(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 757 827**

(51) Int. CI.:

G16B 20/40 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **05.02.2010 PCT/US2010/023312**

(87) Fecha y número de publicación internacional: **12.08.2010 WO10091248**

(96) Fecha de presentación y número de la solicitud europea: **05.02.2010 E 10704279 (8)**

(97) Fecha y número de publicación de la concesión europea: **11.09.2019 EP 2399214**

(54) Título: **Método para seleccionar genes candidatos estadísticamente validados**

(30) Prioridad:

06.02.2009 US 367045

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
30.04.2020

(73) Titular/es:

SYNGENTA PARTICIPATIONS AG (100.0%)
Rosentalstrasse 67
4058 Basel, CH

(72) Inventor/es:

KISHORE, VENKATA, KRISHNA;
GUO, ZHIGANG;
LI, MIN;
WANG, DAOLONG;
GUTIERREZ ROJAS, LIBARDO, ANDRES;
CLARKE, JOSEPH, DALLAS, V. y
BYRUM, JOSEPH

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Método para seleccionar genes candidatos estadísticamente validados

CAMPO DE LA INVENCIÓN

Esta invención se refiere a la genética molecular, particularmente a métodos para evaluar una asociación entre un marcador genético y un fenotipo en una población.

ANTECEDENTES DE LA INVENCIÓN

Se han desarrollado múltiples paradigmas experimentales para identificar y analizar loci de rasgos cuantitativos (QTL) (véase, p. ej., Jansen (1996) Trends Plant Sci 1:89). Un locus de rasgo cuantitativo (QTL) es una región del genoma que codifica una o más proteínas y que explica una proporción significativa de la variabilidad de un fenotipo dado que puede ser controlado por múltiples genes. La mayoría de los informes publicados sobre el mapeo de QTL en especies de cultivos se han basado en el uso del cruce biparental. Típicamente, estos paradigmas implican cruzar uno o más pares parentales, que pueden ser, por ejemplo, un solo par derivado de dos cepas endogámicas, o múltiples parentales relacionados o no relacionados de diferentes cepas o líneas endogámicas, cada una de las cuales exhibe características diferentes con relación al rasgo fenotípico de interés. Típicamente, este protocolo experimental implica derivar de 100 a 300 descendientes de segregación de un solo cruce de dos líneas endogámicas divergentes (p. ej., seleccionadas para maximizar las diferencias de marcadores fenotípicos y moleculares entre las líneas). Los parentales y la progenie segregante se genotipan para múltiples loci marcadores y se evalúan para uno o varios rasgos cuantitativos (p. ej., resistencia a enfermedades). Los QTL se identifican como asociaciones estadísticas significativas entre los valores genotípicos y la variabilidad fenotípica entre la progenie segregante.

Los expertos en la técnica conocen numerosos métodos estadísticos para determinar si los marcadores están genéticamente ligados a un QTL (u otro marcador) e incluyen, p. ej., modelos lineales estándar, tales como ANOVA o mapeo de regresión (Haley y Knott (1992) Heredity 69: 315), métodos de probabilidad máxima, tales como algoritmos de maximización de expectativas (p. ej., Lander y Botstein (1989) Genetics 121:185-199; Jansen (1992) Theor. Appl. Genet., 85:252-260; Jansen (1993) Biometrics 49:227-231; Jansen (1994) En J. W. van Ooijen y J. Jansen (eds.), Biometrics in Plant breeding: applications of molecular markers, págs. 116-124, CPRO-DLO Metherlands; Jansen (1996) Genetics 142:305-311; y Jansen y Stam (1994) Genetics 136:1447-1455). Métodos estadísticos a modo de ejemplo incluyen análisis de marcadores de un solo punto, mapeo de intervalos (Lander y Botstein (1989) Genetics 121: 185), mapeo de intervalos compuestos, análisis de regresión penalizada, análisis de pedigrí complejo, análisis MCMC, análisis MQM (Jansen (1994) Genetics 138: 871), Análisis HAPLO-IM+, análisis HAPLO-MQM y análisis HAPLO-MQM+, MCMC bayesiano, regresión de cresta, análisis de identidad por descendencia y regresión de Haseman-Elston.

La disección de rasgos complejos en muchas especies se ha basado en gran medida en dos enfoques principales, el análisis de enlaces y el mapeo de asociación (Andersson y Georges 2004, *Nat. Rev. Genet.* **5:** 202–212; Flint et al. 2005, *Nat. Rev. Genet.* 6: 271–286; Hirschhorn y Daly 2005, *Nat. Rev. Genet.* **6:** 95-108). Si bien los métodos para el análisis de enlace que utilizan poblaciones de mapeo diseñadas se han empleado durante mucho tiempo (Doerge 2002, *Nat. Rev. Genet.* 3: 43–52), los métodos para el mapeo de asociación con muestras basadas en la población se desarrollaron más recientemente para superar la estructura de la población oculta o relación críptica dentro de las muestras recogidas (Falush et al. 2003, *Genetics* 164: 1567–1587; Yu et al. 2006, *Nat. Genet.* 38: 203-208). Se han estudiado métodos estadísticos para el enlace conjunto y la estrategia de mapeo de desequilibrio de enlace para poblaciones naturales (Wu y Zeng 2001, *Genetics* 157: 899–909; Wu et al. 2002, *Genetics* 160: 779-792) y también se ha examinado cruzar una endogamia con un material heterogéneo (Mott y Flint 2002, *Genetics* 160: 1609-1618). Para un pedigrí complejo general, el mapeo fino mediante la combinación de información de enlace y desequilibrio de enlace en regiones QTL previamente mapeadas ha identificado polimorfismos de genes candidatos (Meuwissen et al. 2002, *Genetics* 161: 373–379; Blott et al. 2003, *Genetics* 163: 253-266). Estudios previos de diseños genéticos con múltiples cruces de líneas han demostrado una potencia mejorada y una resolución de mapeo en una sola población (Rebai y Goffinet 1993, *Genet. Res.* 75: 243–247; Xu 1998, *Genetics* 148: 517–524; Rebai y Goffinet 2000, *Genet. Res.* 75: 243–247; Yi y Xu 2002, *Genetica* 114: 217–230; Jansen et al. 2003, *Crop Sci.* 43: 829–834; Li et al. 2005, *Genetics* 169: 1699–1709; Verhoeven et al. 2006, *Heredity* 96: 139-149). Sin embargo, estos estudios explotaron principalmente la información de enlace de cruces de múltiples líneas.

En el caso de seres humanos, el uso de la genética para identificar genes y vías asociadas con rasgos sigue un paradigma muy estándar. Primero, se realiza un estudio de vinculación de todo el genoma utilizando cientos de marcadores genéticos en datos basados en la familia para identificar regiones amplias vinculadas al rasgo. El resultado de este tipo de análisis de vinculación estándar es la identificación de regiones que controlan el rasgo, lo que restringe la atención de los más de 30.000 genes a unos 500 a 1000 genes en una región particular del genoma que está vinculada al rasgo. Sin embargo, las regiones identificadas mediante el análisis de enlaces siguen siendo demasiado amplias para identificar genes candidatos asociados con el rasgo. Por lo tanto, estos estudios de

vinculación suelen ser seguidos por un mapeo de precisión de las regiones de vinculación utilizando marcadores de mayor densidad en la región de vinculación, aumentando el número de familias en el análisis e identificando poblaciones alternativas para el estudio. Estos esfuerzos restringen aún más la atención a regiones más estrechas del genoma, del orden de 100 genes en una región particular vinculada al rasgo. Incluso con la región de enlace más estrechamente definida, el número de genes para validar sigue siendo irrazonablemente grande. Por lo tanto, la investigación en esta etapa se enfoca en identificar genes candidatos basados en la supuesta función de genes conocidos o predichos en la región y la relevancia potencial de esa función para el rasgo. Este enfoque es problemático porque se limita a lo que se conoce actualmente sobre los genes. A menudo, dicho conocimiento es limitado y está sujeto a interpretación. Como resultado, los investigadores a menudo se desvían y no identifican los genes que afectan al rasgo.

SUMARIO DE LA INVENCIÓN

La invención incluye evaluar o validar asociaciones entre genes candidatos y un rasgo de interés en una población. Los métodos de la invención comprenden una combinación única de análisis de asociación del genoma completo (GWA) y uno o ambos de mapeo de asociación anidada (NAM) y análisis de QTL de expresión (eQTL) para la selección y priorización de marcadores candidatos para su posterior implementación o uso. Los marcadores se seleccionan si se demuestra que se correlacionan positivamente con un rasgo de interés utilizando GWA y una combinación de uno o ambos de NAM y eQTL.

Además se proporcionan nuevos modelos de regresión para el mapeo de asociación anidada. Estos métodos comprenden un modelo de regresión de marcador único (SMR) y un modelo de regresión de marcador múltiple (MRM). Los genotipos no informativos se eliminan antes de evaluar una asociación entre un valor de rasgo y un genotipo marcador utilizando el modelo SMR. La regresión escalonada se utiliza para seleccionar marcadores de cofactor para su inclusión en el modelo MMR. En diversos aspectos de la invención, los marcadores se consideran para una validación adicional si se detecta una asociación utilizando tanto SMR como MMR.

Los marcadores identificados, seleccionados o validados utilizando los métodos de la invención pueden utilizarse en la reproducción y selección asistida por marcadores, como marcadores genéticos para construir mapas de enlace genético, para aislar la secuencia de ADN genómico que rodea una secuencia de ADN codificante o no codificante de genes, para identificar genes que contribuyen a un rasgo de interés y para generar organismos transgénicos que tienen un rasgo deseado.

BREVE DESCRIPCIÓN DE LAS FIGURAS

La Figura 1 es un diagrama de flujo a modo de ejemplo que representa los pasos implicados en el GWA.

La Figura 2 es un diagrama de flujo a modo de ejemplo que representa los pasos implicados en el NAM.

La Figura 3 es un diagrama de flujo a modo de ejemplo que representa los pasos implicados en la combinación del GWA y el NAM para seleccionar y priorizar un marcador candidato para uso posterior.

La Figura 4 es una representación esquemática a modo de ejemplo de selección y priorización basada en marcadores solapantes identificados utilizando el NAM (panel superior) y el GWA (panel inferior).

La Figura 5 muestra los histogramas para 3 rasgos relacionados con el etanol para 600 líneas endogámicas en el panel endogámico. Los datos fenotípicos se ajustaban bien a las distribuciones normales.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

Visión de conjunto

La estimación de las posiciones y los efectos de los loci de rasgos cuantitativos (QTL) es de importancia central para la selección asistida por marcadores. Hasta ahora, esto se ha logrado mediante enfoques de mapeo QTL clásicos (Lander y Botstein (1989) Genetics 121:185–199). Los experimentos necesarios requieren el establecimiento, así como el fenotipado y el genotipado de grandes poblaciones de mapeo y, por lo tanto, requieren mucho tiempo y costos (Parisseaux y Bernardo (2004) Theor Appl Genet 109:508–514).

Se describe en esta memoria un método para descubrir o validar una asociación entre uno o más genes candidatos y un rasgo fenotípico de interés. Como un ejemplo, los marcadores se seleccionan, validan o priorizan para su uso posterior comparando los marcadores correlacionados positivamente identificados utilizando el análisis de asociación del genoma completo (GWA) con los marcadores correlacionados positivamente utilizando otros modelos de asociación como el mapeo de asociación anidada (NAM) y/o análisis de expresión QTL (eQTL). Los marcadores correlacionados positivamente identificados utilizando GWA y uno o ambos análisis NAM y eQTL se colocan en un mapa genético físico de las especies en estudio. Los marcadores se priorizan para su uso posterior si se identifican tanto en los métodos GWA como en uno o ambos NAM y eQTL (es decir, marcadores "solapantes"). Por lo tanto, los métodos descritos en esta memoria facilitan la priorización de marcadores candidatos para la selección e

implementación en procesos posteriores para aumentar las posibilidades de éxito en el desarrollo de marcadores de diagnóstico para la mejora asistida por marcadores y el desarrollo de productos.

Además, en esta memoria se proporcionan nuevos métodos para el mapeo de asociación anidada (NAM). El NAM es un método para evaluar la asociación entre un marcador candidato y un rasgo de interés en una población anidada de organismos. Los métodos comprenden nuevos modelos de regresión simple y múltiple para evaluar una asociación entre un gen candidato y un rasgo de interés en una población anidada de *Zea mays*.

Para los fines de la presente invención, un "gen candidato" es un gen o elemento genético que está siendo ensayado para una asociación entre el gen y un rasgo de interés. El gen candidato puede ser un ortólogo de un gen conocido o del que se sospecha que está asociado con el rasgo de interés en una especie diferente. Como se utiliza en esta memoria, la expresión "asociado con" en relación con una relación entre un marcador genético (SNP*,* haplotipo, inserción/deleción, repetición en tándem, etc.) y un fenotipo se refiere a una dependencia estadísticamente significativa de la frecuencia del marcador con respecto a un escala cuantitativa o gradación cualitativa del fenotipo. Un marcador se correlaciona "positivamente" con un rasgo cuando está vinculado a él y cuando la presencia del marcador es un indicador de que el rasgo o la forma de rasgo deseado ocurrirá en un organismo que comprende el marcador. Un marcador se correlaciona negativamente con un rasgo cuando está vinculado a él y cuando la presencia del marcador es un indicador de que un rasgo o una forma de rasgo deseado no ocurrirá en un organismo que comprende el marcador. Para los fines de la presente invención, el término "marcador" se refiere a cualquier elemento genético que está siendo ensayado para una asociación con un rasgo de interés, y no necesariamente significa que el marcador esté correlacionado positiva o negativamente con el rasgo de interés.

Por lo tanto, un marcador se asocia con un rasgo de interés cuando los genotipos marcadores y los fenotipos de rasgos se encuentran juntos en la progenie de un organismo más a menudo que si los genotipos marcadores y los fenotipos de rasgos se segregan por separado. La expresión "rasgo fenotípico" se refiere a la apariencia, u otra característica de un organismo, p. ej., una planta o un animal, que resulta de la interacción de su genoma con el entorno. El término

"fenotipo" se refiere a cualquier propiedad visible, detectable o de otra manera medible de un organismo. El término "genotipo" se refiere a la constitución genética de un organismo. Esto puede considerarse en total, o con respecto a los alelos de un solo gen, es decir, en un locus genético dado.

En algunas realizaciones, los marcadores son genes o elementos genéticos candidatos directamente atribuibles al rasgo fenotípico. Por ejemplo, un elemento genético directamente atribuible a la acumulación de almidón en una planta puede ser un gen directamente implicado en el metabolismo del almidón de la planta. Alternativamente, el marcador puede encontrarse dentro de un locus genético asociado con el rasgo fenotípico de interés. Un "locus" es una región del cromosoma en donde se encuentra un ácido nucleico polimórfico, un determinante del rasgo, un gen o un marcador. Así, por ejemplo, un "locus genético" es una ubicación cromosómica específica en el genoma de una especie en donde se puede encontrar un gen específico. En diversas realizaciones, los marcadores identificados o validados utilizando los métodos descritos en esta memoria pueden estar asociados con un locus de rasgo cuantitativo (QTL). La expresión "locus de rasgos cuantitativos" o "QTL" se refiere a un locus genético polimórfico con al menos dos alelos que afectan de manera diferencial la expresión de un rasgo fenotípico en al menos un fondo genético.

En algunos aspectos, los genes candidatos identificados o validados utilizando los métodos descritos en esta memoria están vinculados o estrechamente vinculados a los marcadores de QTL. La expresión "estrechamente vinculado", en la presente solicitud, significa que la recombinación entre dos loci unidos se produce con una frecuencia igual o inferior a aproximadamente 10% (es decir, están separados en un mapa genético por no más de 10 cM). En otras palabras, los loci estrechamente vinculados se co-segregan al menos el 90% del tiempo. Los loci marcadores son especialmente útiles en la presente invención cuando demuestran una probabilidad significativa de co-segregación (enlace) con un rasgo deseado. En algunos aspectos, estos marcadores pueden denominarse marcadores QTL vinculados.

Los métodos descritos en esta memoria incorporan una diversidad de ensayos estadísticos y modelos que pueden no describirse explícitamente en esta memoria. Se puede encontrar una descripción detallada de los ensayos estadísticos estándar en libros de texto básicos sobre estadística, tales como, por ejemplo, Dixon, W. J. et al., Introduction to Statistical Analysis, Nueva York, McGraw-Hill (1969) o Steel R. G. D. et al., Principles and Procedures of Statistics: with Special Reference to the Biological Sciences, Nueva York, McGraw-Hill (1960). También hay un cierto número de programas de software para análisis estadísticos que son conocidos por un experto en la técnica.

Población de interés

Los métodos de la presente invención comprenden la identificación o la validación de un marcador candidato mediante la realización de un análisis de asociación del genoma completo (GWA) en una población de organismos (es decir, *Zea mays*), y la comparación de cualesquiera marcadores positivamente correlacionados en el análisis GWA con marcadores que se determina que tienen una correlación positiva con el rasgo de interés en la misma

especie del organismo utilizando uno o ambos del mapeo de asociación anidada (NAM) y análisis de expresión QTL (eQTL). Los marcadores candidatos se priorizan para su posterior uso o implementación (p. ej., mejora genética asistida por marcadores, desarrollo de plantas transgénicas y similares) cuando se demuestra que el marcador tiene una correlación positiva en el análisis GWA y al menos otro método de análisis de enlace, por ejemplo, al menos uno de análisis eQTL, NAM o AEA. No es necesario que se utilice la misma población de mapeo para cada uno de los análisis, siempre que la población de todos los estudios consista en organismos de la misma especie.

La mayoría de los informes publicados sobre el mapeo de QTL en especies de cultivos se han basado en el uso del cruce biparental (Lynch y Walsh (1997) Genetics and Analysis of Quantitative Traits Sinauer Associates, Sunderland). Típicamente, este protocolo experimental implica derivar de 100 a 300 descendientes de segregación de un solo cruce de dos líneas endogámicas divergentes (p. ej., seleccionadas para maximizar las diferencias de marcadores fenotípicos y moleculares entre las líneas). La progenie segregante se genotipa para múltiples loci marcadores y se evalúa para uno o varios rasgos cuantitativos en varios entornos. Los QTL se identifican como asociaciones estadísticas significativas entre los valores genotípicos y la variabilidad fenotípica entre la progenie segregante.

Los métodos proporcionados en este documento son útiles para descubrir o validar asociaciones de marcadores: rasgos en cualquier población. El término "población" o la expresión "población de organismos" indica un grupo de organismos de la misma especie, por ejemplo, de los cuales se toman muestras para evaluación, y/o de los cuales se seleccionan miembros individuales para fines de reproducción. Los miembros de la población a partir de los cuales se evalúan los marcadores no necesitan ser idénticos a los miembros de la población finalmente seleccionados para la reproducción para obtener la progenie, p. ej., la progenie utilizada para los ciclos de análisis posteriores.

