizada de micromatrices.

Las tecnologías de inyección de tinta utilizan elementos piezoeléctricos y otras formas de propulsión para transferir sustancias bioquímicas de boquillas en miniatura a superficies sólidas. Usando piezoelectricidad, la muestra diana se expulsa al pasar una corriente eléctrica a través de un cristal piezoeléctrico que se expande para expulsar la muestra. Las tecnologías de propulsión piezoeléctrica incluyen dispositivos continuos y a demanda. Además de los chorros de tinta piezoeléctricos, se puede utilizar calor para formar y propulsar gotas de fluido utilizando chorros de burbujas o cabezales de chorros de tinta térmica; sin embargo, normalmente, dichos chorros de tinta térmica no son adecuados para la transferencia de materiales biológicos debido al calor que a menudo es estresante en muestras biológicas. Como ejemplo del uso de la tecnología de inyección de tinta se incluye la patente estadounidense No 5.658.802 (expedida el 19 de agosto de 1997).

Una pluralidad de ácidos nucleicos está típicamente inmovilizada sobre o en regiones distintas de un sustrato sólido. El sustrato es poroso para permitir la inmovilización dentro del sustrato, o es sustancialmente no poroso para permitir la inmovilización de la superficie.

El sustrato sólido puede estar fabricado de cualquier material al que se puedan unir los polipéptidos, ya sea directa o indirectamente. Los ejemplos de sustratos sólidos adecuados incluyen vidrio plano, obleas de silicio, mica, cerámica y polímeros orgánicos tales como plásticos, incluyendo poliestireno y polimetacrilato. También es posible utilizar membranas semipermeables, tales como membranas de nitrocelulosa o de nailon, que están ampliamente disponibles. Las membranas semipermeables se instalan en una superficie sólida más fuerte, como el vidrio. Las superficies pueden estar recubiertas opcionalmente con una capa de metal, tal como oro, platino u otro metal de transición.

Preferentemente, el sustrato sólido es generalmente un material que tiene una superficie rígida o semirrígida. En realizaciones preferidas, al menos una superficie del sustrato será sustancialmente plana, aunque en algunas realizaciones es deseable separar físicamente las regiones de síntesis para diferentes polímeros, por ejemplo, con regiones elevadas o fosas grabadas. También se prefiere que el sustrato sólido sea adecuado para la aplicación de secuencias de ADN de alta densidad en zonas distintas normalmente de 50 a 100 µm, lo que da una densidad de 10.000 a 40.000 cm-2.

El sustrato sólido se divide convenientemente en secciones. Esto se logra mediante técnicas tales como el fotograbado o la aplicación de tintas hidrófobas, por ejemplo, tintas con teflón (Cel-line, EE. UU.).

Las distintas posiciones, en las que se ubica cada miembro diferente de la matriz, pueden tener cualquier forma conveniente, por ejemplo, circular, rectangular, elíptica, en forma de cuña, etc.

La unión de los ácidos nucleicos al sustrato puede ser covalente o no covalente, generalmente a través de una capa de moléculas a la que se unen los ácidos nucleicos. Por ejemplo, las sondas/cebadores de ácido nucleico se pueden marcar con biotina y el sustrato se puede recubrir con avidina y/o estreptavidina. Una característica conveniente del uso de sondas/cebadores biotinilados es que la eficacia del acoplamiento al sustrato sólido se determina fácilmente.

Se puede proporcionar una interfaz química entre el sustrato sólido, por ejemplo, en el caso del vidrio, y las sondas/cebadores. Como ejemplos de interfaces químicas adecuadas se incluyen hexaetilenglicol, polilisina. Por ejemplo, la polilisina puede modificarse químicamente utilizando procedimientos estándar para introducir un ligando de afinidad.

Otros métodos para unir las sondas/cebadores a la superficie de un sustrato sólido incluyen el uso de agentes de acoplamiento conocidos en la técnica, por ejemplo, como se describe en el documento WO98/49557.

El sistema de alto rendimiento está diseñado para determinar las apariciones de nucleótidos de un SNP o una serie de SNP. Los sistemas pueden determinar las apariciones de nucleótidos de un mapa de SNP de alta densidad hologenómico.

Los sistemas de alto rendimiento para analizar marcadores, especialmente los SNP, pueden incluir, por ejemplo, una plataforma, tal como la plataforma UHT SNP-IT (Orchid Biosciences, Princeton, NJ, EE. UU.) Sistema MassArray™ (Sequenom, San Diego, California, EE. UU.)), el sistema integrado de genotipado SNP (Illumina, San Diego, California, EE. UU.), TaqMan™ (ABI, Foster City, California, EE. UU.). Son ejemplos de matrices de ácidos nucleicos del tipo descrito los del documento WO 95/11995 donde también se describen sub-matrices optimizadoas para la detección de una forma variante de un polimorfismo pre-caracterizado. Dicha sub-matriz contiene sondas diseñadas para ser complementarias a una segunda secuencia de referencia, que es una variante alélica de la primera secuencia de referencia. La inclusión de un segundo grupo (o grupos adicionales) puede ser particularmente útil para analizar subsecuencias cortas de una secuencia de referencia primaria en la que se espera que se produzcan múltiples mutaciones dentro de una distancia corta proporcional a la longitud de las sondas (por ejemplo, dos o más mutaciones dentro de 9 a 21 bases). Más preferentemente, el sistema de alto rendimiento comprende una micromatriz de SNP como las disponibles en Affymetrix o las descritas, por ejemplo, en el documento US 6.468.743 o en Hacia et al, Nature Genetics, 14: 441, 1996.

