(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 381 457**

(51) Int. CI.:
C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **08867704 .2**
(96) Fecha de presentación: **29.12.2008**
(97) Número de publicación de la solicitud: **2229458**
(97) Fecha de publicación de la solicitud: **22.09.2010**

(54) Título: **Uso de una variación estructural para analizar diferencias genómicas para la predicción de heterosis**

(30) Prioridad:
28.12.2007 US 17227 P

(45) Fecha de publicación de la mención BOPI:
28.05.2012

(45) Fecha de la publicación del folleto de la patente:
28.05.2012

(73) Titular/es:
PIONEER HI-BRED INTERNATIONAL INC.
7100 N.W. 62ND AVENUE
JOHNSTON, IA 50131-1014, US y
E. I. DU PONT DE NEMOURS AND COMPANY

(72) Inventor/es:
BEATTY, Mary;
JANNI, James A.;
LIGHTNER, Jonathan E. y
RAFALSKI, J. Antoni

(74) Agente/Representante:
de Elzaburu Márquez, Alberto

DESCRIPCIÓN

Uso de una variación estructural para analizar diferencias genómicas para la predicción de heterosis

CAMPO DE LA INVENCIÓN

Esta invención se refiere al campo de la biología molecular de plantas y al cultivo de plantas, particularmente a la predicción del grado de fenotipos heteróticos en plantas.

ANTECEDENTES

La producción agrícola ha aumentado drásticamente durante la segunda mitad del siglo veinte. Una gran parte de este aumento se ha atribuido al desarrollo y al uso de variedades de semillas híbridas en cultivos básicos como el maíz, el sorgo, el girasol, la alfalfa, la canola y el trigo. El éxito de las variedades de semillas híbridas se debe a un fenómeno denominado heterosis, por el cual las plantas híbridas presentan un fenotipo más deseable que cualquiera de las dos líneas originales usadas para producir la planta híbrida. La heterosis se ha observado en una serie de rasgos vegetales que incluyen el rendimiento, la altura de la planta, la biomasa, la resistencia a enfermedades e insectos, la tolerancia al estrés, y otros. Estos rasgos heteróticos son de naturaleza poligénica, lo que da como resultado su rango característico de fenotipos, en lugar de los fenotipos mendelianos discretos tradicionales. La naturaleza poligénica de los rasgos da como resultado estructuras complejas de herencia, de tal modo que los componentes subyacentes en los fenotipos heteróticos observados todavía son cuestión de debate en la comunidad de la ciencia de las plantas.

Debido al valor económico de la heterosis, ha habido varios intentos para usar técnicas de biología molecular con el objetivo de aumentar los programas de cultivo de plantas híbridas tradicionales. El grueso de los esfuerzos se ha centrado bien en el ARNm (ARN mensajero) o bien en el ADN genómico. La estrategia de ARNm es extremadamente difícil ya que las comparaciones requieren muestras de tejido seleccionadas de la misma porción de la planta, en el mismo tiempo de desarrollo y en las mismas, o muy similares, condiciones ambientales. El proceso se complica aún más ya que el investigador necesita determinar que porción de la planta o qué etapa de desarrollo dará lugar a los mejores resultados para predecir el grado de un fenotipo heterótico particular de interés. Como consecuencia de estas implicaciones, las predicciones basadas en ARNm frecuentemente tienen niveles elevados de ruido y presentan una baja precisión en la predicción del grado de un fenotipo heterótico.

El uso de ADN genómico para predecir el grado de uno o más fenotipos heteróticos ha sido igualmente decepcionante. Los esfuerzos iniciales usaron hibridación sustractiva o hibridación *in situ* fluorescente a fin de identificar diferencias de número de copia en líneas de plantas cultivadas. Estas técnicas no producen resultados fácilmente cuantificables y sólo pueden detectar diferencias grandes en los números de copia, tal como una doblado o una eliminación completa. Esto es un problema significativo en plantas poliploides, ya que las duplicaciones cromosomales y otros eventos evolucionarios han dado como resultado genes con copias múltiples, algunos de los cuales son seudogenes, a lo largo de todo el genoma de la planta. Estos números de copia más altos reducen enormemente la utilidad de las estrategias de ADN genómico, puesto que son incapaces de detectar de formar precisa la adición o la eliminación de una copia sencilla de un gen representado tres o más veces en el genoma.

Otra estrategia genómica ha sido el uso de marcadores genéticos para predecir la heterosis. En estas técnicas se han usado marcadores RFLP, así como otros marcadores tradicionales. Los investigadores han intentado usar marcadores genéticos para predecir el grado de un fenotipo heterótico con algo de éxito, siempre que las plantes madre potenciales pertenezcan a los mismos grupos heteróticos que se usaron en los cruces iniciales para generar los datos de correlación en los que se basa la predicción. Una vez que se usan plantas de otros grupos heteróticos, la capacidad predictiva de fenotipo heterótico de los marcadores genéticos disminuye enormemente. La razón de la pérdida de capacidad predictiva ha sido atribuida a un enlace insuficiente de los marcadores a sitios de rasgos cuantitativos que controlan el rasgo de interés, y a una carencia de desequilibrio de enlace de fase gamética entre el marcador y alelos de sitio de rasgo cuantitativo. Esta reducción de la capacidad predictiva limita gravemente el uso de marcadores genéticos en los programas de cultivo de plantas.