En realizaciones de la invención, la población de organismos, tal como una población de plantas, comprende o consiste en una población resultante de cruces entre una o más líneas fundadoras y una única línea parental común. En diversos ejemplos, la línea principal común única es una línea de ensayo. La expresión "línea de ensayo" se refiere a una línea que no está relacionada y que es genéticamente diferente de un conjunto de líneas con las que se cruza. El uso de un parental de ensayo en un cruce sexual permite a un experto determinar la asociación del rasgo fenotípico con la expresión de loci de rasgos cuantitativos en una combinación híbrida. La expresión "combinación híbrida" se refiere al proceso de cruzar un parental de ensayo con múltiples líneas. El fin de producir cruces de este tipo es evaluar la capacidad de las líneas de producir fenotipos deseables en la progenie híbrida derivada de la línea por el cruce de ensayo.

La progenie del cruce entre las líneas fundadoras y la línea de ensayo se somete a múltiples rondas de "autofecundación" para generar una población que segrega todos los genes de una manera mendeliana. A esta población de mapeo se la alude en esta memoria como la "población anidada" y es útil para la realización particular de la invención que implementa métodos de mapeo asociado anidado (NAM), por ejemplo, los nuevos métodos NAM descritos en esta memoria. Estas líneas endogámicas recombinantes (RIL) (líneas genéticamente relacionadas; generalmente $>F_5$, desarrolladas a partir de líneas F_2 continuamente autofecundadas hacia la homocigosidad) pueden utilizarse como población de mapeo. La información obtenida de los marcadores dominantes se puede maximizar utilizando RIL, porque todos los loci son homocigotos o casi. En condiciones de enlace estrecho (es decir, aproximadamente <10% de recombinación), los marcadores dominantes y co-dominantes evaluados en poblaciones RIL proporcionan más información por individuo que cualquier tipo de marcador en poblaciones de retrocruzamiento (Reiter et al., Proc. Natl. Acad. Sci. (U.S.A.) 89:1477-1481 (1992)).

El término "cruzado" o "cruce" en el contexto de esta invención significa la fusión de gametos por polinización para producir progenie (p. ej., células, semillas o plantas). El término abarca tanto los cruces sexuales (p. ej., la polinización de una planta por otra, o la fertilización de un gameto por otro) y la autofecundación (p. ej., la autopolinización, p. ej., cuando el polen y el óvulo son de la misma planta). El término "híbrido" se refiere a organismos que resultan de un cruce entre individuos genéticamente divergentes. El término "endogámico" se refiere a organismos derivados de un cruce entre individuos genéticamente relacionados. El término "líneas" en el contexto de esta invención se refiere a una familia de plantas relacionadas derivadas por auto-polinización de una planta endogámica. El término "progenie" se refiere a los descendientes de un organismo particular (p. ej., plantas autofecundadas) o un par de organismos (p. ej., a través de cruce sexual). Los descendientes pueden ser, por ejemplo, de la generación F_1, la F_2 o de cualquier generación posterior.

Los métodos descritos en esta memoria comprenden, además, un cruce híbrido entre una línea de ensayo y una línea de élite. Una "línea de élite" o "cepa de élite" es una línea agronómicamente superior que ha resultado de muchos ciclos de reproducción y selección para un comportamiento agronómico superior. Por el contrario, una "cepa exótica" o un "germoplasma exótico" es una cepa o un germoplasma derivado de un organismo que no pertenece a una línea o cepa de germoplasma de élite disponible. Numerosas líneas de élite están disponibles y son conocidas por los expertos en el técnica de reproducción. Una "población de élite" es una variedad de individuos o líneas de élite que se pueden utilizar para representar el estado de la técnica en términos de genotipos agronómicamente superiores de una especie dada. De manera similar, un "germoplasma de élite" o cepa de germoplasma de élite es un germoplasma agronómicamente superior, típicamente derivado y/o capaz de dar lugar a un organismo con un comportamiento agronómico superior. El término "germoplasma" se refiere a material genético de o procedente de

un individuo (p. ej., una planta o animal), un grupo de individuos (p. ej., una línea de planta, variedad o familia), o un clon derivado de una línea, variedad, especie o cultivo. El germoplasma puede ser parte de un organismo o célula, o puede estar separado del organismo o de la célula. En general, el germoplasma proporciona material genético con una composición molecular específica que proporciona una base física para algunas o todas las cualidades hereditarias de un organismo o cultivo celular.

En algunos casos, una población puede incluir organismos parentales, así como una o más progenie derivada del organismo parental. En algunos casos, una población se deriva de un solo cruce biparental, p. ej., una población de progenie de un cruce entre dos parentales. Alternativamente, una población incluye miembros derivados de dos o más cruces que implican a los mismos o diferentes parentales. La población puede consistir en líneas endogámicas recombinantes, líneas de retrocruzamiento, líneas de cruce de ensayo y similares.

En diversos ejemplos, la población es una población de plantas que consiste en materiales de reproducción de fase temprana. Por material de reproducción de "fase temprana" se pretende que las plantas estén en la generación F2 a F3. El uso de materiales de reproducción de fase temprana encuentra ventaja en que el número de materiales de reproducción disponibles es grande; los datos fenotípicos están disponibles para las líneas de reproducción; y los resultados del mapeo pueden ayudar directamente a la selección. En las primeras etapas de la reproducción, se someten a ensayo múltiples líneas en múltiples ubicaciones.

Debido a que las primeras fases de reproducción implican la evaluación de un gran número de progenies derivadas de cruces múltiples, proporcionan los datos fenotípicos necesarios para identificar y validar marcadores para una amplia gama de rasgos agronómicos. Al integrar análisis de marcadores en los programas de mejora existentes, se puede lograr el poder, la precisión y la exactitud asociados

con un gran número de progenie. Además, se pueden hacer inferencias sobre las asociaciones de marcadores a través del programa de mejora en lugar de limitarse a la muestra de progenie de un solo cruce.

Las poblaciones de retrocruzamiento (p. ej., generadas a partir de un cruce entre una variedad de éxito (parental recurrente) y otra variedad (parental donante) que tiene un rasgo no presente en la primera) pueden utilizarse como una población de mapeo. Se puede hacer una serie de retrocruces con el parental recurrente para recuperar la mayoría de sus rasgos deseables. Por lo tanto, se crea una población que consiste en individuos casi como el parental recurrente, pero cada uno de los individuos lleva cantidades variables o mosaicos de regiones genómicas del parental donante. Las poblaciones de retrocruzamiento pueden ser útiles para mapear marcadores dominantes si todos los loci en el parental recurrente son homocigotos y los parentales donante y recurrente tienen alelos marcadores polimórficos contrastantes (Reiter et al., Proc. Natl. Acad. Sci. (U.S.A.) 89:1477-1481 (1992)). La información obtenida de las poblaciones de retrocruzamiento que utilizan marcadores co-dominantes o dominantes es menor que la obtenida de las poblaciones F_2 porque se muestrean uno, en lugar de dos gametos recombinantes por planta. Sin embargo, las poblaciones de retrocruzamiento son más informativas (con baja saturación de marcadores) en comparación con las RILs, ya que la distancia entre los loci enlazados aumenta en las poblaciones de RIL (es decir, aproximadamente el 0,15% de recombinación). La recombinación incrementada puede ser beneficiosa para la resolución de enlaces estrechos, pero puede ser indeseable en la construcción de mapas con baja saturación de marcadores.

En otro ejemplo, la población consiste en plantas endogámicas agrupadas en pedigrí según los parentales comunes. Una "estructura de pedigrí" define la relación entre un descendiente y cada uno de los antepasados que dio origen a ese descendiente. Una estructura de pedigrí puede abarcar una o más generaciones, describiendo las relaciones entre el descendiente y sus padres, abuelos, bisabuelos, etc.

De acuerdo con la invención, los marcadores se pueden identificar o validar utilizando una población de mapeo existente. Por ejemplo, la población de mapeo descrita en Yu *et al.* (2008) *Genetics* 178:539-551 puede utilizarse particularmente para los métodos NAM. Otras poblaciones de mapeo públicas o privadas pueden ser adecuadas para los métodos descritos en esta memoria.

Marcadores Genéticos

Aunque secuencias de ADN específicas que codifican proteínas generalmente están bien conservadas en una especie, otras regiones de ADN (típicamente no codificantes) tienden a acumular polimorfismo y, por lo tanto, pueden ser variables entre individuos de la misma especie. Dichas regiones proporcionan la base para numerosos marcadores genéticos moleculares.

Después de la generación o selección de uno o más poblaciones en los métodos descritos en esta memoria, se obtiene un valor genotípico para una pluralidad de marcadores para una pluralidad de miembros de la o las poblaciones. El valor genotípico corresponde a la medida cuantitativa o cualitativa del marcador genético. El término "marcador" se refiere a una secuencia de ADN identificable que es variable (polimórfica) para diferentes individuos dentro de una población y facilita el estudio de la herencia de un rasgo o un gen. Un marcador en el nivel de secuencia de ADN está vinculado a una ubicación cromosómica específica única para el genotipo de un individuo y se hereda de manera predecible.

El marcador genético es típicamente una secuencia de ADN que tiene una ubicación específica en un cromosoma que se puede medir en un laboratorio. La expresión "marcador genético" también puede utilizarse para referirse, p. ej., a un ADNc y/o un ARNm codificado por una secuencia genómica, así como a esa secuencia genómica. Para ser útil, un marcador necesita tener dos o más alelos o variantes. Los marcadores pueden ser directos, es decir, están ubicados dentro del gen o locus de interés, o indirectos, que están estrechamente relacionados con el gen o el locus de interés (presumiblemente debido a una ubicación cercana, pero no dentro del gen o locus de interés). Además, los marcadores también pueden incluir secuencias que modifican o no modifican la secuencia de aminoácidos codificada por el gen en el que se encuentran.

En general, cualquier rasgo polimórfico heredado diferencialmente (incluido el polimorfismo de ácido nucleico) que se segrega entre la progenie es un marcador potencial. El término "polimorfismo" se refiere a la presencia en una población de dos o más variantes alélicas. El término "alelo" o "alélico" o la expresión "variante de marcador" se refiere a la variación presente en una posición definida dentro de un marcador o secuencia de marcador específica; en el caso de un SNP, este es el nucleótido real que está presente; para un SSR, es el número de secuencias repetidas; para una secuencia peptídica, es el aminoácido real presente; en el caso de un haplotipo de marcador, es la combinación de dos o más variantes de marcador individuales en una combinación específica. Un "alelo asociado" se refiere a un alelo en un locus polimórfico que está asociado con un fenotipo particular de interés. Variantes alélicas de este tipo incluyen variación de la secuencia en una base única, por ejemplo, un polimorfismo de un solo nucleótido (SNP). Un polimorfismo puede ser una diferencia de un solo nucleótido presente en un locus, o puede ser una inserción o deleción de uno, unos pocos o muchos nucleótidos consecutivos. Se reconocerá que si bien los métodos de la invención se ejemplifican principalmente mediante la detección de SNPs, estos métodos u otros conocidos en la técnica pueden utilizarse de manera similar para identificar otros tipos de polimorfismos, que típicamente implican más de un nucleótido.

La variabilidad genómica puede ser de cualquier origen, por ejemplo, inserciones, deleciones, duplicaciones, elementos repetitivos, mutaciones puntuales, eventos de recombinación o la presencia y secuencia de elementos transponibles. El marcador se puede medir directamente como un polimorfismo de secuencia de ADN, tal como un polimorfismo de un solo nucleótido (SNP), polimorfismo de longitud de fragmento de restricción (RFLP) o repetición en tándem corto (STR), o indirectamente como una variante de secuencia de ADN, como un polimorfismo de conformación de una sola cadena (SSCP). Un marcador también puede ser una variante a nivel de un producto derivado de ADN, tal como un polimorfismo/abundancia de ARN, un polimorfismo de proteínas o un polimorfismo de metabolitos celulares, o cualquier otra característica biológica que tenga una relación directa con la variante de ADN o producto genético subyacente.

Con frecuencia se utilizan dos tipos de marcadores en el mapeo y en los protocolos de reproducción asistida por marcadores, a saber, marcadores de repetición de secuencia simple (SSR, también conocidos como microsatélites) y marcadores de polimorfismo de un solo nucleótido (SNP). El término SSR se refiere generalmente a cualquier tipo de heterogeneidad molecular que resulta en variabilidad de longitud, y más típicamente es un segmento corto (hasta varios cientos de pares de bases) de ADN que consiste en múltiples repeticiones en tándem de una secuencia de dos o tres pares de bases. Estas secuencias repetidas dan como resultado regiones de ADN altamente polimórficas de longitud variable debido a la baja fidelidad de replicación, p. ej., provocada por el deslizamiento de la polimerasa. Los SSR parecen estar dispersos al azar a través del genoma y generalmente están flanqueados por regiones conservadas. Los marcadores SSR también pueden derivarse de secuencias de ARN (en forma de un ADNc, un ADNc parcial o un EST), así como material genómico.

En un ejemplo, el marcador molecular es un polimorfismo de un solo nucleótido. Se han desarrollado diversas técnicas para la detección de SNPs, incluida la hibridación específica de alelos (ASH; véase, p. ej., Coryell et al., (1999) Theor. Appl. Genet., 98:690-696). También se utilizan ampliamente otros tipos de marcadores moleculares, que incluyen, pero no se limitan a etiquetas de secuencia expresada (EST) y marcadores SSR derivados de secuencias EST, polimorfismo de longitud de fragmento amplificado (AFLP), ADN polimórfico amplificado al azar (RAPD) y marcadores de isoenzimas. Un experto en la materia conoce una amplia gama de protocolos para detectar esta variabilidad, y estos protocolos son frecuentemente específicos para el tipo de polimorfismo que están diseñados para detectar. Por ejemplo, la amplificación por PCR, polimorfismos de conformación de cadena sencilla (SSCP) y replicación de secuencia auto-sostenida (3SR; véase Chan y Fox, Reviews in Medical Microbiology 10:185-196).

El ADN para el análisis de marcadores puede recogerse y rastrearse en cualquier tejido conveniente, tales como células, semillas o tejidos a partir de los cuales se pueden cultivar nuevas plantas, o partes de plantas, tales como hojas, tallos, polen o células, que pueden cultivarse en una planta entera. En algunos ejemplos, los datos del marcador se toman de tejidos que se han asociado con el rasgo en estudio. En los ejemplos, los datos del marcador se miden a partir de múltiples tejidos de cada una de las plantas en estudio. Se obtiene un número suficiente de células para proporcionar una cantidad suficiente de muestra para el análisis, aunque solo un tamaño de muestra mínimo

será necesario en los casos en los que la puntuación sea por amplificación de ácidos nucleicos. El ADN, el ARN o la proteína pueden aislarse de la muestra celular mediante técnicas estándar de aislamiento de ácido nucleico conocidas por los expertos en la técnica.

En un ejemplo, los valores genotípicos corresponden a los valores obtenidos para esencialmente todos, o la totalidad de los SNPs de un mapa de los SNP del genoma completo de alta densidad. Este enfoque tiene la ventaja sobre los enfoques tradicionales de que, dado que abarca todo el genoma, identifica las posibles interacciones de productos genómicos expresados a partir de genes ubicados en cualquier parte del genoma sin requerir un conocimiento preexistente con respecto a una posible interacción entre los productos genómicos. Un ejemplo de un mapa de los SNP de genoma completo de alta densidad es un mapa de al menos aproximadamente 1 SNP por cada 10.000 kb, de al menos 1 SNP por cada 500 kb o de aproximadamente 10 SNP por cada 500 kb, o al menos aproximadamente 25 SNPs o más por cada 500 kb. Las definiciones de las densidades de los marcadores pueden cambiar a través del genoma y están determinadas por el grado de desequilibrio de enlace dentro de una región del genoma.

Adicionalmente, un cierto número de plataformas de rastreo de marcadores genéticos están ahora disponibles comercialmente, y se pueden utilizar para obtener datos de marcadores genéticos necesarios para el proceso de los presentes métodos. En muchos casos, estas plataformas pueden adoptar la forma de matrices de ensayo de marcadores genéticos (micromatrices), que permiten el ensayo simultáneo de muchos miles de marcadores genéticos. Por ejemplo, estas matrices pueden someter a ensayo marcadores genéticos en números de más de 1.000, más de 1.500, más de 2.500, más de 5.000, más de 10.000, más de 15.000, más de 20.000, más de 25.000, más de 30.000, más de 35.000, más de 40.000, más de 45.000, más de 50.000 o más de 100.000, más de 250.000, más de 500.000, más de 1.000.000, más de 5.000.000, más de 10.000.000 o más de 15.000.000. Ejemplos de un producto comercialmente disponible de este tipo son aquellos comercializados por Affymetrix Inc ((www.affymetrix.com)) o Illumina (www.illumina.com). En un ejemplo, el valor genotípico se obtiene de al menos 2 marcadores genéticos.

Se apreciará que, debido a la naturaleza de dicha información, puede requerirse un filtrado o procesamiento previo de los datos, es decir, un control de calidad de los datos. Por ejemplo, los datos de marcadores pueden excluirse de acuerdo con un criterio particular (p. ej., duplicación de datos o baja frecuencia; véase, por ejemplo, Zenger et. al (2007) *Anim Genet.* 38(1):7-14). A continuación se describen ejemplos de dicha filtración, aunque también se pueden emplear otros métodos para filtrar los datos como apreciaría el experto en la materia para obtener un conjunto de datos de trabajo en el que se determina la asociación del marcador.

En un ejemplo, los datos del marcador se excluyen del análisis en el que la frecuencia alélica de un marcador particular es inferior a aproximadamente 0,01 o inferior a aproximadamente 0,05. "Frecuencia alélica" o " frecuencia alélica de un marcador" (MAF) se refiere a la frecuencia (proporción o porcentaje) a la que un alelo está presente en un locus dentro de un individuo, dentro de una línea o dentro de una población de líneas. Por ejemplo, para un alelo "A", los individuos diploides del genotipo "AA", "Aa" o "aa" tienen frecuencias alélicas de 1,0, 0,5 o 0,0, respectivamente. Se puede estimar la frecuencia alélica dentro de una línea promediando las frecuencias alélicas de una muestra de individuos de esa línea. De manera similar se puede calcular la frecuencia alélica dentro de una población de líneas promediando las frecuencias alélicas de las líneas que componen la población. Para una población con un número finito de individuos o líneas, una frecuencia alélica puede expresarse como un recuento de individuos o líneas (o cualquier otro grupo especificado) que contenga el alelo.

En diversos ejemplos, el conjunto de marcadores evaluados para un rasgo de interés particular puede ser marcadores aleatorios tal como se describe anteriormente, o pueden ser marcadores que han mostrado o se sospecha que están asociados con el rasgo de interés en una especie de planta diferente. Se conoce un gran número de marcadores moleculares para diversas especies en la técnica y se pueden validar en diferentes especies utilizando los métodos descritos en esta memoria. Por ejemplo, un grupo de genes candidatos que ha sido identificado en función de sus funciones moleculares y/o comportamientos en el maíz puede analizarse en soja. Por lo tanto, los modelos descritos en esta memoria son útiles para validar los efectos de estos genes candidatos en una especie de planta diferente. Al evaluar un conjunto de marcadores candidatos, los marcadores generalmente aleatorios que no tienen asociación conocida también se incluirán en el análisis.