Las matrices de ADN se suelen leer al mismo tiempo mediante una cámara de dispositivo acoplado cargado (CCD) o un sistema de imágenes confocal. Como alternativa, la matriz de ADN puede colocarse para la detección en un aparato adecuado que pueda moverse en una dirección x-y, tal como un lector de placas. De esta manera, el cambio en las características de cada posición distinta se mide automáticamente mediante el movimiento controlado informatizado de la matriz para colocar a su vez cada elemento distinto en línea con los medios de detección.

Los medios de detección pueden consultar cada posición en la matriz de la biblioteca de forma óptica o eléctrica. Ejemplos de medios de detección adecuados incluyen cámaras CCD o sistemas de formación de imágenes confocales.

El sistema puede incluir además un mecanismo de detección para detectar la unión de la serie de oligonucleótidos a la serie de SNP. Dichos mecanismos de detección son conocidos en la técnica.
El sistema de alto rendimiento puede incluir un mecanismo de tratamiento de reactivos que pueden utilizarse para aplicar un reactivo, normalmente un líquido, al soporte sólido.

El sistema de alto rendimiento también puede incluir un mecanismo eficaz para mover un soporte sólido y un mecanismo de detección.

Estimación del valor de reproducción

Para estimar el valor de reproducción en el método de la presente invención, se utiliza cualquiera de una serie de métodos estadísticos, preferentemente utilizando medios informáticos, incluyendo estrategias de remuestreo, por

ejemplo, pruebas de asignación al azar y remuestreo por reposición, que permiten la construcción de intervalos de confianza y pruebas de significación apropiadas, por ejemplo, Best Liner Unbiased Predictors (BLUP; Henderson en: "Applications of Linear Models in Animal Breeding", University of Guelph, Guelph, Ontario, Canada; Lynch y Walsh, En: "Genetics and Analysis of Quantitative Traits", Sunuaer Associates, Sunderland MA, USA, 1998); la estrategia de Monte Carlo de Cadenas de Markov (MCMC) (Geyer et al, Stat. Sci. 7, 73-511, 1992; Tierney et al., Ann. Statist. 22, 1701-1762, 1994; Tanner et al., En: "Tools for Statistical Analysis", Springer-Verlag, Berlin/Nueva York, 1996); el muestreador de Gibbs (Geman et al., IEEE Trans. Pattern Anal. Mach. Intell. 6, 721-741, 1984); la distribución posterior bayesiana (p. ej., Smith et al, J. Royal Statist. Soc. Ser. B55, 3-23, 1993). Dichos métodos son muy conocidos por los expertos en la materia.

Preferentemente, los EBV (e*xpected breeding value,* valores de reproducción esperados) se calculan utilizando un método denominado "Bayes2" de Meuwissen et al. .Genetics 157, 1819-1829 (2001). El método de Bayes 2 permite que algunos segmentos cromosómicos tengan un efecto más grande sobre el rasgo que otros. El modelo estadístico también podría ajustarse al efecto de cada posición en el genoma utilizando, por ejemplo, BLUP para calcular el efecto de cualquier alelo QTL presente en esa posición en todos los gametos representados en la población. Como alternativa, la relación promedio entre los animales puede estimarse a partir de los alelos marcadores que se ha inferido que llevan, posiblemente ponderando cada posición en el genoma por su importancia en el control del rasgo. Esto supone que cada segmento cromosómico procede de un antepasado o fundador clave dentro de una recombinación mínima o ninguna dentro del segmento, una suposición que se mantiene cuando el número de generaciones entre el antepasado o fundador y el individuo de interés es bajo, es decir, menor de aproximadamente 10 generaciones. Por ejemplo, la matriz puede ser una matriz idéntica por descendencia (IBD, *identical-by-descent*) cuyos elementos gis son la esperanza de la cantidad de segmentos cromosómicos transportados por j individual que son IBD con un alelo muestreado aleatoriamente de i individual, condicional a la información de árbol genealógico y a los datos del marcador. Las matrices IBD pueden calcularse para diferentes segmentos cromosómicos, por ejemplo, separados a lo largo del genoma. Las matrices de IBD también pueden promediarse a través de posiciones y cromosomas. Para calcular una matriz IBD pueden utilizarse diferentes números de segmentos cromosómicos. La precisión de la evaluación puede calcularse como la correlación entre los valores de reproducción verdaderos y estimados.

Para calcular el EBV a partir de marcadores de ADN hologenómico, es conveniente considerar que el proceso comprende tres etapas:

1. Utilizar los marcadores para deducir el genotipo de cada animal en cada QTL;
2. Estimar el efecto de cada genotipo QTL en el rasgo; y
3. Sumar los efectos QTL para los candidatos de selección para obtener su EBV genómico (GEBV, *genomic* e*xpected breeding value*).

Estas etapas se describen con más detalle en los siguientes párrafos.