Rasgo de interés

Los métodos de la presente invención son aplicables a cualquier fenotipo con un componente genético subyacente, es decir, cualquier rasgo hereditario. Un "rasgo" es una característica de un organismo que se manifiesta por sí misma en un fenotipo y se refiere a una o más características biológicas, de comportamiento o de cualquier otra característica o características mensurables, que puede ser cualquier entidad que pueda cuantificarse en una muestra biológica u organismo, que luego puede utilizarse sola o en combinación con una o más entidades cuantificadas. Un "fenotipo" es una apariencia externa u otra característica visible de un organismo y se refiere a uno o más rasgos de un organismo.

Se pueden inferir muchos rasgos diferentes por los métodos descritos en esta memoria. El fenotipo puede observarse a simple vista, o por cualquier otro medio de evaluación conocido en la técnica, p. ej., microscopía, análisis bioquímico, análisis genómico, un ensayo de resistencia a una enfermedad en particular, etc. En algunos casos, un fenotipo es directamente controlado por un solo gen o locus genético, es decir, un "rasgo de un solo gen". En otros casos, un fenotipo es el resultado de varios genes. Un "loci de rasgos cuantitativos" (QTL) es un dominio genético que es polimórfico y produce un fenotipo que puede describirse en términos cuantitativos, p. ej., altura,

peso, contenido en aceite, días hasta la germinación, resistencia a enfermedades, etc. y, por lo tanto, se le puede asignar un "valor fenotípico" que corresponda a un valor cuantitativo para el rasgo fenotípico.

Para cualquier rasgo, una característica "relativamente alta" indica mayor que la media, y una característica "relativamente baja" indica menos que la media. Por ejemplo, "rendimiento relativamente alto" indica un rendimiento de planta más abundante que el rendimiento medio para una población de plantas particular. A la inversa, "rendimiento relativamente bajo" indica un rendimiento menos abundante que el rendimiento medio para una población de plantas particular.

En el contexto de un programa ejemplar de fitomejora, los fenotipos cuantitativos incluyen, rendimiento (p. e., rendimiento de grano, rendimiento de ensilaje), estrés (p. ej., estrés de mitad de temporada, estrés terminal, estrés por humedad, estrés por calor, etc.), resistencia, resistencia a enfermedades , resistencia a insectos, resistencia a la densidad, número de grano, tamaño de grano, tamaño de mazorca, número de mazorcas, número de vainas, número de semillas por vaina, madurez, tiempo para florecer, unidades de calor para florecer, días para florecer, resistencia de alojamiento de raíces, alojamiento de tallos resistencia, altura de la mazorca, contenido de humedad del grano, peso de ensayo, contenido de almidón, composición de grano, composición de almidón, composición de aceite, composición de proteínas, contenido de nutracéuticos y similares.

Además, los siguientes valores fenotípicos pueden ser correlacionados con el marcador de interés: color, tamaño, forma, grosor de la piel, densidad de la pulpa, contenido de pigmento, depósitos de aceite, contenido de proteínas, actividad enzimática, contenido de lípidos, contenido de azúcar y almidón, contenido de clorofila, minerales, contenido de sal, acidez, aroma y sabor y otras características. Para cada uno de estos índices, se determina una distribución de parámetros para la muestra determinando una característica (p. ej., el peso) asociada con cada uno de los elementos de la muestra, y luego midiendo los valores de la media y la desviación estándar de la distribución.

Del mismo modo, los métodos son igualmente aplicables a rasgos que son continuamente variables, tales como el rendimiento de grano, la altura, el contenido de aceite, la respuesta al estrés (p. ej., estrés terminal o de mitad de temporada) y similares, o rasgos merísticos que son de multi-categoría, pero se pueden analizar como si fueran continuamente variables, tales como días hasta la germinación, días de floración o fructificación, y los rasgos con se distribuyen de manera no continua (discontinua) o discreta. Sin embargo, debe entenderse que los rasgos análogos u otros únicos pueden caracterizarse utilizando los métodos descritos en esta memoria, dentro de cualquier organismo de interés.

Además de los fenotipos directamente evaluables a simple vista, con o sin la ayuda de uno o más dispositivos manuales o automatizados, incluidos, p. ej., microscopios, escalas, reglas, calibradores, etc., se pueden evaluar muchos fenotipos utilizando productos bioquímicos y/o medios moleculares. Por ejemplo, se puede evaluar el contenido en aceite, el contenido en almidón, el contenido en proteínas, el contenido nutracéutico, así como sus componentes constituyentes, opcionalmente después de una o más etapas de separación o purificación, utilizando uno o más ensayos químicos o bioquímicos. Fenotipos moleculares, tales como los perfiles de metabolitos o los perfiles de expresión, ya sea a nivel de proteína o de ARN, también son susceptibles de evaluación de acuerdo con los métodos de la presente invención. Por ejemplo, los perfiles de metabolitos, ya sean metabolitos de moléculas pequeñas o biomoléculas grandes producidas por una vía metabólica, proporcionan información valiosa sobre fenotipos de interés agronómico. Perfiles de metabolitos de este tipo pueden evaluarse como medidas directas o indirectas de un fenotipo de interés. De manera similar, los perfiles de expresión pueden servir como medidas indirectas de un fenotipo, o pueden servir por sí mismos directamente como el fenotipo sujeto al análisis para fines de correlación de marcadores. Los perfiles de expresión se evalúan con frecuencia a nivel de productos de expresión de ARN, p. ej., en un formato de matriz, pero también se pueden evaluar a nivel de proteína utilizando anticuerpos u otras proteínas de unión.

Además, en algunas circunstancias es deseable emplear una relación matemática entre los atributos fenotípicos en lugar de correlacionar la información del marcador de forma independiente con múltiples fenotipos de interés. Por ejemplo, el objetivo final de un programa de mejora puede ser obtener plantas de cultivo que produzcan un alto rendimiento en condiciones de poco agua, es decir, de sequía. En lugar de correlacionar independientemente el marcador de rendimiento y resistencia a las condiciones de poco agua, un indicador matemático del rendimiento y la estabilidad del rendimiento sobre las condiciones del agua puede correlacionarse con marcadores. Un indicador matemático puede adoptar formas que incluyen; un valor de índice estadísticamente derivado, basado en contribuciones ponderadas de valores de un cierto número de rasgos individuales, o una variable que es un componente de un modelo de crecimiento y desarrollo de cultivos o un modelo ecofisiológico (al que se alude colectivamente como modelos de crecimiento de cultivos) de respuestas de rasgos de plantas a través de múltiples condiciones ambientales. Estos modelos de crecimiento de cultivos son conocidos en la técnica y se han utilizado para estudiar los efectos de la variación genética de los rasgos de las plantas y el mapa QTL para las respuesta de los

rasgos de las plantas. Véanse las referencias de Hammer et al. 2002 European Journal of Agronomy 18: 15-31, Chapman et al. 2003. Agronomy Journal 95: 99-113, y Reymond et al. 2003. Plant Physiology 131: 664-675.

Análisis de Asociación

Los métodos descritos en esta memoria implicaron la comparación de marcadores positivamente asociados identificados o validados por múltiples estrategias de análisis de enlaces. En diversos ejemplos, los marcadores se someten a ensayo utilizando estrategias de mapeo de asociación de genoma completo (GWA). Los marcadores correlacionados positivamente se alinean en un mapa genético físico de la especie que está siendo ensayada. Marcadores correlacionados positivamente identificados o validados utilizando otros métodos, tales como el análisis eQTL o NAM, también se alinean en el mapa físico. Los marcadores candidatos se seleccionan para su uso posterior si los marcadores se identifican o validan utilizando GWA y uno o ambos de eQTL o NAM.

Los datos genéticos se han utilizado en el campo del análisis de rasgos para intentar identificar los genes que afectan a este tipo de rasgos. Un desarrollo clave en tales actividades ha sido el desarrollo de grandes colecciones de marcadores moleculares/genéticos, que pueden utilizarse para construir mapas genéticos detallados de especies. Estos mapas se utilizan en metodologías de mapeo de Locus de Rasgos Cuantitativos (QTL) como mapeo de un solo marcador, mapeo de intervalos, mapeo de intervalos compuestos y mapeo de rasgos múltiples. Las metodologías de mapeo QTL proporcionan análisis estadísticos de la asociación entre fenotipos y genotipos con el propósito de comprender y diseccionar las regiones de un genoma que afectan a los rasgos.

Un mapeo de asociación utiliza marcadores dentro de genes candidatos, que son genes que se cree que están implicados funcionalmente en el desarrollo del rasgo debido a información, tal como bioquímica, fisiología, perfiles transcripcionales y experimentos genéticos inversos en organismos modelo. En la definición más simple, el mapeo de asociación es la utilidad del desequilibrio de enlace, también conocido como desequilibrio de la fase gamética, en poblaciones naturales para identificar marcadores con diferencias significativas de frecuencia alélica entre individuos con el rasgo de interés e individuos que no exhiben el rasgo de interés. El análisis de la asociación de genoma completo (GWA) es un enfoque que implica escanear rápidamente marcadores a través de los conjuntos completos (o casi completos) de ADN o genomas de organismos de la población para encontrar variaciones genéticas asociadas con un rasgo particular. Se considera que una asociación estadística entre genotipos en un locus marcador y el rasgo de interés es evidencia de una estrecha vinculación física entre el marcador y los QTL que controla ese rasgo (Pritchard et al., 2000).

Si bien los enfoques de mapeo genético clásico son útiles en la exploración del genoma completo para el control de loci QTL, el mapeo de asociación se está convirtiendo en una herramienta líder para la estimación precisa de las posiciones de los QTL. Por ejemplo, este método se ha utilizado para identificar genes para rasgos complejos en genética médica (Lander y Schork, 1994; Risch, 2000), y su aplicación se está moviendo gradualmente a otros campos, tales como la genética de las plantas. Dado que el mapeo de asociación utiliza poblaciones naturales, han transcurrido muchas generaciones (y, por lo tanto, meiosis), por lo que la recombinación habrá eliminado la asociación entre un QTL y cualquier marcador que no esté estrechamente vinculado a él. El mapeo de asociación permite un mapeo mucho más fino que los enfoques cruzados bi-parentales estándar.

Los datos de marcadores a intervalos regulares en todo el genoma en estudio o en las regiones del gen de interés se utilizan para controlar la segregación o detectar asociaciones en una población de interés. En algunos ejemplos, estos intervalos definidos regularmente se definen en Morgans o, más típicamente, centimorgans (cM). Un Morgan es una unidad que expresa la distancia genética entre marcadores en un cromosoma. Un Morgan se define como la distancia en un cromosoma en el que se espera que ocurra un evento de recombinación por gameto por generación. En algunos ejemplos, cada uno de los intervalos definidos regularmente es inferior a 100 cM. En otros ejemplos, cada uno de los intervalos definidos regularmente es menos de 10 cM , menos de 5 cM , menos de 2,5 cM , menos de 2 cM , menos de 1,5 cM o menos de 1 cM.

Modelos de vinculación para la Asociación de Genoma Completo

El objetivo del mapeo genético es identificar marcadores simplemente heredados en estrecha proximidad de factores genéticos que afectan a los rasgos cuantitativos, es decir, QTL. Esta localización se basa en procesos que crean una asociación estadística entre el marcador y los alelos de QTL y procesos que reducen selectivamente esa asociación en función de la distancia del marcador desde los QTL. Pueden utilizarse varios tipos de conocidos análisis estadísticos para inferir la asociación marcador/rasgo a partir de los datos del fenotipo/genotipo, pero una idea básica es la de detectar los marcadores, es decir, polimorfismos, para los cuales los genotipos alternativos tienen significativamente diferentes fenotipos medios. Por ejemplo, si un locus A de un marcador dado tiene tres genotipos alternativos (AA, Aa y aa), y si esas tres clases de individuos tienen fenotipos significativamente diferentes, entonces se infiere que el locus A está asociado con el rasgo. La importancia de las diferencias en el fenotipo puede someterse a ensayo mediante varios tipos de ensayos estadísticos estándar, tales como la regresión lineal de genotipos de marcadores en el fenotipo o el análisis de varianza (ANOVA). El mapa genético se crea colocando marcadores genéticos en orden de mapa genético (lineal) de modo que se entiendan las relaciones posicionales entre los marcadores.

Se pueden utilizar muchos programas conocidos para realizar análisis de asociación de acuerdo con este aspecto de la invención. Uno de estos programas es MapMaker/QTL, que es el programa complementario de MapMaker y es el software de mapeo QTL original. MapMaker/QTL analiza los datos del marcador mediante el mapeo de intervalos estándar. Otro de estos programas es QTL Cartographer, que realiza regresión de marcador único, mapeo de intervalos (Lander y Botstein , Id.), mapeo de intervalos múltiples y mapeo de intervalos compuestos (Zeng, 1993,

PNAS 90: 10972-10976; y Zeng, 1994, Genetics 136: 1457-1468). QTL Cartographer permite el análisis de poblaciones F_2 o de retrocruzamiento. QTL Cartographer está disponible en statgen.ncsu.edu/qtlcart/cartographer.html (North Carolina State University). Otro programa que se puede utilizar es Qgene, que realiza el mapeo de QTL mediante regresión de un solo marcador o regresión por intervalos (Martinez y Curnow 1994 Heredity 73: 198-206). Utilizando Qgene, se pueden analizar múltiples tipos de población diferentes (todos derivados de la endogamia). Qgene está disponible de www.qgene.org. Aún otro programa es MapQTL, que realiza el mapeo de intervalos estándar (Lander y Botstein , Id.), el mapeo múltiple de QTL (MQM) (Jansen, 1993, Genetics 135: 205-211; Jansen, 1994, Genetics 138: 871-881) y el mapeo no paramétrico (prueba de suma de rango de Kruskal-Wallis). MapQTL puede analizar una diversidad de tipos de pedigrí, incluidos los pedigríes exógamos (polinizadores cruzados). MapQTL está disponible de Plant Research International, Plant Research International, P.O. Box 16, 6700 AA Wageningen, Holanda; www.plant.wageningen-ur.nl/default.asp?section=products). Aún otro programa que puede utilizarse en algunos ejemplos es Map Manager QT, que es un programa de mapeo de QTL (Manly y Olson, 1999, Mamm Genome 10: 327-334). Manager QT realiza análisis de regresión de un solo marcador, mapeo de intervalos simple basado en regresión (Haley y Knott, 1992, Heredity 69, 315-324), mapeo de intervalos compuestos (Zeng 1993, PNAS 90: 10972-10976), y ensayos de permutación. Una descripción de Map Manager QT es proporcionada por la referencia Manly y Olson, 1999, *Mammalian Genome* 10: 327-334.

Aún otro programa que se puede utilizar para realizar análisis de enlace es MultiCross QTL, que mapea QTL de cruces que proceden de líneas endogámicas. MultiCross QTL utiliza un enfoque de modelo de regresión lineal y maneja diferentes métodos, tales como mapeo de intervalos, mapeo de todos los marcadores y mapeo de QTL múltiple con cofactores. El programa puede manejar una amplia diversidad de poblaciones de mapeo simples para especies endogámicas y exogámicas. MultiCross QTL está disponible de Unite de Biometrie et Intelligence Artificielle, IRA, 31326 Castanet Tolosan, Francia.

Aún otro programa que se puede utilizar para realizar análisis de enlace es QTL Cafe. El programa puede analizar la mayoría de las poblaciones derivadas de cruces de líneas puras, tale como cruces F_2, retrocruces, líneas endogámicas recombinantes y líneas haploides duplicadas. QTL Cafe incorpora una implementación de Java de la regresión de marcadores flanqueantes de Haley & Knotts, así como la regresión de marcadores, y puede manejar múltiples QTL. El programa permite tres tipos de análisis QTL ANOVA de marcador único, regresión de marcadores (Kearsey y Hyne, 1994, Theor. Appl. Genet., 89: 698-702), y mapeo de intervalos por regresión (Haley y Knott, 1992, Heredity 69: 315-324). QTL Cafe está disponible en web.bham.ac.uk/g.g.seaton/.

Aún otro programa que puede utilizarse para realizar análisis de enlace es MAPL, que realiza análisis QTL mediante mapeo de intervalos (Hayashi y Ukai, 1994, Theor. Appl. Genet. 87:1021-1027) o análisis de varianza. Se pueden analizar diferentes tipos de población, incluyendo F_2, retrocruzamiento, endogamias recombinantes derivadas de F_2 o retrocruzamiento después de una determinada generación de autofecundación. Es posible agrupar y ordenar automáticamente numerosos marcadores mediante escalamiento multidimensional métrico. MAPL está disponible del Institute of Statistical Genetics on Internet (ISGI), Yasuo, UKAI, web.bham.ac.uk/g.g.seaton/.

Otro programa que se puede utilizar para el análisis de enlaces es R/qtl. Este programa proporciona un entorno interactivo para mapear QTLs en cruces experimentales. R/qtl hace uso de la tecnología oculta del modelo de Markov (HMM) para tratar los datos de genotipo que faltan. R/qtl ha implementado muchos algoritmos HMM, teniendo en cuenta la presencia de errores de genotipado, retrocruces, entrecruces y cruces de cuatro vías conocidos en fase. R/qtl incluye funcione para estimar mapas genéticos, identificar errores de genotipado y realizar exploraciones de genoma de QTL simple y exploraciones de genoma bidimensional de dos QTL, mediante mapeo de intervalos con regresión de Haley-Knott e imputación múltiple. R/qtl está disponible de Karl W. Broman, Johns Hopkins University, biosun01.biostat.jhsph.edu/.about.kbroman/qtl/.

El software basado en Java TASSEL (Análisis de Rasgos por Asociación, Evolución y Vinculación) se puede utilizar para determinar asociaciones de marcador rasgo. Véase Yu et al. (2005) *Nature Genetics* 38:203-208. TASSEL permite que se calculen y visualicen gráficamente las estadísticas de desequilibrio de enlace. TASSEL es capaz de fusionar datos de diferentes fuentes en un solo conjunto de datos de análisis, imputar datos que faltan utilizando un algoritmo k-vecinos más próximos (Cover y Hart (1967) *Proc IEEE Trans Inform Theory* 13) y realizar análisis de componentes principales (PCA) para reducir un conjunto de fenotipos correlacionados. El código fuente abierto para el paquete de software TASSEL está disponible en: sourceforge.net/projects/tassel.

TASSEL se puede utilizar con la prueba cuantitativa de desequilibrio de pedigrí endogámico (QIPDT). QIPDT es un ensayo de mapeo de asociación basada en la familia con líneas endogámicas de programas de mejora de plantas Véase Stich et al. (2006) *Theor Appl Genet* 113:1121-1130. QIPDT es un método de detección de QTL para datos recopilados de forma rutinaria en programas de fitomejora. QIPDT es un ensayo de asociación basado en la familia aplicables a la información genotípica de líneas parentales puras y la información genotípica y fenotípica de sus descendientes endogámicos. El QIPDT extiende el QPDT, un ensayo de asociación basado en la familia. Familias nucleares que consisten en dos líneas endogámicas parentales y al menos una línea endogámica de descendencia se pueden combinar para formar pedigríes extendidos, la base del QIPDT, si están relacionadas las líneas parentales de diferentes familias nucleares. QIPDT también tiene en cuenta la corrección de Martin et al. (2001) *Am J Hum Genet* 68:1065-1067 en relación con la prueba de desequilibrio de pedigrí.

También se puede utilizar el modelo de regresión mejorado QIPDT2. QIPDT2 adopta los mismos métodos para la codificación de marcadores y el ajuste fenotípico que se utilizan en QIPDT1, con dos mejoras: 1) se ajusta un modelo de regresión para el marcador y los datos fenotípicos, que permite estimar los efectos genéticos y las contribuciones fenotípicas para los marcadores en cuestión; 2) extender el enfoque a los híbridos de endogamias con diferentes probadores cultivados en múltiples ubicaciones, mientras que el enfoque original es aplicable solo a las endogamias. Dicha extensión se logra mediante la extracción de valores genéticos de endogamias de un modelo mixto que tiene en cuenta los efectos del probador y efectos no genéticos (p. ej., ubicaciones). QIPDT2 se describe en la Solicitud de Patente de EE.UU. Nº 12/328.689, presentada el 4 de diciembre de 2008.

Paquetes de software estadístico disponibles comercialmente adicionales que se utilizan comúnmente para hacer este tipo de análisis incluyen SAS Enterprise Miner (SAS Institute Inc., Cary, N.C.) y Splus (Insightful Corporation. Cambridge, Mass.). Los expertos en la técnica apreciarán que existen varios otros programas y algoritmos que se pueden utilizar en las etapas de los métodos de la presente invención, en donde se necesita un análisis genético cuantitativo, y todos esos programas y algoritmos están dentro del alcance de la presente invención.

Mapeo de Asociación Anidada

En diversas realizaciones, los marcadores candidatos se identifican o validan comparando marcadores correlacionados positivamente identificados utilizando GWA con marcadores correlacionados positivamente utilizando mapeo de asociación anidada (NAM) y seleccionando para uso adicional cualquier marcador que se muestre positivamente correlacionado utilizando ambos métodos. La estrategia de NAM aborda

la disección de rasgos complejos en un nivel fundamental mediante la generación de un recurso de mapeo común que permite a los investigadores explotar eficientemente las herramientas genéticas, genómicas y de biología de sistemas.

Basándose en los principios genéticos en estrategias y métodos de mapeo genómico anteriores (Meuwissen et al. 2002 *Genetics* 161: 373–379; Mott y Flint 2002, *Genetics* 160: 1609–1618; Darvasi y Shifman 2005, *Nat. Genet.* 37: 118–119), NAM tiene las ventajas de una menor sensibilidad a la heterogeneidad genética y una mayor potencia, así como una mayor eficiencia en el uso de la secuencia del genoma o marcadores densos, manteniendo una alta riqueza de alelos debido a diversos fundadores. NAM crea una población de mapeo integrada específicamente diseñada para una exploración completa del genoma con alta potencia para loci de rasgos cuantitativos (QTL) con efectos de diferentes tamaños.

El procedimiento en NAM implica primero seleccionar diversos fundadores y desarrollar un gran conjunto de progenies de mapeo relacionadas. En diversas realizaciones, las progenies relacionadas consiste en un conjunto de líneas endógamas recombinantes (RIL) derivadas de cruce entre un único parental común y un conjunto de diversas líneas fundadoras. Los RIL se desarrollan mediante múltiples rondas de autofecundación. El efecto de fondo genético de estos fundadores parentales en el mapeo de QTL individuales se minimiza sistemáticamente al reorganizar los genomas de los dos parentales de cada uno de los cruces durante el desarrollo de RIL, así como mediante el análisis combinado de todos los RIL a través de múltiples cruces. En general, la estrategia de proyectar información de secuencia, anidada dentro de marcadores informativos, desde los individuos más conectados hasta los individuos restantes, es aplicable a una amplia gama de especies, incluidos seres humanos, ratones, Arabidopsis y arroz.

Seguidamente, las líneas fundadoras se secuencian por completo o se genotipan densamente, y un número menor de marcadores de marcado tanto en los fundadores como en las progenies se genotipan para definir la herencia de segmentos cromosómicos y proyectar la información del marcador de alta densidad de los fundadores a la progenie. Las progenies son fenotipadas para diversos rasgos, y el genoma - el análisis de asociación amplia se lleva a cabo para relacionar rasgos fenotípicos con marcadores de alta densidad proyectados de las progenies. Véase Yu *et al.* 2008, Genetics 178:539-551.

Como en el mapeo de asociación general, la resolución de mapeo ofrecida por NAM depende en gran medida del desequilibrio de enlace entre los individuos fundadores. Estudios empíricos con genes candidatos de maíz secuenciados a través de diversas líneas han demostrado una rápida descomposición de LD de más de 2000 pb (Wilson et al. 2004, Plant Cell 16: 2719-2733). Recientes análisis del genoma completo en diversos accesos de Arabidopsis (Nordborg et al. 2005, *PLoS Biol.* 3**:** e196) y razas de perros (*Canis familiaris*) (Lindblad-Toh et al. 2005, *Nature* 438**:** 803–819) concuerdan con este patrón. LD se descompone rápidamente a través de germoplasma genéticamente diverso. Con la estrategia NAM, esta ventaja en la resolución se utiliza plenamente sin el inconveniente acoplado - la necesidad de buenos genes candidatos o un gran número de marcadores - al proyectar la información genómica de los fundadores a las RILs.

Modelos para NAM

En la presente invención, la estrategia NAM para identificar o validar marcadores candidatos emplea modelos de regresión para detectar una asociación entre un rasgo de interés y un marcador. En estadística, el análisis de regresión es un nombre colectivo para técnicas para el modelado y el análisis de datos numéricos que consisten en valores de una variable dependiente (variable de respuesta) y de una o más variables independientes (variables

explicativas). La variable dependiente en la ecuación de regresión se modela como una función de las variables independientes, los parámetros correspondientes ("constantes") y un término de error. El término de error se trata como una variable aleatoria. Representa una variación inexplicable en la variable dependiente. Los parámetros se estiman para dar un "mejor ajuste" de los datos. Más comúnmente, el mejor ajuste se evalúa utilizando el método de mínimos cuadrados, pero también se han utilizado otros criterios.

Los mínimos cuadrados se pueden interpretar como un método de ajuste de datos. El mejor ajuste en el sentido de mínimos cuadrados es el caso del modelo para el cual la suma de los residuos cuadrados tiene su menor valor, siendo un residuo la diferencia entre un valor observado y el valor dado por el modelo. Los mínimos cuadrados corresponden al criterio de máxima probabilidad si los errores experimentales tienen una distribución normal y también pueden derivarse como un método de estimador de momentos. El análisis de regresión está disponible en la mayoría de los paquetes de software estadísticos.

Para los fines de la presente invención, se puede utilizar cualquier método de regresión adecuado para identificar QTL en una población anidada. Ejemplos de modelos de regresión se describen en esta memoria. Además se proporcionan dos nuevos modelos de regresión (SMR y MMR) que pueden utilizarse para identificar, validar o priorizar para el uso posterior de un marcador asociado con un rasgo de interés.

Regresión sobre un Único Marcador (SMR):

En esta memoria se proporciona una nueva herramienta de regresión de único marcador (SMR) para realizar el mapeo de asociación anidada. El método es similar a la regresión de único marcador utilizada en el análisis de vinculación QTL estándar, con dos modificaciones clave. Una es que la información de fondo poligénica se incorpora de cada una de las subpoblaciones en el modelo. Al hacerlo, la variación genética provocada por diferentes antecedentes genéticos se puede separar del modelo, mejorando así la potencia de mapeo de QTL. Al mismo tiempo, la inclusión de la información de fondo genético elimina el efecto de estratificación de la población en el mapeo de QTL, minimizando falsas tasas de descubrimiento positivos. La segunda mejora sobre los métodos existentes es la exclusión de datos de marcadores de poblaciones distorsionadas. Esta característica permite que el modelo evite la influencia de la distorsión de segregación de marcadores en la detección de QTL, lo que puede crear desafíos en el mapeo de la asociación. Este modelo se beneficia adicionalmente del diseño experimental de NAM, que es una combinación de mapeo de vinculación y asociación. La presente invención utiliza un modelo lineal único para describir la relación entre los valores de rasgos y los genotipos de marcadores que se pueden escribir como:

$$y_{ij} = \mu + x_{ij}a + g_i u_i + e_{ij}$$

en que y_{ij} es el valor fenotípico del j individual en la subpoblación i; μ es la media general; a es el efecto aditivo de QTL; g_i es la variable indicadora de la subpoblación i; u_i es el efecto de la subpoblación i; e_{ij} es el error residual que se supone que sigue una distribución normal con ceros medios y varianza σ^2. De acuerdo con la presente invención, el genotipo x_{ij} se define como 1 si el j individual lleva el alelo del parental común y -1 si el j individual lleva el alelo del otro parental. Esta definición se basa en que solo hay dos alelos distintos para cada uno de los marcadores. Para explotar la simplicidad de regresión, el efecto de fondo genético u_i se supone que es un efecto fijo. Tal como se utiliza en esta memoria, la expresión "efectos fijos" preferiblemente se refiere a influencias estacionales, espaciales, geográficas, ambientales o de gestión que provocan un efecto sistemático sobre el fenotipo o sobre aquellos efectos con niveles que fueron deliberadamente dispuestos por el experimentador, o el efecto de un gen o marcador que es consistente en toda la población evaluada. Por lo tanto, la invención incluye el efecto de fondo genética u_i en el modelo para tener en cuenta la influencia de la estratificación de la población y, por lo tanto, reducir la varianza residual.

Este método SMR difiere del método de regresión original basado en marcadores para NAM (Yu *et al* (2006) *Nature Genetics* 38(2):203-208) en el uso de marcadores polimórficos. A partir de los datos de marcadores NAM, se ve fácilmente que algunos marcadores muestran polimorfismo en algunas subpoblaciones, pero no en otras. En este caso, la inclusión de marcadores no informativos puede conducir a la distorsión de la segregación de los genotipos del marcador en ese locus, y la distorsión podría causar la reducción de la eficiencia, potencia y precisión del de mapeo de QTL. Para evitar el problema, la presente invención utiliza un procedimiento filtrado por marcadores, incorporado en al modelo SMR para reducir el riesgo potencial debido a la distorsión del marcador. Este procedimiento filtrado por marcadores significa que solo los datos fenotípicos y genotípicos de esas subpoblaciones con genotipos segregados de un marcador se incluyen en cada uno de los análisis. Por lo tanto, en la invención, las subpoblaciones con genotipos no informativos se excluyen antes del análisis SMR. El procedimiento permite a SMR identificar aquellos alelos con muy baja frecuencia (menos del 5%) en NAM.

Mapeo de Intervalos Compuestos

Cuando están presentes múltiples QTL vinculados, los métodos de intervalo y marcador único actuales colocan a menudo a QTL en la ubicación incorrecta, por ejemplo generando un QTL fantasma en la posición entre los dos QTL reales. Un enfoque para tratar con múltiples QTLs es modificar el mapeo de intervalos estándar para incluir marcadores adicionales como cofactores (también denominados en esta memoria "covariables") en el análisis. En general, el uso de cofactores reduce el sesgo y el error de muestreo de las posiciones estimadas de QTL (Utz y

Melchinger, Biometrics in Plant Breeding, Proceedings of the Ninth Meeting of the Eucarpia Section Biometrics in Plant Breeding, Holanda, 1994). Utilizando los marcadores no vinculados apropiados se puede explicar en parte la varianza de segregación generada por los QTLs no vinculados, mientras que los efectos de los QTLs vinculados se pueden reducir al incluir marcadores vinculados al intervalo de interés. Este enfoque general de añadir cofactores marcadores a un análisis de intervalos estándar, a menudo denominado "mapeo de intervalos compuestos" o CIM, da como resultado aumentos sustanciales en la potencia para detectar un QTL y en la precisión de las estimaciones de la posición de QTL.

CIM maneja múltiples QTL incorporando información de marcadores multilocus de organismos modificando el mapeo de intervalos estándar para incluir marcadores adicionales como cofactores para el análisis. En estos métodos se realiza un mapeo de intervalos utilizando un subconjunto de loci marcadores como covariables. Estos marcadores sirven como representantes de otros QTLs para aumentar la resolución del mapeo de intervalos, contabilizando los QTLs vinculados y reduciendo la variación residual. Modelos de CIM a modo de ejemplo se describen, por ejemplo, en Jansen, 1993, Genetics 135, pág. 205; Zeng, 1994, Genetics 136, pág. 1457.

Se pueden utilizar modelos adicionales. Se ha informado de muchas modificaciones y enfoques alternativos para el mapeo de intervalos, incluido el uso de métodos no paramétricos (Kruglyak y Lander, Genetics, 121:1421-1428, 1995). También se pueden utilizar métodos o modelos de regresión múltiple, en los cuales el rasgo es regresado en un gran número de marcadores (Jansen et al., Theor. Appl. Genet., 91:33-37, 1995; Weber y Wricke, Advances in Plant Breeding, Blackwell, 1994).

Regresión de marcadores múltiples (MMR)

Para tener en cuenta las influencias de otros QTL, se describe en esta memoria un nuevo método de regresión de marcadores múltiples (MMR). Este método utiliza marcadores de cofactor para absorber el efecto de otros QTLs. El modelo lineal para MMR es:

$$y_{ij} = \mu + x_{ij}a + \Sigma(k=1, m)c_{ijk}b_k + g_iu_i + e_{ij}$$

en que y_{ij} es el valor fenotípico del *j* individual en la subpoblación *i*; μ es la media general; x_{ij} es el genotipo de QTL; *a* es el efecto aditivo de QTL; c_{ijk} es el marcador de cofactor *k* para el *j* individual en la subpoblación *i*, y b_k es el efecto del marcador de cofactor *k*; g_i es la variable indicadora de la subpoblación *i*; u_i es el efecto de la subpoblación *i*, y e_{ij} es el error residual que se supone que sigue una distribución normal con ceros medios y varianza σ^2. Este modelo MMR es similar al mapeo de intervalos compuesto (Zeng 1993, 1994, *infra*).

El problema clave con CIM ha sido la elección de loci marcadores adecuados para servir como covariables; una vez que se han elegido, CIM convierte el problema de selección del modelo en un escaneo unidimensional. Antes de la presente invención, la elección de cofactores marcadores no se había resuelto. En la presente invención, la regresión escalonada se utiliza para seleccionar marcadores de cofactor basados en el nivel de significancia 0,01. El modelo lineal utilizado para elegir cofactores es:

$$y_{ij} = \mu + x_{ij}a + c_{ijk}b_k + g_iu_i + e_{ij.}$$

Este modelo de regresión gradual es diferente del utilizado para el mapeo de intervalos compuesto convencional (Zeng 1993, 1994) y el utilizado originalmente para NAM (Yu et al 2008). En el presente modelo MMR, la regresión escalonada se utiliza para una población de NAM con la inclusión de las bases genéticas de diferentes subpoblaciones en el modelo. Este método de selección se centra en seleccionar aquellos QTLs que tienen efectos estables a través de múltiples subpoblaciones. Por lo tanto, reduce efectivamente el número de cofactores incluidos en el modelo, evitando el problema de la sobresaturación.

Con los marcadores de cofactor es posible obtener un perfil LOD mucho más claro de MMR que SMR. El uso de marcadores de cofactor es para reducir el error residual y, por lo tanto, aumentar la importancia del ensayo de hipótesis QTL. El nuevo modelo MMR proporcionado en esta memoria muestra la capacidad de separar QTL estrechamente vinculado y localizar un QTL dentro de una región genómica estrecha. En diversas realizaciones, todos los datos genotípicos de todas las subpoblaciones se utilizan para el análisis de datos.

Es esperar que SMR y MMR proporcionarán resultados similares para esos marcadores sin segregación distorsionada, mientras que pueden mostrar diferencias en marcadores con la segregación genotípica sesgada. Por lo tanto, tanto SMR como MMR se realizan en combinación de complemento para el conjunto de datos NAM. SMR y MMR se pueden realizar por separado para el mismo conjunto de datos fenotípicos de rasgos y datos de marcadores. Después se pueden comparar los resultados obtenidos de cada uno de los métodos. Para aquellos QTL no identificados consistentemente tanto por SMR como por MMR, se puede realizar el análisis de segregación de marcadores. Este análisis se puede realizar para determinar si la inconsistencia de los QTL es provocada por la distorsión del marcador. La distorsión del genotipo marcador puede dar como resultado la falta de QTL verdadero (falso negativo), o puede resultar en la detección de QTL falso si el genotipo marcador se correlaciona con la tendencia del rasgo. Para aquellos QTL identificados consistentemente por SMR y MMR, el análisis de segregación de marcadores puede no ser necesario. Sin embargo, el uso conjunto de SMR y MMR bajo cualquier circunstancia probablemente resulte en QTL con una potencia mejorada y una tasa de falsos positivos disminuida. Por lo tanto, en

este aspecto de la invención, los marcadores correlacionados positivamente se consideran identificadas tanto por SMR como por MMR.

Ensayo del efecto de QTL

A menudo, el objetivo de un estudio de asociación no es simplemente detectar asociaciones de marcador/rasgo, sino estimar la ubicación de los genes que afectan el rasgo directamente (es decir, QTLs) en relación con las ubicaciones de los marcadores. En un enfoque simple para este objetivo, se hace una comparación entre los loci marcadores de la magnitud de la diferencia entre genotipos alternativos o el nivel de importancia de esa diferencia. Se infiere que los genes de rasgos se ubican más cerca del o de los marcadores que tienen la mayor diferencia genotípica asociada. En un análisis más complejo, como el mapeo de intervalos (Lander y Botstein, Genetics 121:185-199, 1989), cada una de las muchas posiciones a lo largo del mapa genético (por ejemplo, a intervalos de 1 cM) se somete a ensayo la probabilidad de que un QTL esté ubicado en esa posición. Los datos de genotipo/fenotipo se utilizan para calcular cada una de las posiciones de ensayo una puntuación LOD (log de la relación de probabilidad). Cuando la puntuación LOD

excede de un valor umbral crítico, hay una evidencia significativa de la ubicación de un QTL en esa posición en el mapa genético (que se caerá entre dos loci de marcador particulares).

Las hipótesis para probar el efecto QTL se pueden formular como H_0 $a = 0$ y H_1 $a_1 \neq 0$. Los parámetros bajo H_0 o H_1 se estiman por el método de mínimos cuadrados según el modelo de regresión dependiendo de si el efecto QTL está incluido en el modelo. A continuación, se puede obtener la relación de probabilidad (LR). La relación de probabilidad es la relación de la probabilidad máxima de un resultado bajo dos hipótesis diferentes. Un ensayo de relación de probabilidad es un ensayo estadístico para tomar una decisión entre dos hipótesis en base al valor de esta relación. Siendo una función de los datos x, la LR es, por lo tanto, una estadística. El ensayo de la relación de probabilidad rechaza la hipótesis nula si el valor de esta estadística es demasiado pequeño. Cuán pequeño es demasiado pequeño depende del nivel de significancia del ensayo, es decir, de qué probabilidad de error de Tipo I se considera tolerable (los errores de "Tipo I" consisten en el rechazo de una hipótesis nula que es verdadera).