Uso de marcadores para deducir el genotipo de cada animal en cada QTL

El método más sencillo para deducir genotipos de QTL es tratar los marcadores como si fueran QTL y estimar los efectos de los alelos o genotipos marcadores. El parámetro clave aquí es la proporción de la variación QTL definida por los marcadores (r^2). Esto depende del LD entre el QTL y un marcador o una combinación lineal de marcadores. La extensión de LD y por lo tanto de r^2 es muy variable. El promedio de r^2 disminuye a medida que aumenta la distancia entre los dos locus. Por ejemplo, en ganado Holstein el promedio de r^2 cuando los locus están separados por 50 kb es de 0,35. Para obtener una separación promedio de 50 kb se requieren 60.000 marcadores separados de manera uniforme. Como es poco probable que los marcadores estén separados de manera uniforme, y debido a la naturaleza variable del LD, aún no podemos esperar que todos los QTL tengan un SNP en LD completo con ellos. Esto sugiere que se necesitan marcadores más densos que los disponibles actualmente. Se dispone de tecnología para conseguir esto (por ejemplo, Parks et al. Nature Genet, publicación en línea del 6 de junio de 2007).

Una alternativa al uso de genotipos de un marcador individual es construir haplotipos basándose en diversos marcadores. Un QTL que no esté en LD completo con cualquier marcador individual puede estar en LD completo con un haplotipo de marcador múltiple. Por ejemplo, utilizando 9323 genotipos SNP de ganado Angus, y considerando un SNP elegido al azar como un sustituto para un QTL, la proporción de variación explicada por un haplotipo de marcadores circundantes puede aumentar de 0,2 para el marcador más cercano a 0,58 para un marcador de haplotipo 6. El uso de genotipos de marcadores múltiples, pero sin deducir haplotipos, por ejemplo, con regresión de marcadores múltiples, generalmente estará entre estos dos límites. Normalmente, hay muchos haplotipos presentes en una población y, por lo tanto, se reduce la cantidad de datos con que estimar el efecto de cada uno y esto reducirá la precisión con la que se estima cada efecto de haplotipo. Sin embargo, el aumento en la variación de QTL explicado por el uso de haplotipos marcadores compensa la disminución en la precisión de la estimación de un mayor número de efectos de haplotipos, de modo que los haplotipos predicen el efecto de los alelos QTL con mayor precisión que un solo marcador. La ventaja de los haplotipos sobre los marcadores individuales disminuye a medida que aumenta r^2 entre los marcadores adyacentes. A un valor de r^2= 0,215 entre marcadores adyacentes, la estrategia de haplotipo y de marcador único proporcionan precisiones muy similares.

A medida que aumenta el número total de animales con fenotipos y genotipos de marcadores, la precisión de la estimación de los efectos del genotipo marcador se aproximará a 1,0 y, por lo tanto, la precisión de la estimación de los efectos de los haplotipos. Pero la precisión para el haplotipo se acercará a 1,0 1.0 más lentamente que la precisión de la estimación de los efectos de SNP porque hay más de 2 efectos de haplotipos por QTL a estimar. Por lo tanto, la ventaja de los haplotipos sobre los marcadores individuales aumenta a medida que aumenta la cantidad de datos para la estimación, especialmente a densidades de marcadores más bajas. La precisión de utilizar marcadores individuales puede ser mayor que la de utilizar haplotipos de marcadores si hay un número limitado de registros fenotípicos para estimar los efectos y el nivel de LD entre los marcadores individuales y los QTL es muy alto.

Una alternativa al tratamiento de un haplotipo de marcadores como si fuera un alelo QTL es tratar cada gameto como portador de un alelo QTL diferente, pero estimar la correlación entre los efectos de estos alelos en función de los marcadores circundantes. Un análisis de ligamiento rastrea los alelos QTL a través del árbol genealógico conocido utilizando los marcadores y calcula la probabilidad de que dos alelos sean idénticos por descendencia (IBD) de un antepasado o fundador común dentro del árbol genealógico. La probabilidad de que dos alelos QTL sean IBD debido a un antepasado o fundador común fuera del árbol genealógico puede evaluarse a partir de la similitud de los alelos marcadores que rodean al QTL suponiendo un modelo evolutivo para el desequilibrio de ligamiento entre los marcadores y los QTL. El análisis de ligamiento y el análisis de LD se pueden combinar para estimar una matriz de probabilidades de IBD entre todos los alelos QTL y esto puede utilizarse para estimar los efectos de todos los alelos QTL. Los errores en el posicionamiento de los marcadores en el genoma reducirán la precisión de inferir haplotipos y, por lo tanto, la precisión de los GEBV resultantes de los enfoques de haplotipos e IBD.

A bajas densidades de marcadores (por ejemplo, r^2 entre marcadores adyacentes menor de 0,2), se prefiere la estrategia de IBD sobre la estrategia de haplotipo o la estrategia de marcador individual. A altas densidades de marcadores, los tres métodos proporcionan aproximadamente la misma precisión.

Estimación del efecto de cada genotipo QTL en el rasgo

La ganancia genética es mayor si la estimación del valor de reproducción (g) tiene la propiedad GEBV = E (g | “datos”). Dado que el EBV se calcula sumando los efectos estimados de todos los QTL (u), la propiedad deseada para el EBV se logra al estimar cada efecto QTL de la siguiente manera:

$$\hat{\mathbf{u}} = E(\mathbf{u} \mid \text{"datos"}),$$

en donde el estimador apropiado es:

$$\hat{\mathbf{u}} = \frac{\int \mathbf{u} * p(\text{datos} \mid \mathbf{u}) p(\mathbf{u}) du}{\int p(\text{datos} \mid \mathbf{u}) p(\mathbf{u}) du}$$

en donde:

1. p(datos|u) es una probabilidad; y
2. p(u) es una distribución previa de los efectos de QTL.

Por consiguiente, el estimador de los efectos de QTL depende de la distribución previa de los efectos de QTL. Dado que normalmente se analiza un QTL en muchas posiciones (por ejemplo, 10.000 SNP), se espera que no haya QTL en la mayoría de esas posiciones. Por lo tanto, la distribución previa p(u) debe tener una alta probabilidad de p(0) para el rasgo en cuestión. Por ejemplo, se estima que los rasgos de producción de leche en el ganado lechero están determinados por al menos aproximadamente 150 QTL y se estima que la distribución de sus efectos es aproximadamente exponencial.

Como alternativa, se emplea un método de mínimos cuadrados para estimar el efecto de cada QTL en un rasgo. Las estimaciones de los mínimos cuadrados corresponden a suponer una distribución previa de los efectos QTL con una variación infinitamente grande. Utilizando los mínimos cuadrados, solo se detectará y utilizará un QTL con un gran efecto, y por lo tanto los marcadores no capturarán toda la variación genética. Al suponer que los efectos QTL se obtienen de la distribución normal con una variación constante entre los segmentos cromosómicos, se obtiene una estimación de BLUP en la que todos los efectos se estiman simultáneamente, por lo que se obtienen estimaciones que se correlacionan mejor con los BV verdaderos. Sin embargo, se prefiere un análisis Bayesiano que utiliza una distribución previa más apropiada de los efectos de QTL. Para situaciones en las que la mayoría de los 'QTL' tienen

efecto cero, los mínimos cuadrados y el resultado de BLUP en estos cero se estima que los efectos son pequeños pero no nulos y su efecto acumulativo agrega ruido a las estimaciones.

Se obtienen mejores estimaciones cuando se estima que muchos QTL posibles tienen un efecto cero o, equivalentemente, se excluyen del modelo. Si todos los efectos de QTL eran de una distribución exponencial reflejada (es decir, sin peso adicional en cero), se prefiere un estimador denominado LASSO (Tibshirani et al., J. Royal Stat. Soc. Ser. B 58, 267-288, 1996) . Sin embargo, en la situación en la que muchos efectos verdaderos son cero, LASSO aún estima demasiados efectos distintos de cero. Una alternativa pragmática es excluir del modelo todos los efectos excepto los más altamente significativos, por ejemplo, estableciendo un umbral de significación de tal manera que se espera que solo un positivo falso por genoma proporcione un EBV altamente correlacionado con BV. Sin embargo, si los efectos de estos QTL significativos se estiman por mínimos cuadrados, los efectos aún se sobreestimarán y pueden requerir corrección utilizando validación cruzada, por ejemplo, como describen Whittaker et al., Genet. Res. 69, 137-144, (1997). Esto implica estimar los efectos en dos partes independientes de los datos y calcular la regresión de un conjunto de soluciones en el otro. Después, las soluciones se retroregresan por este coeficiente de regresión para dar estimaciones imparciales. La validación cruzada también puede utilizarse para elegir entre modelos de la competencia. Dentro de un conjunto de datos, agregar QTL adicionales aumenta la precisión de la predicción, pero la precisión de los GEBV en un conjunto de datos independiente se puede usar para juzgar si la precisión realmente ha aumentado. y calculando la regresión de un conjunto de soluciones sobre el otro. Las soluciones son luego regresadas por este coeficiente de regresión para dar estimaciones imparciales. La validación cruzada también se puede utilizar para elegir entre modelos de competencia. Dentro de un conjunto de datos, la adición de extra QTL aumenta la precisión de la predicción, pero la precisión de los GEBV en un conjunto de datos independiente puede utilizarse para juzgar si la precisión ha aumentado realmente.

Preferentemente, se emplea un explícito anterior en el que se supone que los efectos de QTL se extraen de una distribución normal pero que la variación de esa distribución varía entre los QTL, y que la distribución de variaciones sigue una distribución de chi-cuadrado invertida. Una ventaja de utilizar el explícito anterior es que las estimaciones de los QTL más grandes o más significativos no están sobreestimadas. Esto significa que los efectos pueden estimarse a partir de todos los datos disponibles, independientemente de si los datos formaban parte de los utilizados para descubrir el QTL o no. Esto proporciona una ventaja importante a medida que la selección genómica se implementa en la industria y se vuelve imposible distinguir claramente entre datos de descubrimiento (donde las estimaciones de mínimos cuadrados están sesgadas) e independientes, datos de validación (donde son imparciales).

En todos los métodos anteriores para estimar el BV, se puede añadir un término poligénico al modelo para explicar la variación genética no explicada por los marcadores. Cuando se prueba la importancia de un marcador a la vez, la omisión del término poligénico del modelo produce aproximadamente el doble de positivos falsos que se indican por el umbral de significación. Esto se debe a que, dentro de un conjunto de datos, todos los marcadores y QTL están correlacionados a través de la relación de árbol genealógico entre los animales. En consecuencia, cualquier marcador puede, por casualidad, correlacionarse con un QTL a cierta distancia o incluso en otro cromosoma, por lo que parece tener un efecto que en realidad es un artefacto de la estructura de árbol genealógico. Incluso cuando todos los QTL se ajustan simultáneamente, puede ser conveniente ajustar un efecto poligénico, ya que esto capturará, hasta cierto punto, aquellos QTL no asociados con marcadores o haplotipos a altos niveles de r^2.