Valores más bajos de la relación de probabilidad significan que es menos probable que ocurra el resultado observado bajo la hipótesis nula. Los valores más altos significan que es más probable que ocurra el resultado observado bajo la hipótesis nula. La LR se puede obtener de los modelos de regresión tales como $LR = -2(\ell_{reducida} - \ell_{completa})$, en que $\ell_{reducida}$ es la probabilidad logarítmica del modelo reducido, correspondiente a H_0, y $\ell_{completa}$ es la del modelo completo, correspondiente a H_1 (Lander y Botstein 1989).

A partir de la LR, se calcula un logaritmo de la puntuación de probabilidades (LOD). Una puntuación LOD es una estimación estadística de si es probable que dos loci se encuentren cerca uno del otro en un cromosoma y, por lo tanto, es probable que estén genéticamente vinculados. En el presente caso, una puntuación LOD es una estimación estadística de si una posición dada en el genoma en estudio está vinculada al rasgo cuantitativo correspondiente a un gen dado. En un ejemplo, la puntuación LOD se calcula como $LR/(2 \ln 10)$. La puntuación LOD indica esencialmente cuánto más probable es que hayan surgido los datos suponiendo la presencia de unos QTL en comparación con su ausencia. El valor umbral de LOD para evitar un falso positivo con una confianza dada, digamos 95%, depende del número de marcadores y de la longitud del genoma. Gráficos que indican los umbrales LOD se recogen en Lander y Botstein, Genetics, 121:185-199 (1989), y se describen más detalladamente por Ars y Moreno-Gonzalez, Plant Breeding, Hayward, Bosemark, Romagosa (eds.) Chapman & Hall, Londres, págs. 314-331 (1993).

En general, una puntuación LOD de tres o más sugiere que dos loci están genéticamente vinculados, una puntuación LOD de 4 o más es una fuerte evidencia de que dos loci están genéticamente vinculados,

y una puntuación LOD de 5 o más es una evidencia muy fuerte de que dos loci están genéticamente vinculados. Sin embargo, la importancia de cualquier puntuación LOD en realidad varía de una especie a otra dependiendo del modelo utilizado.

Ensayos de permutación para NAM

El método original de regresión múltiple para NAM (Yu et al 2008) utilizó un nivel de significancia muy bajo 10^{-7} como umbral para la detección de QTL. Este método no es apropiado para determinar el umbral de LOD en un nivel de significancia dado, especialmente basado en un mapa de enlace denso. Para resolver este problema, la presente invención proporciona un método nuevo de ensayo de permutación para determinar el umbral empírico de LOD en el nivel de significancia dado de 0,05 y 0,01. El método de permutación de la invención reordena los valores fenotípicos dentro de cada una de las subpoblaciones sin destruir la estructura de las subpoblaciones y la correlación entre los diferentes rasgos de interés. Para lograr esto, SMR y MMR se realizan en los datos fenotípicos aleatorios y los datos del marcador original, y luego se calcula la puntuación LOD máxima en todos los marcadores en el genoma. Este tipo de análisis se repite 1000 veces y se registra la puntuación máxima de LOD de cada uno de los análisis. Finalmente, estas puntuaciones de LOD se clasifican en orden ascendente. El valor de LOD en la posición $(1 - \alpha) * n$ es el umbral empírico de LOD en el nivel de significancia α. En algunos ejemplos, el umbral de 0,01 puede no ser estable debido al número limitado de ensayos de permutación. Por lo tanto, se recomiendan 10000 permutaciones a este nivel de significancia. Sin embargo, se entiende que es posible un número diferente de permutaciones y todavía

se obtiene el nivel de significancia deseado. Por ejemplo, se pueden realizar aproximadamente 2000, aproximadamente 3000, aproximadamente 4000, aproximadamente 5000, aproximadamente 6000, aproximadamente 7000, aproximadamente 8000, aproximadamente 9000 o más permutaciones.

Análisis de expresión de QTL

Otro enfoque abarcado por la presente invención para priorizar genes candidatos para aplicaciones posteriores es la combinación de técnicas GWA y DGE (Expresión Digital de Genes) para priorizar aún más los genes para su implementación o validación a través de la resolución de eQTL. Determinadas plataformas de descubrimiento/genotipado de marcadores están diseñadas de tal manera que proporcionen marcadores suficientes para GWA junto con el Perfil de Expresión de cada uno de los marcadores genotipados (p. ej., la plataforma de descubrimiento/genotipado SNP Solexa).

Por lo tanto, los análisis QTL clásicos se combinan con el perfil de expresión génica, es decir, mediante micromatrices de ADN. QTLs de expresión (e-QTL) de este tipo describen elementos de control cis y trans para la expresión de genes asociados con un rasgo de interés. Estos métodos son capaces de determinar la relación entre los marcadores en el mapa de enlace y la expresión de uno o más marcadores para identificar QTLs estadísticamente significativos. La expresión se puede monitorear y correlacionar con el rasgo de interés bajo una diversidad de condiciones, tales como la fase de desarrollo, la exposición ambiental y similares. Una relación de este tipo puede determinarse utilizando cualquier método de asociación descrito en esta memoria o conocido por un experto en la técnica, por ejemplo, pero no limitado a, ANOVA de un solo punto, regresión simple, mapeo de intervalos, mapeo de intervalos compuestos y NAM.

Por lo tanto, el análisis eQTL comienza con datos de expresión génica (p. ej., de un estudio de expresión génica o un estudio proteómico) y datos de genotipo de una población en estudio. En un aspecto de la presente invención, el nivel de expresión de un gen en un organismo en la población de interés se determina midiendo una cantidad de al menos un constituyente celular que corresponde al gen en una o más células del organismo. Tal como se utiliza en esta memoria, la expresión "constituyente celular" comprende genes individuales, proteínas, ARNm que expresa un gen y/o cualquier otro componente celular variable o actividad proteica, grado de modificación de la proteína (p. ej., fosforilación), por ejemplo, que se mide típicamente en un experimento biológico realizado por los expertos en la técnica.

El nivel de expresión de una secuencia de nucleótidos en un gen puede medirse mediante cualquier técnica de alto rendimiento. Sin embargo, medido, el resultado es la cantidad absoluta o relativa de transcripciones o datos de respuesta, incluidos, pero no limitados a valores que representan abundancias o relaciones de abundancia. La medición del perfil de expresión se puede realizar mediante hibridación con matrices de transcripción (p. ej. , "matrices de transcritos" o "matrices de perfilado") . Las matrices de transcritos se pueden emplear para analizar el perfil de expresión en una muestra celular y especialmente para medir el perfil de expresión de una muestra celular de un tipo de tejido particular o fase de desarrollo o un tipo celular expuesto a una condición ambiental particular.

Los datos de expresión se transforman en una estadística de expresión que se utiliza para tratar la abundancia de cada constituyente celular en los datos de expresión génica como un rasgo cuantitativo. Luego, para cada uno de los genes en una pluralidad de genes expresados por un organismo en la población, se realiza un análisis de loci de rasgos cuantitativos (QTL) utilizando el mapa de marcadores genéticos para producir datos de QTL. Un conjunto de estadísticas de expresión representa el rasgo cuantitativo utilizado en cada uno de los análisis de QTL.

Las estadísticas de expresión comúnmente utilizadas como rasgos cuantitativos en los análisis incluyen, pero no se limitan a la relación logarítmica media, la intensidad logarítmica y la intensidad corregida de fondo. Otros tipos de estadísticas de expresión también pueden utilizarse como rasgos cuantitativos. Por ejemplo, la transformación puede realizarse utilizando un módulo de normalización. En este tipo de ejemplos, el nivel de expresión de una pluralidad de genes en cada uno de los organismos en estudio está normalizado. Se puede utilizar cualquier rutina de normalización. Rutinas de normalización representativas incluyen, pero no se limitan a puntuación Z de intensidad, intensidad mediana, log de intensidad mediana, puntuación Z de log de desviación estándar de intensidad, desviación absoluta media de la puntuación Z del conjunto de genes de ADN de calibración log, corrección de intensidad mediana de relación y corrección de fondo de intensidad. Además, se pueden ejecutar combinaciones de rutinas de normalización.

En la última década, varias tecnologías han permitido monitorizar el nivel de expresión de un gran número de transcripciones en cualquier momento (véase, p. ej., Schena et al., 1995, Science 270:467-470; Lockhart et al., 1996, Nature Biotechnology 14:1675-1680; Blanchard et al., 1996, Nature Biotechnology 14, 1649; Pat. de EE.UU. Nº 5.569.588). Por ejemplo, la expresión se puede medir utilizando la expresión digital de genes (DGE). DGE proporciona un análisis cuantitativo, global y libre de hipótesis del transcriptoma completo. Esta solicitud analiza el nivel de expresión de prácticamente todos los genes en una muestra mediante el recuento del número de moléculas de ARNm individuales producidas a partir de cada uno de los genes. No existe requisito de que los genes puedan identificarse y caracterizarse antes de la realización de un experimento. Plataformas DGE están disponibles comercialmente, por ejemplo, a través de Helicos Biosciences (Cambridge, MA) e Illumina, Inc. (San Diego, CA).

La capacidad de realizar análisis de polimorfismo de un solo nucleótido (SNP) en el genoma completo ha hecho posible los estudios de GWA para la identificación de variantes de rasgos comunes. Estudios del genoma completo del epigenoma, o información no basada en la secuencia heredada durante la división celular, se ha quedado atrás. Parte de la razón es la diversa naturaleza de los elementos de control epigenéticos, tales como la metilación del ADN y las modificaciones de cromatina múltiple. El análisis estándar de expresión génica indiscriminada basada en alelos puede revelar cambios epigenéticos en genes individuales, o simplemente puede reflejar cambios dinámicos en la expresión génica mediada por componentes reguladores que actúan en *trans*, tales como factores de transcripción. La capacidad de discriminar la expresión específica de alelos (ASE) de los dos alelos de genes puede revelar cambios en el control epigenético, ya que los dos alelos están afectados por los mismos factores de transcripción, aunque diferirían en los elementos de control que actúan en *cis*.

Por lo tanto, el análisis eQTL abarca la evaluación de la expresión específica para alelos. En principio, los métodos estándar de QTL o Asociación de Marcadores vinculan un segmento discreto de ADN, tal como un haplotipo, con un porcentaje de varianza fenotípica en algún nivel de importancia. Habitualmente, el fenotipo es una medida cuantificable del comportamiento de la planta, tal como el rendimiento. Del mismo modo, un análisis de eQTL considera que la expresión génica es un fenotipo cuantificable que puede asociarse con un segmento discreto de ADN. Métodos de este tipo se utilizan para vincular los patrones de expresión específicos para una ubicación específica en el genoma, pero no justifican las secuencias de actuación cis/trans o influencias epigenéticos en la expresión génica.

Queda abarcado en esta memoria un método por el cual la expresión cuantificable de cada uno de los genes de cada uno de los individuos de una población definida se subdivide en intervalos de valores de expresión basados en el haplotipo. Por ejemplo, si el Gen ABC tiene 8 haplotipos, entonces a cada uno de los haplotipos se le asigna un intervalo de expresión basado en la expresión colectiva de cada uno de los haplotipos a través de cada uno de los individuos de la población. Posteriores análisis de asociación se pueden realizar entonces entre la expresión del haplotipo y tanto la secuencia del haplotipo como el fenotipo cuantificable.

En términos generales, los resultados de este tipo de análisis revelan uno de tres patrones: (1) cada uno de los haplotipos de un solo gen tiene su propio intervalo de expresión único, que puede indicar la expresión del gen específico para el alelo que actúa en cis; (2) cada uno de los haplotipo de un solo gen tiene el mismo intervalo de expresión, lo que puede indicar una regulación conservada del gen en cuestión; o (3) un haplotipo específico de un solo gen tiene múltiples intervalos de expresión, lo que puede indicar una expresión específica de alelo de acción en trans o regulación epigenética.

En algunos casos, este tipo de análisis puede proporcionar una confirmación independiente de una asociación de un haplotipo de genes con un rasgo de interés. Por ejemplo, si un haplotipo específico se asocia con un rendimiento incrementado y un valor de expresión específico del mismo haplotipo, también se asocia con un rendimiento incrementado, existe una indicación más fuerte de que el haplotipo está asociado con el rasgo de interés.

Alternativamente, o además, este análisis puede facilitar la identificación y asociación de influencias epigenéticas o específicas para el alelo cis/trans sobre un rasgo de interés. Por ejemplo, en condiciones normales, cada uno de los haplotipos únicos de un solo gen tiene el mismo intervalo de expresión. En tales casos, cualquier asociación de un haplotipo específico a un valor específico del rasgo de interés (p. ej., rendimiento incrementado en una planta) puede atribuirse a la variación del ADN en ese locus. Alternativamente, cada uno de los haplotipos podría tener su o sus propios intervalos de expresión única. En tales casos, la asociación de un haplotipo específico y un intervalo de expresión solo o en combinación con un rendimiento incrementado podría atribuirse a influencias epigenéticas o específicas para alelos cis/trans sobre el rendimiento de la planta.

Métodos para examinar la ASE se describen, por ejemplo, en Lo et al. (2003) *Genome Res.*13(8):1855-62; Pant et al. (2006) *Genome Res.* 16(3):331-9; y Bjornsson *et al.* (2008) *Genome Research* 18:771–779.

Métodos Implementados en Computadora

Los métodos arriba descritos para evaluar la asociación marcador:rasgo se pueden realizar, total o parcialmente, con el uso de un programa informático o un método implementado por computadora. Los programas de computadora están configurados adecuadamente para realizar las operaciones descritas en esta memoria.

Los programas de computadora y los productos de programas de computadora de la presente invención comprenden un medio utilizable por computadora que tiene una lógica de control almacenada en el mismo para hacer que una computadora ejecute los algoritmos descritos en esta memoria. Sistemas informáticos de la presente invención comprenden un procesador, operativo para determinar, aceptar, verificar y visualizar datos, una memoria para almacenar datos acoplados a dicho procesador, un dispositivo de visualización acoplado a dicho procesador para visualizar datos, un dispositivo de entrada acoplado a dicho procesador para introducir datos externos; y un script legible por computadora con al menos dos modos de operación ejecutables por dicho procesador. Un script legible por computadora puede ser un programa de computadora o lógica de control de un producto de programa de computadora de una realización de la presente invención.

No es crítico para la invención que el programa informático se escriba en cualquier lenguaje informático particular o que opere en cualquier tipo particular de sistema informático o sistema operativo. El programa informático se puede escribir, por ejemplo, en lenguaje de programación C++, Java, Perl, Python, Ruby, Pascal o Basic. Se entiende que se puede crear un programa de este tipo en uno de los muchos lenguajes de programación diferentes. En un aspecto de esta invención, este programa está escrito para funcionar en una computadora que utiliza un sistema operativo Linux. En otro aspecto de esta invención, el programa está escrito para funcionar en una computadora que utiliza un sistema operativo MS Windows o MacOS.

Un experto en la técnica entenderá que los códigos pueden realizarse en cualquier orden, o simultáneamente, de acuerdo con la presente invención, siempre que el orden siga un flujo lógico.

Uso intermedio de marcadores

Los marcadores identificados o validados utilizando los métodos descritos en esta memoria pueden utilizarse para técnicas de diagnóstico y selección basadas en el genoma; para rastrear la progenie de un organismo; para determinar la hibridación de un organismo; para identificar la variación de rasgos fenotípicos vinculados, rasgos de expresión de ARNm o rasgos de expresión tanto fenotípicos como de ARNm; como marcadores genéticos para construir mapas de enlace genético; para identificar la progenie individual de un cruce en donde la progenie tiene una contribución genética deseada de un donante parental, parental receptor, o tanto el donante parental como el parental receptor; para aislar la secuencia de ADN genómico que rodea una secuencia de ADN codificante o no codificante, por ejemplo, pero no limitada a un promotor o una secuencia reguladora; en la selección asistida por marcadores, la clonación basada en mapas, la certificación híbrida, huellas digitales, genotipado y marcadores específicos para alelos; para el desarrollo de plantas transgénicas; y como un marcador en un organismo de interés.

La motivación principal para desarrollar tecnologías de marcadores moleculares desde el punto de vista de los obtentores ha sido la posibilidad de aumentar la eficiencia de mejora a través de la mejora asistida por marcadores. Después de que se hayan identificado marcadores positivos a través de los modelos estadísticos arriba descritos, los alelos marcadores genéticos correspondientes se pueden utilizar para identificar plantas que contienen el genotipo deseado en múltiples loci y sería de esperar que transfieran el genotipo deseado junto con el fenotipo deseado a su progenie. Un alelo marcador molecular que demuestra un desequilibrio de enlace con un rasgo fenotípico deseado (p. ej., un locus de rasgo cuantitativo o QTL) proporciona una herramienta útil para la selección de un rasgo deseado en una población de plantas (es decir, mejora asistida por marcadores).

Un "locus marcador" es un locus que puede utilizarse para rastrear la presencia de un segundo locus vinculado, p. ej., un locus vinculado que codifica o contribuye en la expresión de un rasgo fenotípico. Por ejemplo, un locus marcador puede utilizarse para controlar la segregación de alelos en un locus, tal como un QTL, que están genética o físicamente vinculado al locus marcador. Por lo tanto, un "alelo marcador", alternativamente un "alelo de un locus marcador" es uno de una pluralidad de secuencias de nucleótidos polimórficos que se encuentran en un locus marcador en una población que es polimórfica para el locus marcador. En algunos aspectos, la presente invención proporciona métodos para la identificación o validación de loci marcadores correlacionados con un rasgo fenotípico de interés. Se espera que cada uno de los marcadores identificados se encuentre en una estrecha proximidad física y genética (que resulte en un enlace físico y/o genético) con un elemento genético, p. ej., un QTL que contribuya al rasgo de interés.

En diversas realizaciones de la presente invención, los marcadores que se identifican utilizando los métodos descritos en esta memoria se utilizan para seleccionar plantas y enriquecer la población de plantas para individuos que tienen rasgos deseados. El obtentor puede utilizar ventajosamente marcadores moleculares para identificar a los individuos deseados mediante la identificación de alelos marcadores que muestran una probabilidad estadísticamente significativa de co-segregación con un fenotipo deseado. Mediante la identificación y la selección de un alelo del marcador (o alelos deseados de múltiples marcadores) que está optimizado para el fenotipo deseado, el obtentor es capaz de seleccionar rápidamente un fenotipo deseado mediante la selección para el alelo marcador molecular apropiado.

La presencia y/o ausencia de un alelo marcador genético particular en el genoma de una planta que exhibe un rasgo fenotípico preferido se determina mediante cualquier método arriba mencionado, p. ej., RFLP, AFLP, SSR, amplificación de secuencias variables y ASH. Si los ácidos nucleicos de la planta se hibridan con una sonda específica para un marcador genético deseado, la planta puede ser autofecundada para crear una verdadera línea de reproducción con el mismo genoma o puede ser introgresada en una o más líneas de interés. El término "introgresión" se refiere a la transmisión de un alelo deseado de un locus genético de un fondo genético a otro. Por ejemplo, la introgresión de un alelo deseado en un locus específico puede transmitirse a al menos una progenie a través de un cruce sexual entre dos parentales de la misma especie, en que al menos uno de los parentales tiene el alelo deseado en su genoma. Alternativamente, por ejemplo, la transmisión de un alelo puede ocurrir por recombinación entre dos genomas donantes, p. ej., en un protoplasto fusionado, en donde al menos uno de los protoplastos donantes tiene el alelo deseado en su genoma. El alelo deseado puede ser, p. ej., un alelo seleccionado de un marcador, un QTL, un transgen o similar. En cualquier caso, la descendencia que comprende el alelo deseado puede retrocruzarse repetidamente a una línea que tiene un fondo genético deseado y puede seleccionarse para el alelo deseado, para dar como resultado que el alelo se fije en un fondo genético seleccionado.

Los loci marcadores identificados o validados utilizando los métodos de la presente invención también se pueden utilizar para crear un mapa genético denso de marcadores moleculares. Un "mapa genético" es una descripción de las relaciones de enlace genético entre loci en uno o más cromosomas (o grupos de enlace) dentro de una especie dada, generalmente representada en forma de diagrama o tabla. El "mapeo genético" es el proceso de definir las relaciones de enlace de loci mediante el uso de marcadores genéticos, poblaciones segregantes para los marcadores y principios genéticos estándar de frecuencia de recombinación. Una "ubicación del mapa genético" es una ubicación en un mapa genético con relación a los marcadores genéticos circundantes en el mismo grupo de enlace en donde se puede encontrar un marcador específico dentro de una especie dada. Por el contrario, un mapa físico del genoma se refiere a distancias absolutas (por ejemplo, medidas en pares de bases o fragmentos genéticos contiguos aislados y solapantes, p. ej., contigs). Un mapa físico del genoma no tiene en cuenta el comportamiento genético (p. ej., frecuencias de recombinación) entre diferentes puntos en el mapa físico.

En determinadas aplicaciones, es ventajoso producir o clonar ácidos nucleicos grandes para identificar ácidos nucleicos unidos más distantemente a un marcador dado, o aislar ácidos nucleicos unidos o responsables de QTL tal como se identifica en esta memoria. Se apreciará que un ácido nucleico genéticamente unido a una secuencia de nucleótidos polimórficos opcionalmente reside hasta aproximadamente 50 centimorgans del ácido nucleico polimórfico, aunque la distancia precisa variará dependiendo de la frecuencia de cruce de la región cromosómica particular. Distancias típicas de un nucleótido polimórfico están en el intervalo de 1-50 centimorgans, por ejemplo, a menudo menos de 1 centimorgan, menos de aproximadamente 1-5 centimorgans, aproximadamente 1-5, 1, 5, 10, 15, 20, 25, 30, 35, 40, 45 o 50 centimorgans, etc.

Se conocen muchos métodos para producir ácidos nucleicos ARN y ADN recombinantes grandes, incluidos plásmidos recombinantes, fagos lambda recombinantes, cósmidos, cromosomas artificiales de levadura (YACs), cromosomas artificiales P1, cromosomas artificiales bacterianos (BAC) y similares. Una introducción general a los YACs, BACs, PACs y MACs como cromosomas artificiales se describe en Monaco & Larin, Trends Biotechnol. 12:280-286 (1994). Ejemplos de técnicas de clonación apropiadas para producir ácidos nucleicos grandes e instrucciones suficientes para dirigir a personas expertas a través de muchos ejercicios de clonación se encuentran también en Berger, Sambrook, y Ausubel, todos supra.

Además, cualquiera de las estrategias de clonación o amplificación descritas en esta memoria es útil para crear contigs de clones superpuestos, proporcionando así ácidos nucleicos solapantes que muestran la relación física a nivel molecular para los ácidos nucleicos genéticamente unidos. Un ejemplo común de esta estrategia se encuentra en proyectos de secuenciación de organismos completos, en los que los clones superpuestos se secuencian para proporcionar la secuencia completa de un cromosoma. En este procedimiento, se crea una colección de ADNc o ADN genómico del organismo de acuerdo con procedimientos estándar descritos, p. ej., en las referencias anteriores. Los clones individuales se aíslan y secuencian, y la información de secuencia solapante se ordena para proporcionar la secuencia del organismo.

Una vez que se han identificado uno o más QTLs que están significativamente asociados con la expresión del gen de interés, entonces cada uno de estos loci y marcadores vinculados también se pueden caracterizar adicionalmente para determinar el gen o genes implicados con la expresión del gen de interés, por ejemplo, utilizando métodos de clonación basados en mapas como sería conocido por un experto en la técnica. Por ejemplo, uno o más genes reguladores conocidos pueden mapearse para determinar si la ubicación genética de estos genes coincide con los QTLs que controlan la expresión de ARNm del gen de interés. La confirmación de que un gen regulador de este tipo coincidente está afectando la expresión de uno o más genes de interés puede obtenerse utilizando técnicas estándar en la técnica, por ejemplo, pero no limitadas a transformación genética, complementación de genes o técnicas de inactivación de genes, o sobreexpresión. El mapa de enlace genético también se puede utilizar para aislar el gen regulador, incluidos cualesquiera genes reguladores nuevos, a través de enfoques de clonación basados en mapas que se conocen en la técnica mediante los cuales los marcadores posicionados en el QTL se utilizan para caminar hacia el gen de interés utilizando contigs de clones genómicos de inserción grandes. La clonación posicional es uno de esos métodos que puede utilizarse para aislar uno o más genes reguladores, tal como se describe en Martin et al. (Martin et al., 1993, Science 262: 1432-1436).

La "clonación posicional de genes" utiliza la proximidad de un marcador genético para definir físicamente un fragmento cromosómico clonado que está vinculado a un QTL identificado utilizando los métodos estadísticos de esta memoria. Clones de ácidos nucleicos unidos tienen una diversidad de usos, incluidos como marcadores genéticos para la identificación de QTLs enlazados en protocolos de mejora asistidos por marcadores posteriores, y para mejorar las propiedades deseadas en plantas recombinantes en donde la expresión de las secuencias clonadas en una planta transgénica afecta a un rasgo identificado. Secuencias enlazadas comunes que se clonan de manera deseable incluyen marcos de lectura abiertos, p. ej., que codifican ácidos nucleicos o proteínas que proporcionan una base molecular para un QTL observado. Si los marcadores están próximos al marco de lectura abierto, pueden hibridarse con un clon de ADN dado, identificando así un clon en el que se encuentra el marco de lectura abierto. Si los marcadores flanqueantes están más distantes, se puede identificar un fragmento que contiene el marco de lectura abierto construyendo un contig de clones solapantes. Sin embargo, también se pueden utilizar otros métodos adecuados como reconoce un experto en la técnica. De nuevo, la confirmación de que un gen regulador coincidente está afectando la expresión de uno o más genes de interés puede obtenerse mediante transformación genética y complementación o mediante técnicas de inactivación descritas más adelante.

Tras la identificación de uno o más genes responsables o que contribuyen a un rasgo de interés, se pueden generar plantas transgénicas para lograr el rasgo deseado. Plantas que exhiben el rasgo de interés pueden incorporarse a las líneas de plantas mediante reproducción o mediante tecnologías comunes de ingeniería genética. Enfoques y técnicas de reproducción son conocidos en la técnica. Véase, por ejemplo, Welsh J. R., Fundamentals of Plant Genetics and Breeding, John Wiley & Sons, NY (1981); Crop Breeding, Wood D. R. (Ed.) American Society of Agronomy Madison, Wis. (1983); Mayo O., The Theory of Plant Breeding, Segunda Edición, Clarendon Press, Oxford (1987); Singh, D. P., Breeding for Resistance to Diseases and Insect Pests, Springer-Verlag, NY (1986); y Wricke y Weber, Quantitative Genetics and Selection Plant Breeding, Walter de Gruyter and Co., Berlin (1986).

Técnicas relevantes incluyen, pero no se limitan a hibridación, endogamia, reproducción por retrocruzamiento, reproducción multilínea, endogamia dihaploide, mezcla de variedades, hibridación interespecífica, técnicas aneuploides, etc.

En algunos ejemplos, puede ser necesario modificar genéticamente las plantas para obtener un rasgo de interés utilizando métodos rutinarios de ingeniería de plantas. En este ejemplo, pueden introducirse en la planta una o más secuencias de ácidos nucleicos asociadas con el rasgo de interés. Las plantas pueden ser homocigóticas o heterocigotas para la o las secuencias de ácidos nucleicos. La expresión de esta secuencia (transcripción y/o traducción) da como resultado una planta que exhibe el rasgo de interés. Métodos para la transformación de plantas son bien conocidos en la técnica.

Los siguientes Ejemplos se ofrecen a modo de ilustración y no a modo de limitación.

EJEMPLOS EXPERIMENTALES

Ejemplo 1 Detección de QTL en una población anidada

El NAM se realiza utilizando SMR y MMR en combinación con el método de permutación descrito más adelante para determinar el umbral de LOD para NAM.

Regresión de Marcador Único (SMR):

El modelo lineal para describir la relación entre los valores de los rasgos y los genotipos de marcadores es:

$$y_{ij} = \mu + x_{ij}a + g_i u_i + e_{ij} \qquad \text{(modelo 1)}$$

en que y_{ij} es el valor fenotípico del *j* individual en la subpoblación *i*; μ es la media general; *a* es el efecto aditivo de QTL; g_i es la variable indicadora de la subpoblación *i*; u_i es el efecto de la subpoblación *i*; e_{ij} es el error residual; y en donde x_{ij} se define como 1 si el *j* individual lleva el alelo del parental común y -1 si el *j* individual lleva el alelo del otro parental.

La definición se basa en el hecho de que solo hay dos alelos distintos para cada uno de los marcadores. Para explotar la simplicidad de regresión, el efecto de fondo genético u_i se supone que es un efecto fijo. Su inclusión en el modelo es para justificar la influencia de la estratificación de la población y, por lo tanto, reducir la varianza residual.

Las hipótesis para probar el efecto QTL se pueden formular como H_0 $a = 0$ y H_1 $a_1 \neq 0$. Los parámetros bajo H_0 o H_1 se estiman por el método de mínimos cuadrados basado en el modelo de regresión dependiendo de si el efecto QTL está incluido en el modelo SMR. $LR = -2(\ell_{reducida} - \ell_{completa})$, en que $\ell_{reducida}$ es la probabilidad logarítmica del modelo reducido, correspondiente a H_0, y $\ell_{completa}$ es la del modelo completo, correspondiente a H_1 (Lander y Botstein 1989). Ambos se calculan a partir del modelo SMR y una puntuación LOD se calcula como $LR/(2 \ln 10)$. Téngase en cuenta que el siguiente método MMR utiliza la misma prueba de hipótesis y el método para calcular LOD.

Este método SMR difiere del método de regresión original basado en marcadores para NAM (Yu *et al* (2006) *Nature Genetics* 38(2):203-208) en el uso de marcadores polimórficos. A partir de los datos de marcadores NAM, algunos marcadores muestran polimorfismo en algunas subpoblaciones, pero no en otras. En este caso, la inclusión de marcadores no informativos puede conducir a la distorsión de la segregación de los genotipos del marcador en ese locus, y la distorsión podría provocar la reducción de la eficiencia, potencia y precisión del mapeo de QTL. Para evitar el problema, un procedimiento filtrado por marcador se incorpora en el modelo SMR para reducir el riesgo potencial debido a la distorsión del marcador. De acuerdo con la presente invención, solo los datos fenotípicos y genotípicos de esas subpoblaciones con genotipos segregados de un marcador se incluyen en cada uno de los análisis. Por lo tanto, en la invención, las subpoblaciones con genotipos no informativos se excluyen antes del análisis SMR. El procedimiento permite a SMR identificar aquellos alelos con muy baja frecuencia (menos del 5%) en NAM.

Regresión de Marcador Múltiple (MMR)

Para justificar las influencias de otros QTLs, se desarrolló un método MMR utilizando marcadores de cofactor para absorber el efecto de otros QTLs. El modelo lineal para MMR es

$y_{ij} = \mu + x_{ij}a + \Sigma(k=1, m)c_{ijk}b_k + g_iu_i + e_{ij}$ (modelo 2)

en que y_{ij} es el valor fenotípico del *j* individual en la subpoblación *i*; en donde μ es la media general; en donde x_{ij} es el genotipo de QTL; en donde *a* es el efecto aditivo de QTL; en donde c_{ijk} es el marcador cofactor *k* para el *j* individual en la subpoblación *i*; en donde b_k es el efecto del marcador cofactor *k*; en donde g_i es la variable indicadora de la subpoblación *i*; en donde u_i es el efecto de la subpoblación *i*; y en donde e_{ij} es el error residual.

Otro aspecto de la invención es el uso de regresión escalonada para seleccionar marcadores de cofactor basados en el nivel de significancia 0,01. El modelo lineal utilizado para elegir cofactores es

$y_{ij} = \mu + c_{ijk}b_k + g_iu_i + e_{ij}$ (modelo 3)

en que y_{ij} es el valor fenotípico del *j* individual en la subpoblación *i*; en donde μ es la media general; en donde c_{ijk} es el marcador cofactor *k* para el *j* individual en la subpoblación *i*; en donde b_k es el efecto del marcador cofactor *k*; en donde g_i es la variable indicadora de la subpoblación *i*; en donde u_i es el efecto de la subpoblación *i*; y en donde e_{ij} es el error residual. Este modelo de regresión gradual es diferente del utilizado para el mapeo de intervalos compuesto convencional (Zeng 1993, 1994) y el utilizado originalmente para NAM (Yu et al 2008). Un aspecto de la presente invención realiza la regresión escalonada para una población NAM con la inclusión de fondos genéticos de diferentes subpoblaciones en el modelo 3. Este método selecciona aquellos QTL que tienen efectos estables a través de múltiples subpoblaciones. Los efectos estables se refieren a los efectos que se observan en múltiples poblaciones. La presente invención también reduce efectivamente el número de cofactores incluidos en el modelo, evitando el problema de una sobresaturación.

Con los marcadores de cofactor es posible obtener un perfil LOD mucho más claro de MMR que SMR. El uso de marcadores de cofactor es para reducir el error residual y, por lo tanto, aumentar la importancia de la prueba de hipótesis QTL. MMR muestra la capacidad de separar el análisis QTL estrechamente vinculado y localizar un QTL dentro de una región genómica estrecha.

Sin embargo, MMR tiene dificultades para utilizar un procedimiento de filtro de marcadores. Este problema es provocado por la singularidad de la matriz de diseño utilizada para el modelo de regresión. Por lo tanto, en lugar de filtrar marcador no informativo, la presente invención utiliza todos los datos genotípicos de todas las subpoblaciones para el análisis de datos. En base a esto, el SMR y MMR proporcionarán resultados similares para aquellos marcadores sin segregación distorsionada, mientras que pueden mostrar resultados diferentes para los marcadores con segregación genotípica sesgada. Por lo tanto, la invención está diseñada para realizar tanto SMR como MMR como una combinación complementaria para el conjunto de datos de NAM.

Ensayos de permutación para NAM:

El método original de regresión múltiple para NAM (Yu et al 2008) utilizó un nivel de significancia muy bajo 10^{-7} como umbral para la detección de QTL. Este método no es apropiado para determinar el umbral de LOD en un nivel de significancia dado, especialmente basado en un mapa de enlace denso. Para resolver este problema, la presente invención utiliza un método nuevo de ensayos de permutación para determinar el umbral empírico de LOD en el nivel de significancia dado de 0,05 y 0,01. El método reordena los valores fenotípicos dentro de cada una de las subpoblaciones sin destruir la estructura de las subpoblaciones y la correlación entre los diferentes rasgos de interés. Se recomiendan el uso de 1000 permutaciones para SMR y MMR. De estas permutaciones, se determina el umbral de LOD a niveles de 0,05 y 0,01. Téngase en cuenta que el umbral de 0,01 puede no ser estable debido al número limitado de ensayos de permutación (se recomiendan 10000 permutaciones).

Ejemplo 2 Método de selección de iniciativas candidatas a una validación adicional después del amplio mapeo de asociación del genoma

Con el advenimiento de los ‘omics, la identificación de candidatos clave entre los miles de genes en un genoma que juegan un papel en un fenotipo o un proceso biológico complejo se ha convertido paradójicamente en uno de los principales obstáculos. De hecho, contrariamente a algunas preocupaciones iniciales de que la carencia de datos globales suficientes seguiría siendo un factor limitante, es precisamente lo contrario, una generosidad de información la que ahora plantea un desafío para los científicos. Esto se ha traducido en la necesidad de herramientas sofisticadas para extraer, integrar y priorizar cantidades masivas de información. La presente invención ayudará a priorizar las iniciativas candidatas identificadas por el mapeo de asociación amplia del genoma (utilizando, p. ej., secuencias de la tecnología Solexa) para una validación e implementación adicionales en la reproducción asistida por marcadores.

La población de mapeo de asociación anidada desarrollado por el grupo de Diversidad Funcional de Maíz (Yu et al. Genetics 2008, 178: 539-551) se utiliza para los rasgos de mapeo de QTL de interés. Dado que el mapa de enlace tiene una resolución de ~ 1 cM (es decir, una densidad de marcador de 1 cM), los QTL identificados en esta población deben ser muy precisos. El mapeo de QTL se realiza utilizando la información de alelos compartida de los parentales que se utilizaron para desarrollar la población. Las secuencias que se utilizan para el mapeo de asociación amplia del genoma están situadas en el mapa físico del maíz. Los marcadores en el mapa de enlace de NAM también se colocan en el mapa físico del maíz.

Los QTL identificados en la población de NAM se alinean en el mapa físico siempre que las secuencias de Solexa y los QTL de la población de NAM se solapen entre sí. Esas secuencias de la secuenciación de Solexa se priorizan para una validación adicional de las secuencias que no se solapan con los QTLs identificados en la población de NAM. Véase la Figura 4.

Ejemplo 3 Detección de QTL utilizando NAM, SMR y MMR

Diseño experimental y preparación de datos fenotípicos y genotípicos

Las líneas NAM RIL se plantaron en cinco ubicaciones en el espacio de dos años. Los rasgos de interés, que incluían principalmente el almidón y la proteína en el proyecto de etanol de maíz, se evaluaron en diferentes ubicaciones y años. Los datos fenotípicos de cada una de las ubicaciones no están equilibrados. La estructura de datos desequilibrada indica que es necesario obtener los datos genotípicos correspondientes para esas líneas. Para ello, los datos del genotipo se descargaron para todos los marcadores (www.panzea.org/lit/data_sets.html) y la información genotípica se extrajo para las líneas NAM evaluadas. Además, para llevar a cabo SMR y MMR, se encontró un mapa de enlace de consenso desde el mismo sitio web y se descargó para el uso adicional.

Métodos para el análisis de datos

Se utilizaron NAM, SMR y MMR para detectar el QTL responsable del almidón y la proteína en el maíz. Los detalles de estos métodos se describen en el Ejemplo 1. Tanto SMR como MMR se utilizan para el mapeo de QTL. SMR tiene una ventaja para disminuir las influencias de la distorsión de segregación de marcadores, mientras que MMR puede localizar un QTL dentro de una región estrecha en el cromosoma. La combinación de SMR y MMR maximiza la capacidad de detección de QTL, al tiempo que minimiza el riesgo de perder cualquier QTL con efectos menores.