Preferentemente, se emplean grandes cantidades de animales con genotipos marcadores y fenotipos para estimar los efectos de QTL, con preferencia aproximadamente de 2000 registros o un número mayor.

La presente invención se describe adicionalmente con referencia al siguiente ejemplo no limitativo.

Ejemplo 1

Método modelo de selección artificial para una población de ganado Holstein

Razonamiento

Muchas razas de ganado bovino tienen un tamaño de población eficaz pequeño de 50 a 100, incluida la población de ganado Holstein. Esto significa que la mayoría de los segmentos cromosómicos encontrados en animales de la generación actual se remontan a uno de menos de aproximadamente 100 antepasados clave en unas pocas generaciones. Este corto tiempo de coalescencia significa que los segmentos cromosómicos son grandes y podrían reconocerse por su haplotipo en un grupo de marcadores. Por consiguiente, puede realizarse genotipado informático de la siguiente manera:

1. Genotipar antepasados clave para determinar un conjunto de marcadores densos;
2. Genotipar individuos de la población/generación actual para obtener marcadores suficientes que permitan que los segmentos cromosómicos se correspondan con los segmentos portados por los antepasados clave; y
3. Inferir genotipos de individuos en la población/generación actual para que sean iguales a los del antepasado o fundador clave del segmento cromosómico correspondiente.

Por este medio, es posible genotipar grandes cantidades de individuos en la población actual para un número moderado de marcadores, aunque obteniendo genotipos informáticos para un gran número de marcadores. A medida que disminuye el coste de la (re)secuenciación del genoma, este genotipo informático es extensible a la secuenciación informático. Es decir, los antepasados clave se secuencian y después se imputa la secuencia completa del genoma para los segmentos cromosómicos en los animales actuales que se corresponden con el segmento cromosómico en antepasados clave.

El genotipado y la secuenciación *informáticos* de la presente invención utiliza las relaciones conocidas entre individuos de la población actual y los antepasados clave, reduciendo así el número de marcadores que deben ser genotipados en los individuos de la población actual.

Este método reduce el coste de la selección genómica al reducir el número de marcadores que se necesita categorizar en los candidatos de selección. El método también identifica los polimorfismos causales subyacentes de QTL. Si cada QTL se ataca por separado, la secuenciación del genoma se dirige a una región particular. Sin embargo, dado que hay muchos QTL que afectan a muchos rasgos de interés, esta es una estrategia muy poco eficaz y, por lo tanto, es deseable realizar una (re)secuenciación completa del genoma y buscar muchos QTL simultáneamente. Realizando los métodos informáticos de la presente invención, será posible inferir las secuencias de miles de genomas en 5 años, por ejemplo, genotipando una gran muestra de animales que se hayan registrado con respecto a muchos rasgos de interés (por ejemplo, rasgos de enfermedad) para un número moderado de marcadores, secuenciando los genomas de los antepasados clave, rastreando los segmentos cromosómicos de los individuos de la población actual hasta sus antepasados, infiriendo la secuencia completa del genoma en cada animal y realizando un análisis de secuencia hologenómico (GWAS, *genome-wide análisis of sequence*) basado en la secuencia inferida completa del genoma. Partiendo de esta base, el método de la presente invención es útil para identificar grandes cantidades de mutaciones que afectan a la susceptibilidad de enfermedades u otros rasgos.

Métodos y resultados

La presente divulgación proporciona un método de selección artificial que comprende:

1. Identificar el conjunto mínimo de antepasados clave que representa la mayoría de los segmentos cromosómicos en una población actual;
2. Genotipar los antepasados clave para un conjunto de marcadores densos;
3. Genotipar uno o más individuos de una población actual con marcadores suficientes para permitir que los segmentos cromosómicos se correspondan con los segmentos portados por los antepasados clave;
4. Rastrear los segmentos cromosómicos de uno o más individuos de la población actual de ganado hasta un antepasado clave;
5. Inferir los genotipos de marcadores dentro de uno o más segmentos cromosómicos del uno o más individuos en la población actual para que sean los mismos que los del antepasado clave; y
6. Utilizar el genotipo inferido del uno o más individuos en la población actual para estimar el valor de reproducción de dicho uno o más individuos.

Estas etapas se describen con más detalle con referencia a la selección artificial de ganado Holstein.

1 Identificación del conjunto mínimo de antepasados clave que representa la mayoría de los segmentos cromosómicos en una población actual

Los antepasados clave de una población de ganado Holstein se determinan, por ejemplo, basándose en árboles genealógicos conocidos y/o estimando las relaciones entre los animales a partir del análisis del ligamiento del marcador de ADN. La estimación de antepasados clave basada en marcadores de ADN proporciona una estimación más precisa de las relaciones entre los animales que el árbol genealógico 'conocido', que a menudo está incompleto y contiene errores. Preferentemente, para identificar a los antepasados clave, se utiliza una matriz A estimada a partir de los marcadores de ADN, y/o la matriz A procedente de árbol genealógico.

Utilizando la matriz de relación aditiva (A) descrita anteriormente en el presente documento, se obtuvieron los antepasados clave proporcionados en las Tablas 2 y 3 para una población de 2300 vacas Holstein.

2. Genotipado de los antepasados clave para un conjunto de marcadores densos

Como no teníamos genotipos sobre los antepasados clave de la población de ganado Holstein en estudio, modelamos la población de antepasados utilizando una población fundadora de 425 animales sin padre ni madre conocidos. En la práctica, se necesitarían menos antepasados clave que estos porque existen árboles genealógicos más completos o las relaciones se deducen más fácilmente a partir de los datos del marcador.