Se han desarrollado métodos de permutación para SMR y MMR de modo que los umbrales empíricos de LOD para los dos métodos se puedan determinar en el nivel de significancia dado de 0,05. En este análisis, se utilizaron 1000 permutaciones para llevar a cabo ensayos de permutación para uno cualquiera de ellos.

Resultados del mapeo de QTL

Se encontraron once QTL para el rasgo de almidón de maíz y diez QTL para proteínas. Entre estos QTL, seis QTL para el almidón se identificaron consistentemente en todas las ubicaciones y cinco para las proteínas. Además, se encontró que 6 QTL controlan tanto el almidón como la proteína, lo que indica los posibles efectos pleiotrópicos para ambos rasgos. Se descubrió que estos seis QTL tienen grandes efectos en los rasgos individuales. La identificación de estos QTL pleiotrópicos probablemente explica la fuerte correlación fenotípica entre el almidón y la proteína en el maíz.

Conclusión

Como era de esperar, SMR y MMR identificaron el QTL principal y pleiotrópico para el almidón y la proteína basado en el diseño experimental de NAM. Se ha demostrado que ambos métodos son una herramienta poderosa para la detección de QTL en NAM. Los métodos de permutación para cualquiera de los métodos proporcionaron umbrales LOD para la detección de QTL.

Ejemplo 4 Un ejemplo de análisis de asociación amplia del genoma en combinación con mapeo de enlace con poblaciones de mapeo de asociación anidadas para priorizar los genes candidatos para la validación/implementación biológica

Introducción

El análisis de asociación amplia del genoma (GWA) es una herramienta poderosa para identificar variantes genéticas comunes en una población que afectan los rasgos de interés, que ofrece una alta resolución de mapeo hasta un cambio de un solo nucleótido. El estudio de asociación aprovecha los eventos de recombinación en el genoma acumulado durante muchas generaciones, que han segmentado el genoma en trozos de bloques de desequilibrio de enlace (LD) en la población. Los marcadores en cada uno de los bloques LD exhiben habitualmente asociaciones significativas con cambios funcionales en los genes en los mismos bloques y, por lo tanto, se pueden tomar como indicadores de los cambios funcionales relevantes en la fitomejora, o como una base para identificar con mayor precisión los genes responsables.

El objetivo de GWA es detectar marcadores que estén físicamente vinculados a los cambios funcionales relevantes. Sin embargo, es común detectar asociaciones de marcadores que están desvinculados o enlazados de forma distante de los cambios, que habitualmente se consideran falsos positivos. Si bien muchos otros factores genéticos de la población (tales como la migración, la mutación, la deriva genética, el apareamiento no aleatorio) también pueden contribuir a la tasa de falsos positivos, la estratificación de la población o la estructura de la población se han identificado como una de las principales preocupaciones que podrían provocar una gran cantidad de falsos positivos en GWA. La estructura de la población existe cuando las frecuencias alélicas son sistemáticamente diferentes entre

las subpoblaciones en una población, lo que puede ser provocado por la migración y el apareamiento no aleatorio, etc.

Un ejemplo para GWA

Muestras y Datos

1) Panel endogámico para GWA: Se ensambló un panel endogámico de maíz para incluir 600 líneas endogámicas seleccionadas para maximizar la diversidad genética de una plataforma de aproximadamente 3000 líneas endogámicas de maíz. Se sabe que 450 de las líneas en el panel se derivan de 3 subpoblaciones, a saber, los subgrupos de tallo no rígido (NSS), de tallo rígido (SS) y tropical-subtropical (TS); las 150 líneas restantes no tienen identidad de subgrupo disponible por varias razones en la práctica.

2) Datos genotípicos en 500.000 SNP: Se empleó la técnica de secuenciación de Solexa para examinar las colecciones de ADNc de genoma completo de 600 líneas endogámicas diversas en el panel endogámico para SNP de todo el genoma, que identificaron aproximadamente 500.000 SNP de alta calidad.

3) Datos fenotípicos sobre 3 rasgos relacionados con el etanol: El contenido porcentual de almidón, aceite y proteína en los granos de maíz, los 3 rasgos principales relacionados con el etanol, se evaluaron con la máquina del Espectroscopio de Infrarrojo Cercano (NIR) para cada una de las 600 endogamias en el panel endogámico desarrollado 2 ubicaciones.

Procesamiento previo de datos

Los datos fenotípicos se evaluaron para eliminar el fenotipo de los puntos de datos sospechosos empírica y estadísticamente, tales como los valores atípicos. La distribución estadística de los datos fenotípicos también se evaluó para determinar si la transformación de datos o las permutaciones eran necesarias para reclamar asociaciones significativas de marcador-rasgo. Tal como se muestra en el histograma para el rasgo (Figura 5), estos 3 rasgos se distribuyen aproximadamente de manera normal, lo que sugiere que los valores de p estimados a partir de los ensayos de asociación que se basan en la suposición normal serán básicamente válidos.

Los datos genotípicos se evaluaron para identificar errores obvios, por ejemplo, más de 2 alelos para un marcador SNP y SNP no informativos (monomórficos o aquellos con una menor frecuencia de alelo < 0,05). Entre los 500.000 SNPs, 1200 SNPs no fueron informativos y quedaron excluidos de los datos. No se pudo analizar el equilibrio de Hardy-Weinberg para una población endogámica porque existe una deficiencia inherente de heterocigotos en la población.

Ajuste de datos fenotípicos para líneas endogámicas individuales

Se utilizó un enfoque de modelo lineal mixto para obtener el valor genético total para cada una de las líneas endogámicas en la muestra con control de los efectos de las ubicaciones y los bloques aleatorios. En este modelo, el efecto genético total para cada una de las líneas endogámicas se toma como aleatorio porque las líneas endogámicas utilizadas en el panel se consideran como una muestra aleatoria de todo el germoplasma; los bloques aleatorizados también se consideran aleatorios; las ubicaciones se toman como efectos fijos, que son las ubicaciones diana para cualquier híbrido futuro que se cultive. El modelo estadístico para el análisis se puede escribir como

$$Y_{hijk} = \mu + G_h + L_i + B_{j(i)} + e_{hijk}$$

en que μ es la media general; G_h es el efecto genético aleatorio total para la *h*-ésima endogamia; L_i es el efecto fijo para la ubicación *i*; $B_{j(i)}$ es el efecto de bloque aleatorio para el bloque j en la ubicación *i*; e_{hijk} es el residuo aleatorio que se supone que se distribuye normalmente. Este modelo está equipado en paquete R estadístico con la colección *lme4* (www.r-project.org).

Los mejores valores pronosticados no sesgados lineales (BLUP) para G_h se obtienen y utilizan como datos fenotípicos en el enfoque de asociación de modelo lineal mixto tal como se implementa en el software TASSEL.

Estimación de la estructura de la población

La inclusión de la estructura de la población en modelos estadísticos puede reducir de manera eficaz los falsos positivos en el análisis de asociación. TASSEL incorpora la estructura de la población como un factor modelo en el modelo lineal mixto para lograr este objetivo.

Como se mencionó previamente, hay 3 subpoblaciones conocidas (SS, NSS, y TS) en el panel endogámico, pero el 25% de las líneas endogámicas no tienen identidad de subpoblación. Una forma de evitar esto es estimar la estructura de la población para el panel endogámico con los datos de SNP para las líneas endogámicas. Se seleccionó un conjunto aleatorio de 2000 SNPs de todos los SNPs informativos, y se utilizaron para la estimación de la estructura de la población.

Tal como se describe en la Solicitud de Patente de EE.UU. 12/328.689, presentada el 4 de diciembre de 2008, el análisis de componentes principales (PCA) ofrece una precisión similar en la estimación de la estructura de la población al enfoque bayesiano en ESTRUCTURA. El PCA se realizó con todos los datos de SNP y se obtuvieron los principales 50 componentes principales (PC) y se obtuvieron sus vectores propios. Los componentes principales que contribuyen específicamente en el rasgo de las 50 PCs se seleccionaron con un análisis de regresión escalonada, el cual ha demostrado que proporciona un mejor control de los efectos de la estructura de la población que simplemente utilizando algunos PCs principales en el modelo mixto de asociación.

Estimación de coeficientes de parentesco

El coeficiente de parentesco es una medida de la relación entre dos individuos. Representa la probabilidad de que dos genes, muestreados al azar de cada uno de los individuos, sean idénticos por descendencia. Existe un cierto número de métodos para estimar los coeficientes de parentesco con datos de marcadores, teniendo cada uno ventajas y desventajas. La proporción de alelos compartidos en todos los loci SNP se eligió como la medida del coeficiente de parentesco entre un par de líneas endogámicas, que es esencialmente la probabilidad de que dos genes aleatorios sean idénticos por estado. Coeficientes de parentesco de este tipo se calcularon para todos los pares de líneas posibles.

Análisis de asociación con un enfoque de modelo mixto

Se han aplicado modelos lineales mixtos al mapeo de asociación en plantas (Yu et al. 2006, Nature Genetics), que ha demostrado ser superior en el control de la estructura de la población. Este enfoque fue implementado en ASReml (Gilmour *et al.* (1995) *Biometrics* 51:1440-1450), un paquete de software comercial para ejecutar modelos mixtos generales, a través de un completo script Perl que proporciona una automatización completa del análisis de datos para múltiples rasgos. En comparación con TASSEL, el software que implementa el enfoque de modelo mixto de Yu et al. (2006, Nature Genetics, Vol. 38: 203-208), ASReml es mucho más rápido y el script Perl minimiza la atención del usuario.

El modelo lineal mixto implementado en ASReml es el mismo que en TASSEL, que se puede escribir en forma de matriz como

$$y = X\beta + S\alpha + Qv + Zu + e \qquad var(y) = ZKZ'\sigma^2_v + R\sigma\,\sigma^2_e$$

en que **y** es el vector para los valores fenotípicos de todas las líneas endogámicas únicas; **β** es el vector para todos los efectos experimentales fijos, **α** es el vector para los efectos genéticos del supuesto QTL en la posición de ensayo; **v** es el vector para los efectos de subpoblación; **u** es el vector para los efectos poligénicos de las endogamias individuales; **e** es el vector residual aleatorio. **X**, **S**, **Q** y **Z** son matrices de incidencia conocidas.

En este análisis, los datos fenotípicos ajustados (efectos genéticos totales) se utilizaron como **y**, los 10 vectores propios de PCA se utilizaron como matriz **Q**; la matriz **X** es esencialmente un vector de 1s para la media general; **S** es la matriz de genotipo bajo modelo genético aditivo para cada uno de los SNP de ensayo; **Z** es la matriz de incidencia para el conjunto de líneas endogámicas únicas.

Resultados de la asociación

Se calculó un valor p para probar la importancia de cada uno de los SNP informativos en el análisis de asociación, junto con la contribución fenotípica R cuadrado y unas pocas de otras estadísticas. Tanto la tasa de descubrimiento falso (FDR) como la corrección de Bonferroni se utilizaron para controlar los falsos positivos inflados con múltiples ensayos. El valor nominal p en el nivel de significancia (alfa) con corrección de Bonferroni se calculó como alfa / número de ensayos (SNPs); el umbral FDR se derivó de la distribución estimada del valor p. Se utilizó la media entre los dos umbrales, al mismo nivel de significancia (alfa).

Alfa = 0,05 fue elegido como el nivel de significancia para todos los ensayos. Esto dio como resultado 102 SNPs asociados significativamente con el contenido de almidón, 134 SNPs asociados con proteínas y 97 SNPs asociados con aceite. Se descubrió que estos SNPs eran de 30, 35 y 23 bloques de desequilibrio de enlace en el genoma, respectivamente, para el almidón, la proteína y el aceite.

Superposición de asociaciones GWA con resultados de mapeo de enlaces de NAM

Asociaciones estadísticamente significativas no siempre pueden indicar verdaderas asociaciones biológicas, posiblemente debido a errores de muestreo. Por lo tanto, la evidencia de fuentes independientes puede ser útil para validar las asociaciones detectadas.

Las poblaciones de mapeo de asociación anidada (NAM) en el maíz, como un nuevo tipo de población de mapeo, se pusieron a disposición del público (Yu et al. 2008, Genetics, y Vol. 178: 539 551). La ventaja de este tipo de población es que ofrece una promesa de mayor poder estadístico y resolución de mapeo que el mapeo de enlace, pero menos falsos positivos que el mapeo de asociación con muestras de una población general. El estudio de

mapeo de ligamiento para el almidón con las poblaciones de NAM se ha realizado previamente (Ejemplo 3 arriba mencionado), a partir del cual se identificaron 11 regiones QTL para almidón, 10 para proteína y 8 para aceite.

El método para superponer los SNPs asociados de GWA a las regiones QTL detectadas a partir del análisis de enlace NAM fue colocar los SNPs y los marcadores asociados en las regiones QTL en el mismo mapa físico y un mapa genético consensuado (véase la Figura 4) . La Tabla 1 indica que el 55% de todos los SNPs asociados estaba contenido en 8 regiones QTL para almidón; el 31,1% de los SNPs asociados estaba contenido en 6 QTL para proteínas, el 27,8 de todos los SNPs asociados estaba contenido en 3 QTLs.

Tabla 1 Sumario de QTL superpuestos y SNPs asociados

Rasgo	Todos los QTLs de NAM	Todos los SNPs asociados a GWA	QTL solapante	SNPs solapantes	% SNP solapante
Almidón (%)	11	102	8	41	40,2
Proteína (%)	10	134	6	33	24,6
Aceite (%)	8	97	3	27	27,8

Los SNPs de los genes que se solapan con los QTLs detectados en la población de NAM se les da una prioridad más alta y se utilizan para una validación adicional biológica. Estos SNPs también se utilizan para la aplicación posterior, tal como la reproducción asistida por marcador.

Todas las publicaciones y solicitudes de patentes mencionadas en la memoria descriptiva son indicativas del nivel de habilidad de los expertos en la técnica a los que pertenece esta invención.

REIVINDICACIONES

1. Un método implementado por computadora para identificar un marcador genético asociado con un rasgo de interés en una población anidada de plantas de *Zea mays*, que comprende:

a) proporcionar un valor genotípico para cada uno de una pluralidad de marcadores genéticos para cada uno de los miembros de dicha población anidada, en donde dicha población comprende miembros que exhiben dicho rasgo de interés;

b) proporcionar un valor fenotípico para dicho rasgo de interés para cada uno de los miembros de dicha población;

c) determinar si uno o más de dichos marcadores están asociados con el rasgo de interés utilizando una computadora adecuadamente programada para realizar un mapeo de asociación anidado que comprende una combinación de una regresión de marcador único, SMR, y una regresión de marcador múltiple, MMR, en donde:

i) los genotipos no informativos se eliminan antes de evaluar una asociación entre un valor de rasgo y un genotipo marcador utilizando el modelo SMR; y,

ii) la regresión escalonada se utiliza para seleccionar marcadores de cofactores para su inclusión en el modelo MMR;

en donde un marcador genético se considera que está asociado con el rasgo de interés si los dos modelos de regresión detectan una asociación;

d) seleccionar una planta de Zea mays con el marcador genético asociado identificado en la etapa c).

2. El método de la reivindicación 1, en el que dicho modelo SMR comprende:

$$y_{ij} = \mu + x_{ij}a + g_i u_i + e_{ij}$$

en donde y_{ij} es el valor fenotípico del *j* individual en la subpoblación *i*;

en donde μ es la media general;

en donde *a* es el efecto aditivo de QTL;

en donde g_i es la variable indicadora de la subpoblación *i*;

en donde u_i es el efecto de la subpoblación *i*;

en donde e_{ij} es el error residual; y

en donde x_{ij} se define como 1 si el *j* individual lleva el alelo del parental común y -1 si el el individual lleva el alelo del otro parental.

3. El método de la reivindicación 1, en el que dicho modelo MMR comprende:

$$y_{ij} = \mu + x_{ij}a + \Sigma(k=1, m)c_{ijk}b_k + g_i u_i + e_{ij}$$

en donde y_{ij} es el valor fenotípico del *j* individual en la subpoblación *i*;

en donde μ es la media general;

en donde x_{ij} es el genotipo de QTL;

en donde *a* es el efecto aditivo de QTL;

en donde c_{ijk} es el marcador de cofactor *k* para el *j* individual en la subpoblación *i*;

en donde b_k es el marcador de cofactor *k*;

en donde g_i es la variable indicadora de la subpoblación *i*;

en donde u_i es el efecto de la subpoblación *i*; y

en donde e_{ij} es el error residual.

4. El método de la reivindicación 3, en el que los marcadores de cofactor se seleccionan en base a un nivel de significancia definido.

5. El método de la reivindicación 4, en el que dicho nivel de significancia es menor que o igual a 0,1.

6. El método de la reivindicación 5, en el que los cofactores se seleccionan utilizando un modelo que comprende:

$$y_{ij} = \mu + c_{ijk}b_k + g_iu_i + e_{ij}$$

en donde y_{ij}es el valor fenotípico del *j* individual en la subpoblación *i*;

en donde μ es la media general;

en donde c_{ijk} es el marcador de cofactor *k* para el *j* individual en la subpoblación *i*;

en donde b_k es el marcador de cofactor *k*;

en donde g_i es la variable indicadora de la subpoblación *i*;

en donde u_i es el efecto de la subpoblación *i*;
y

en donde e_{ij} es el error residual.

7. El método de la reivindicación 1, en el que dicha población anidada es una población endogámica que resulta de un cruce entre una única línea parental común y cada una de una pluralidad de líneas fundadoras.

8. El método de la reivindicación 7, en el que dicha población comprende una población que resulta de una o más rondas de auto-cruzamiento de la progenie del cruce entre dicho progenitor común único y cada una de dicha pluralidad de líneas fundadoras.

9. El método de la reivindicación 1, en donde la etapa a) comprende aislar material genético de cada uno de los miembros de dicha población para determinar el valor genotípico para cada uno de los marcadores.

10. El método de la reivindicación 1, que comprende, además, cruzar al menos un miembro de dicha población con otro organismo de la misma especie y seleccionar de la progenie del mismo cualquier organismo que tenga uno o más de los marcadores asociados identificados en la etapa c).

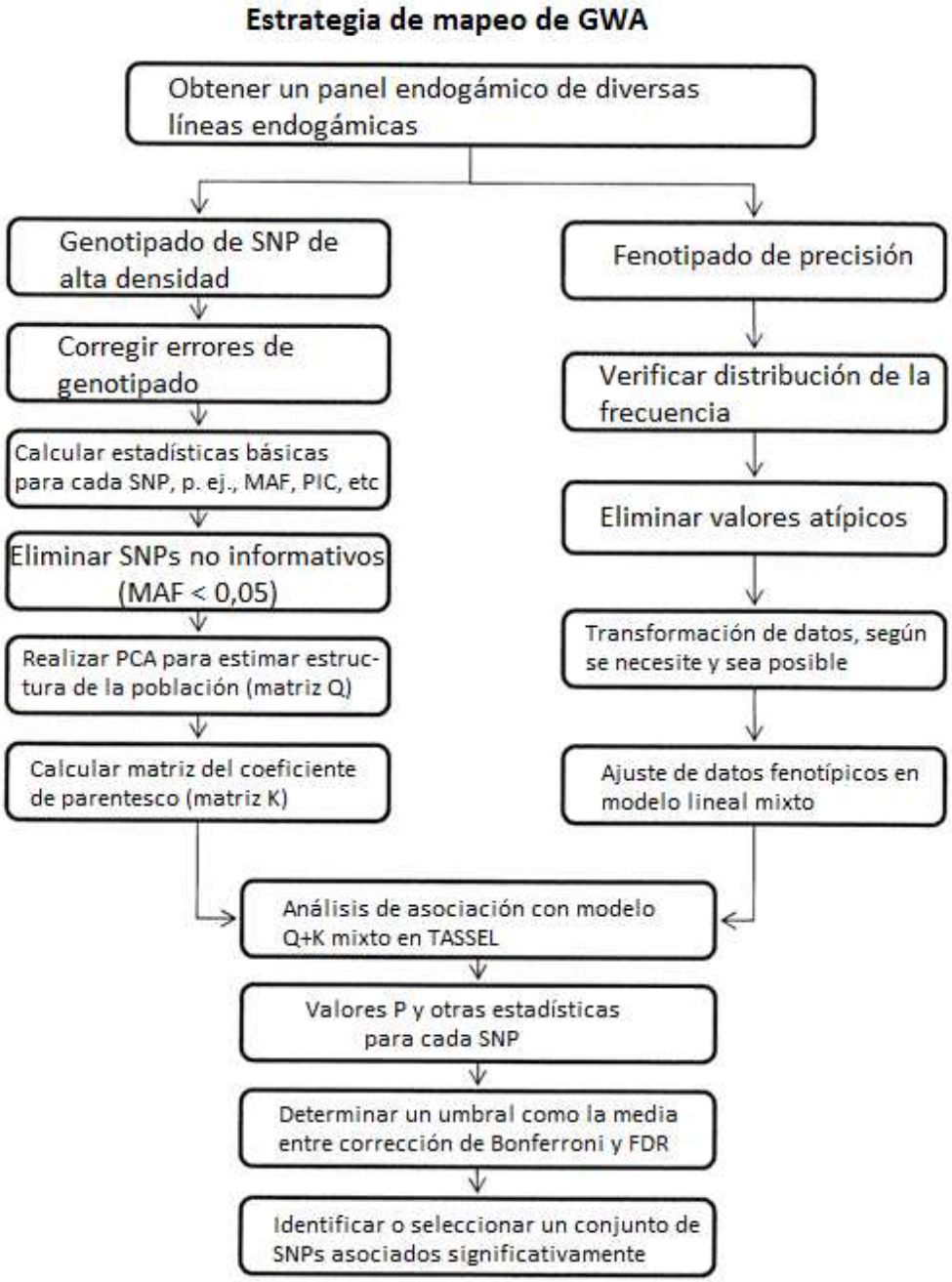

Estrategia de mapeo de GWA
Obtener un panel endogámico de diversas líneas endogámicas
Genotipado de SNP de alta densidad
Corregir errores de genotipado
Calcular estadísticas básicas para cada SNP, p. ej., MAF, PIC, etc
Eliminar SNPs no informativos (MAF < 0,05)
Realizar PCA para estimar estructura de la población (matriz Q)
Calcular matriz del coeficiente de parentesco (matriz K)
Fenotipado de precisión
Verificar distribución de la frecuencia
Eliminar valores atípicos
Transformación de datos, según se necesite y sea posible
Ajuste de datos fenotípicos en modelo lineal mixto
Análisis de asociación con modelo Q+K mixto en TASSEL
Valores P y otras estadísticas para cada SNP
Determinar un umbral como la media entre corrección de Bonferroni y FDR
Identificar o seleccionar un conjunto de SNPs asociados significativamente

FIG. 1

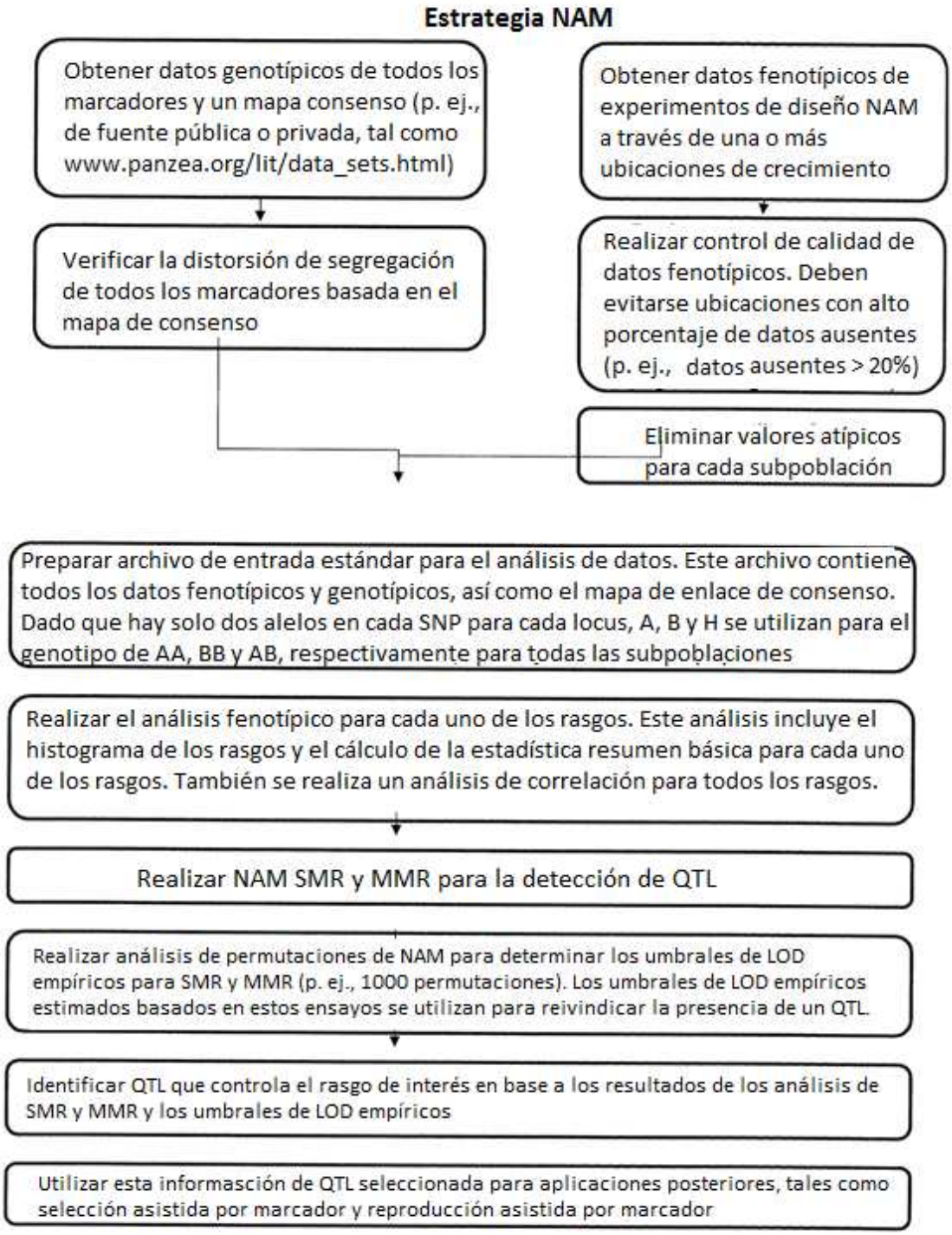

Estrategia NAM
Obtener datos genotípicos de todos los marcadores y un mapa consenso (p. ej., de fuente pública o privada, tal como www.panzea.org/lit/data_sets.html)
Obtener datos fenotípicos de experimentos de diseño NAM a través de una o más ubicaciones de crecimiento
Verificar la distorsión de segregación de todos los marcadores basada en el mapa de consenso
Realizar control de calidad de datos fenotípicos. Deben evitarse ubicaciones con alto porcentaje de datos ausentes (p. ej., datos ausentes > 20%)
Eliminar valores atípicos para cada subpoblación
Preparar archivo de entrada estándar para el análisis de datos. Este archivo contiene todos los datos fenotípicos y genotípicos, así como el mapa de enlace de consenso. Dado que hay solo dos alelos en cada SNP para cada locus, A, B y H se utilizan para el genotipo de AA, BB y AB, respectivamente para todas las subpoblaciones
Realizar el análisis fenotípico para cada uno de los rasgos. Este análisis incluye el histograma de los rasgos y el cálculo de la estadística resumen básica para cada uno de los rasgos. También se realiza un análisis de correlación para todos los rasgos.
Realizar NAM SMR y MMR para la detección de QTL
Realizar análisis de permutaciones de NAM para determinar los umbrales de LOD empíricos para SMR y MMR (p. ej., 1000 permutaciones). Los umbrales de LOD empíricos estimados basados en estos ensayos se utilizan para reivindicar la presencia de un QTL.
Identificar QTL que controla el rasgo de interés en base a los resultados de los análisis de SMR y MMR y los umbrales de LOD empíricos
Utilizar esta informasción de QTL seleccionada para aplicaciones posteriores, tales como selección asistida por marcador y reproducción asistida por marcador

FIG. 2

Superposición de resultados de GWA con resultados de enlace de poblaciones de NAM

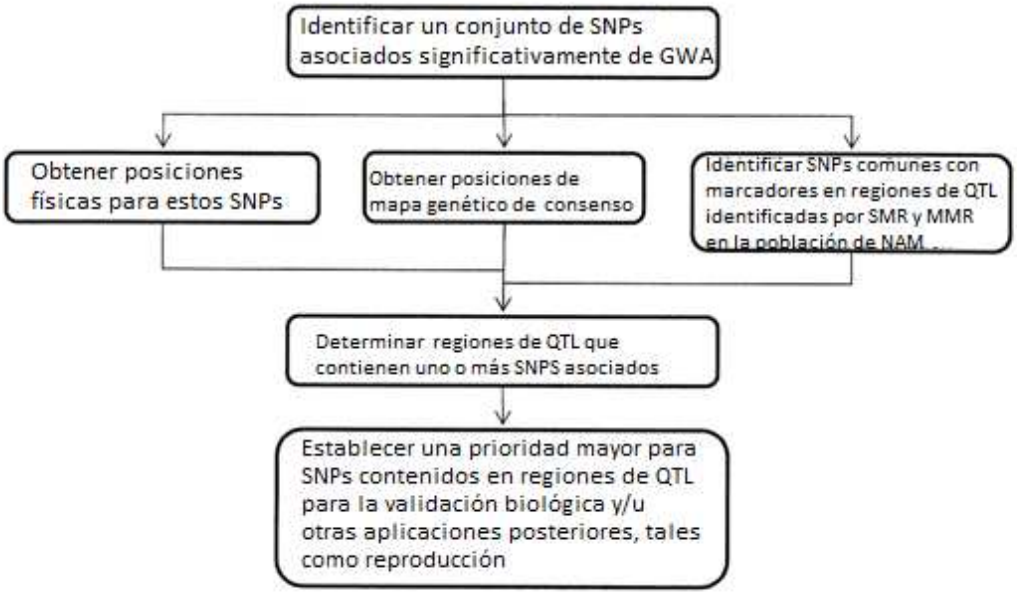


FIG. 3

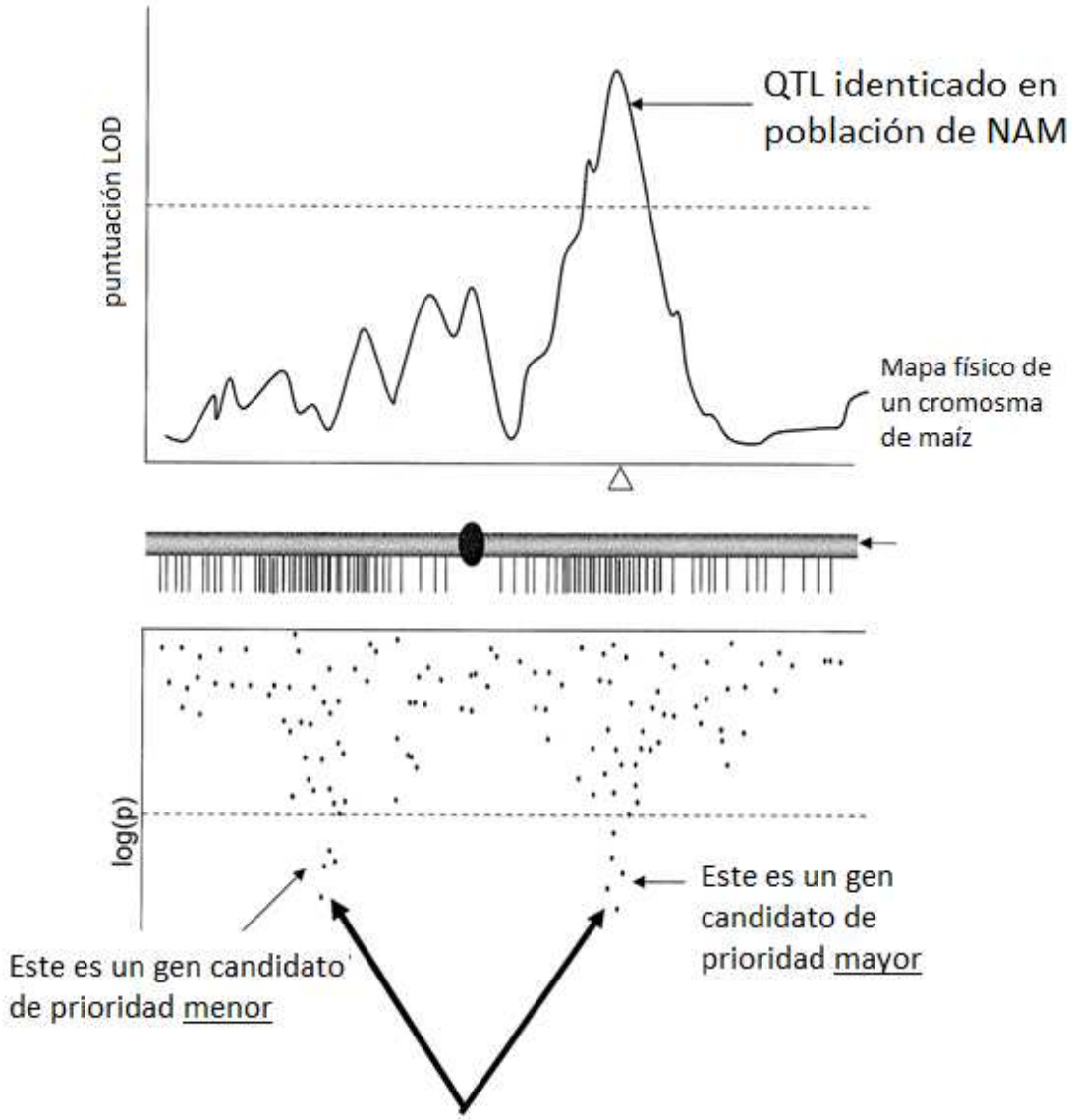
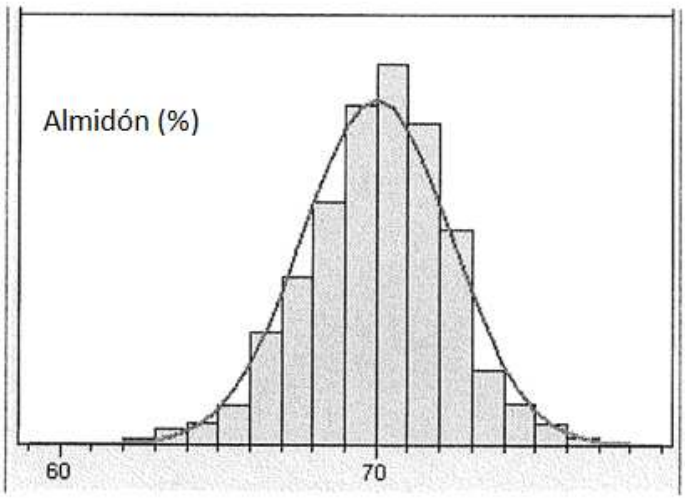
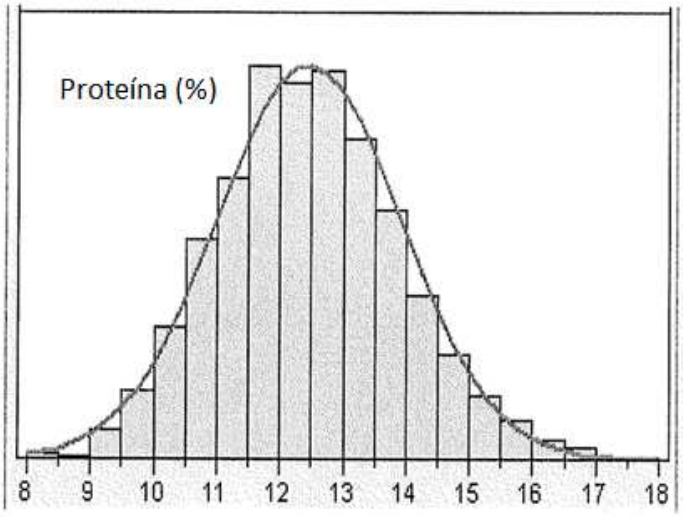

puntuación LOD
QTL identicado en población de NAM
Mapa físico de un cromosma de maíz
log(p)
Este es un gen candidato de prioridad menor
Este es un gen candidato de prioridad mayor
Asociaciones significativas identificadas por mapeo de asociación del genoma completo utilizando secuenciación Solexa

FIG. 4


Almidón (%)
60
70


Proteína (%)
8
9
10
11
12
13
14
15
16
17
18

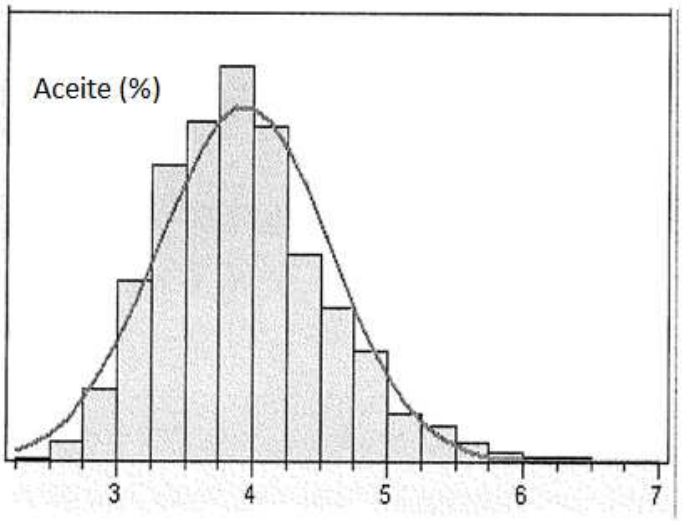

Aceite (%)
3
4
5
6
7

FIG. 5