Después, inferimos los genotipos de los fundadores basados en genotipos de sus parientes para 11 marcadores microsatélite que abarcaban 2,9 cM del cromosoma 21, utilizando un método de Monte Carlo de Cadenas de Markov (MCMC) descrito originalmente por Schafer, JL (1997), Analysis of Incomplete Multivariate Data, Nueva York:

Chapman y Hall. En esta estrategia, el genotipo de cada animal se muestrea desde una distribución posterior condicional a los genotipos de padres, abuelos, descendientes, grandes descendientes y compañeros. Se utilizó una muestra de la distribución posterior de genotipos y los genotipos inferidos en los animales fundadores se consideraron como los genotipos reales.

Preferentemente, el conjunto de datos de genotipos de los antepasados enumerados en la Tabla 2 o 3 procedería de un conjunto de marcadores densos mediante estrategias estándar de genotipado que utilizan el semen como fuente de material genético para el genotipado y/o basándose en datos genotípicos disponibles, y opcionalmente combinados con un método MCMC para inferir o imputar los valores ausentes.

3. Genotipado de uno o más individuos de una población actual con marcadores suficientes para permitir que los segmentos cromosómicos se correspondan con los segmentos portados por los antepasados clave

Utilizamos un conjunto de datos de 2300 vacas Holstein que se habían genotipado para hasta 11 marcadores microsatélite que abarcaban 2,9 cM del cromosoma 21. Diez de estos marcadores se trataron como los marcadores que se habían genotipado en la población actual y en los fundadores. El marcador restante (11º) se trató como un marcador que se había genotipado solo en la población fundadora, junto con otros marcadores desconocidos que se genotiparon solo en fundadores. Razonamos que esto era suficiente para probar rigurosamente el método, porque este 11º marcador tiene cinco (5) alelos casi igualmente comunes y, como consecuencia, representó un marcador difícil en el que predecir un genotipo. Los genotipos en otros marcadores, como los SNP, serían más fáciles de predecir que los del ejemplo proporcionado en el presente documento.

Para probar la precisión del método en este sistema modelo, definimos un subconjunto de "candidatos de selección" de la población actual como aquellos animales que no tienen descendencia pero que tienen un padre y una madre conocidos. Los candidatos de selección se habían genotipado en el marcador 11 en la región 2.9cM del cromosoma 21, sin embargo, el genotipo marcador conocido fue enmascarado u ocultado del análisis, de modo que se supuso que los candidatos de selección eran conocidos por un máximo de solo 10 marcadores microsatélite en esta región del cromosoma. Dicho análisis con ocultación de los marcadores para los candidatos de selección, se realizó para permitirnos comparar el genotipo del marcador verdadero con el genotipo del marcador esperado por el análisis, es decir, cuando se ocultó el valor ausente.

4. Rastrear los segmentos cromosómicos del uno o más individuos de la población actual de ganado hasta un antepasado clave

En el sistema modelo, debido a que se empleó un segmento específico del cromosoma 21 para modelar la precisión del método, no fue estrictamente necesario rastrear los segmentos cromosómicos de los candidatos seleccionados a un antepasado clave.

A pesar de esta limitación, comparamos los genotipos de los candidatos de selección dentro de la región de 2,9 cM del cromosoma 21 para un subconjunto limitado de marcadores para los cuales habían sido genotipos con los genotipos inferidos de los fundadores en la misma región de cromosoma y alineamos los marcadores para por lo tanto, rastrear los segmentos cromosómicos de uno o más individuos de la población actual de ganado hasta un fundador en particular.

Los segmentos cromosómicos se rastrearon a través del árbol genealógico desde los candidatos seleccionados hasta los fundadores basándose en los 10 marcadores que utilizan el mismo programa MCMC que el anterior.

5. Inferir los genotipos de marcadores dentro de uno o más segmentos cromosómicos de uno o más individuos en la población actual para que sean los mismos que los del antepasado clave.

El programa MCMC infiere el genotipo oculto, es decir, ausente, en los candidatos seleccionados porque rastrea el origen de cada posición del cromosoma en un candidato de selección hasta uno de los fundadores en el que se conocía el genotipo del marcador 11. En el 96 % de los casos, el genotipo esperado coincidió con el genotipo verdadero. Los datos están disponibles según petición.

En este ejemplo, los genotipos marcadores de los fundadores se infirieron realmente a partir de genotipos de familiares, pero idealmente eran conocidos. En este ejemplo, se infirieron de familiares porque no se disponía del ADN para el genotipado de los fundadores. Por lo tanto, probamos el método en condiciones desfavorables y concluimos que en condiciones más favorables donde se conocen los genotipos de los fundadores o antepasados, el método ofrecería resultados que son iguales a, o mejores que, los del presente ejemplo.

Se podrían emplear diversos métodos analíticos distintos para inferir los genotipos ausentes de los candidatos de selección, por ejemplo, un algoritmo de despeje basado en árbol genealógico que incluye despeje iterativo múltiple como el descrito por Meuwissen et al, Genetics 161, 373-379, 2002).

Como alternativa, o además, un método diseñado para animales no relacionados, por ejemplo, utilizando el algoritmo fastPHASE disponible en la Universidad de Washington, Ann Arbor, MI 48109-2029, EE. UU. El algoritmo fastPHASE implementa métodos para estimar haplotipos y genotipos ausentes a partir de datos de genotipos de SNP de la población. Cuando se utiliza en animales relacionados, fastPHASE reconoce haplotipos comunes procedentes de antepasados clave. Por ejemplo, también hemos probado la precisión de fastPHASE en el método de la invención, utilizando el mismo conjunto de datos anterior. Analizamos un conjunto de datos de 680 animales para el cual se conocieron 6 o más de los 11 genotipos marcadores. La mitad de estos animales se utilizó como un conjunto experimental y la mitad fueron candidatos de selección. El genotipo en el marcador 5 se ocultó del análisis fastPHASE. En esta variación, utilizando fastPHASE, se predijo correctamente el 91 % de los genotipos ausentes.

La presente divulgación también incluye el uso de dos etapas para inferir genotipos en candidatos de selección a partir de aquellos en antepasados clave. Por ejemplo, pueden genotiparse 100 antepasados clave para todos los marcadores conocidos (por ejemplo, 1.000.000 marcadores), o pueden secuenciarse por completo. Todos los machos utilizados para la reproducción o todos los machos sementales utilizados para la reproducción, pueden genotiparse individualmente para un subconjunto de los marcadores conocidos, por ejemplo, 50.000 marcadores, y los candidatos de selección pueden genotiparse para algunos marcadores, por ejemplo, solo 2000 marcadores. Los segmentos cromosómicos en los candidatos de selección pueden rastrearse durante una o unas pocas generaciones hasta los machos reproductores y pueden rastrearse hasta los antepasados clave. Esto hace uso del genotipado de alto rendimiento (50.000 marcadores) en una pequeña fracción de la población total.

6. Uso del genotipo inferido de uno o más individuos en la población actual para estimar el valor de reproducción de dicho uno o más individuos.

Los métodos estándar como se describen en este documento se utilizan para predecir el valor de reproducción del candidato de selección a partir de los genotipos inferidos de los marcadores. Todos ellos utilizan una ecuación que predice el BV a partir de genotipos marcadores que proceden del análisis de una muestra de animales que tienen genotipos y valores de reproducción estimados (EBV) o registros fenotípicos. Un método preferido descrito en el presente documento calcula el valor esperado del BV condicional en los genotipos marcadores y en una distribución previa de los efectos de los genes en el rasgo de interés. Los métodos para estimar esta distribución previa están disponibles para el público.

Conclusiones

El modelo descrito en el presente documento para una población de ganado Holstein se extrapola fácilmente y es aplicable a estudios de hologenómica, empleando antepasados en lugar de fundadores. En el método, un grupo de antepasados clave sería genotipado para muchos marcadores; los candidatos de selección serían genotipados para un número más pequeño de marcadores; los segmentos cromosómicos de los candidatos de selección se rastrearían hasta los de los antepasados clave y esto permitiría que todos los genotipos marcadores conocidos en los antepasados clave sean imputados a los candidatos de selección. En los ejemplos proporcionados en este documento, tratamos a todos los animales con genotipos conocidos de hasta 11 marcadores como candidatos de selección, y después rastreamos el árbol genealógico conocido de estos animales en la medida de lo posible para identificar a 425 fundadores que no tenían dos padres conocidos. Estos 425 fundadores son verdaderamente representativos de antepasados clave en este ejemplo. Este es un número mayor de antepasados clave de lo normal, sin embargo, el tamaño más grande de la población fundadora es consecuencia de datos de genealogía incompletos y datos de genotipo deficientes para los antepasados mostrados en las Tablas 2 y 3. Diez de los marcadores se consideraron equivalentes a los pequeños números de marcadores clasificados en los candidatos de selección. Uno de esos marcadores se trató como un ejemplo de los muchos marcadores clasificados en los antepasados clave que deseamos imputar a los candidatos seleccionados. Debido a que los antepasados clave no se han genotipado para los 11 marcadores, utilizamos el modelo MCMC para deducir los genotipos marcadores de los 425 fundadores. Después, ejecutamos el programa MCMC nuevamente con el genotipo del marcador 11, eliminado en los candidatos de selección y utilizamos el programa MCMC para imputar el genotipo ausente. Inferimos solo un genotipo ausente, pero es un ejemplo típico porque si los fundadores hubieran sido genotipados para 110 marcadores, habría sido posible inferir los 100 marcadores ausentes con la misma precisión que el marcador realmente inferido por la iteración del proceso. Por ejemplo, el método puede probarse en un conjunto de datos adicional de los genotipos de aproximadamente 700 toros Holstein para 50.000 marcadores de SNP utilizando un ensayo *Illumina*, que forma los candidatos de selección, de los cuales se ocultan o enmascaran datos de todos menos de 2000 SNP; y los genotipos de los antepasados clave de estos toros se determinaron y utilizaron para imputar / inferir los 48,000 genotipos ausentes en los candidatos seleccionados.

Esperamos que el método de la invención tenga un mejor desempeño que en este modelo porque, en nuestro modelo, la necesidad de inferir los genotipos de los antepasados clave a partir de los de sus familiares significa que los genotipos de los antepasados clave pueden contener algunos errores. En una aplicación ideal, el árbol genealógico se conocería y/o los genotipos de los antepasados clave se conocerían, lo que permitiría que los candidatos de selección se remontaran a un número menor de antepasados clave. Por lo tanto, los ejemplos ilustrativos del presente documento demuestran que el método funcionará incluso en una situación desfavorable en la que los genotipos de los antepasados clave deben inferirse de los de sus familiares.

REIVINDICACIONES

1. Un método de selección artificial que comprende:

(i) genotipar a un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en una pluralidad de segmentos cromosómicos, en el que la población es una población de plantas o de animales que tiene un tamaño de población eficaz pequeño menor de 1000 individuos o de menos individuos,
(ii) rastrear los linajes de la pluralidad de segmentos cromosómicos hasta uno o más antepasados y/o fundadores de los que proceden,
(iii) inferir que los genotipos de cada segmento cromosómico son los mismos que para uno o más antepasados y/o fundadores de los que proceden los segmentos cromosómicos,
(iv) estimar el valor de reproducción del individuo basándose en los genotipos inferidos,
en donde cada antepasado o fundador es un antepasado o fundador que proporciona a la población actual, al menos el 0,1 % de la variación genética total y donde se conocen los genotipos de uno o más antepasados y/o fundadores de uno o más marcadores informativos, y
(v) seleccionar un individuo que tenga un valor de reproducción estimado elevado,
y en el que la población o el individuo no es un ser humano.

2. El método de acuerdo con la reivindicación 1, donde dicho método comprende:

(i) determinar un conjunto mínimo de antepasados y/o fundadores representativos de la población actual;
(ii) genotipar uno o más antepasados y/o fundadores de marcadores conocidos;
(iii) genotipar a un individuo en una población actual para determinar la presencia o la ausencia de uno o más marcadores informativos en una o una pluralidad de segmentos cromosómicos;
(iv) rastrear los linajes de una o una pluralidad de segmentos cromosómicos hasta uno o más antepasados y/o fundadores de los que proceden;
(v) inferir que los genotipos de cada segmento cromosómico son los mismos que para uno o más antepasados y/o fundadores de los que proceden los segmentos cromosómicos,
(vi) estimar el valor de reproducción del individuo basándose en los genotipos inferidos, y
(vii) seleccionar un individuo que tenga un valor de reproducción estimado elevado.

3. El método de acuerdo con la reivindicación 1 o la reivindicación 2, donde la población es una población de plantas o animales seleccionados de ganado, ovejas, cerdos, aves de corral, pescado, crustáceos o ganado Holstein.

4. El método de acuerdo con una cualquiera de las reivindicaciones 1 a 3, donde los marcadores, dentro de cada segmento cromosómico, están en desequilibrio de ligamiento.

5. El método de acuerdo con una cualquiera de las reivindicaciones 1 a 4, donde los marcadores se seleccionan de un alelo, un haplotipo, un halogrupo, un locus, un locus de rasgo cuantitativo, polimorfismos, tal como polimorfismos mononucleotídicos (SNP), STR y combinaciones de los mismos.

6. El método de acuerdo con una cualquiera de las reivindicaciones 1 a 5, que adicionalmente comprende inferir el árbol genealógico del individuo a partir de marcadores que se utilizan para trazar segmentos cromosómicos.

7. El método de acuerdo con una cualquiera de las reivindicaciones 1 a 6, que adicionalmente comprende caracterizar a los antepasados y/o fundadores mediante el genotipado de uno o más antepasados y/o fundadores para determinar marcadores conocidos.

8. El método de acuerdo con una cualquiera de las reivindicaciones 1 a 7, que adicionalmente comprende identificar a los antepasados y/o fundadores tal como determinando un conjunto mínimo de antepasados y/o fundadores representativos de la población actual.

9. El método de acuerdo con una cualquiera de las reivindicaciones 1 a 8, en el que se conocen las secuencias genómicas de los antepasados y/o fundadores clave y en el que dicho método comprende inferir secuencias genómicas de individuos en la población actual.

10. El uso del método de acuerdo con una cualquiera de las reivindicaciones 1 a 9, para seleccionar un individuo o material reproductivo o regenerativo del individuo, en el que el individuo no es un ser humano.

11. Un proceso para producir ganancia genética en una población que comprende realizar el método de acuerdo con una cualquiera de las reivindicaciones 1 a 9 y seleccionar un individuo de una población que tenga un valor de reproducción estimado elevado, en el que el individuo no es un ser humano.

12. Un método de selección artificial en ganado bovino que comprende:

(i) identificar el conjunto mínimo de antepasados clave que representan la mayoría de los segmentos cromosómicos en una población actual que tiene un tamaño de población eficaz pequeño menor de 1000 individuos;
(ii) genotipar los antepasados clave para un conjunto de marcadores densos;
(iii) genotipar uno o más individuos de una población actual para determinar marcadores suficientes para permitir así que los segmentos cromosómicos coincidan con los segmentos llevados por los antepasados clave;
(iv) rastrear los segmentos cromosómicos de uno o más individuos de la población actual hasta un antepasado clave;
(v) inferir los genotipos de los marcadores dentro de uno o más segmentos cromosómicos del uno o más individuos en la población actual para que sean los mismos que los del antepasado clave;
(vi) utilizar el genotipo inferido del uno o más individuos en la población actual para estimar el valor de reproducción de dicho uno o más individuos; y
(vii) seleccionar un individuo que tenga un valor de reproducción estimado elevado
y en el que la población o el individuo no es un ser humano.

13. El método de la reivindicación 12, en el que el ganado bovino es ganado Holstein.