En base a estos esfuerzos, la aplicación de técnicas de biología molecular para la predicción del grado de un fenotipo heterótico ha sido como mínimo problemática. A pesar de los años de investigación, todavía no se ha desarrollado un método satisfactorio.

La Hibridación de Genoma Comparativa (CGH, en sus siglas en inglés) es una técnica que ha sido empleada para estudiar anormalidades cromosomales en células de animales. Un área principal de uso de la CGH ha sido el análisis de mutaciones de cáncer en un esfuerzo para identificar mejor células cancerígenas con el objetivo de seleccionar cursos de terapia más efectivos. La CGH es particularmente efectiva en células animales ya que hay normalmente dos copias de cada gen en el genoma (una de cada padre). Adicionalmente, actualmente se conocen genomas completos de mamíferos. Los investigadores han sido capaces de aprovechar la baja duplicación y la información de secuencia de genoma para identificar regiones cromosomales duplicadas y eliminadas. Esta información se puede usar después para identificar los cambios que han transformado células normales en células cancerosas. Sin embargo, en la actualidad se desconoce la secuencia de genoma completa de varios cultivos principales. Como resultado, se ha hecho poco uso de la CGH en plantas y hacerlo requiere superar las numerosas

diferencias que surgen cuando se trabaja con genómica de plantas.

RESUMEN

La presente invención se refiere al uso de análisis de variación estructural del genoma, tal como el análisis de variación del número de copias, detectado por ejemplo mediante el uso de hibridación genómica comparativa, para predecir el grado de una progenie de fenotipo heterótico en plantas. En un aspecto de la invención, se ponen en contacto grupos de moléculas sonda de oligonucleótidos y ADN genómico de planta y la mezcla resultante de sondas hibridadas y ADN genómico se analiza para determinar las sondas que muestran niveles de hibridación diferentes entre dos padres diferentes. A continuación los resultados se usan para predecir el grado de un fenotipo heterótico de plantas progenie derivadas de las dos líneas parentales. El grado predicho de un fenotipo heterótico puede usarse en el desarrollo de plantas híbridas. Se puede seleccionar un subconjunto de moléculas sonda de oligonucleótidos que son buenas predictoras del grado de un fenotipo heterótico a partir de una población mayor de moléculas sonda de oligonucleótidos, y el subconjunto seleccionado puede usarse entonces en ensayos futuros para predecir el grado de un fenotipo heterótico.

Por consiguiente, la invención proporciona un método que comprende:

(a) seleccionar una primera planta y una segunda planta, en donde la primera planta y la segunda planta pueden cruzarse para producir una planta progenie fértil que presente un fenotipo relacionado con heterosis en comparación con las plantas parentales;

(b) detectar variaciones estructurales de ADN entre un genoma de la primera planta y un genoma de la segunda planta; y

(c) relacionar las variaciones estructurales con el fenotipo relacionado a heterosis usando una estrategia computacional evolucionaria iterada,

identificando con ello las variaciones estructurales que predicen el grado esperado del fenotipo relacionado a heterosis en la planta progenie;

en donde, en la etapa (b),

(i) las variaciones estructurales son detectadas usando un método de hibridación genómica comparativo; o

(ii) las variaciones estructurales son variaciones del número de copias.

La invención también proporciona un método para desarrollar un sistema de oligonucleótidos para la predicción de un fenotipo relacionado a heterosis en una planta que comprende:

(a) seleccionar una pluralidad de líneas parentales en las que se ha cuantificado el fenotipo relacionado a heterosis en una pluralidad de los cruces F1 de dichas líneas parentales;

(b) poner en contacto ADN genómico de cada una de la pluralidad de dichas líneas parentales con una pluralidad de moléculas sonda de oligonucleótidos, en donde dichas pluralidades de moléculas sonda de oligonucleótidos tiene al menos un subconjunto de moléculas sonda de oligonucleótidos en común;

(c) detectar las intensidades de hibridación para moléculas sonda de oligonucleótidos individuales en las pluralidades de moléculas sonda de nucleótidos;

(d) determinar las medidas relativas de intensidad de hibridación para una pluralidad de las moléculas sonda de oligonucleótidos en dicho subconjunto de moléculas sonda de oligonucleótidos;

(e) seleccionar moléculas sonda de oligonucleótidos que muestran diferentes intensidades de hibridación entre dichas líneas parentales;

(f) relacionar dichas intensidades de hibridación de dichas moléculas sonda de oligonucleótidos con el fenotipo relacionado a heterosis de las plantas progenie; y

(g) crear un sistema de oligonucleótidos especializado para la predicción de un fenotipo relacionado a heterosis que comprende dichas moléculas sonda de oligonucleótidos que se relacionan con un fenotipo relacionado a heterosis.

BREVE DESCRIPCIÓN DE LAS FIGURAS

Figura 1: muestra las predicciones de rendimiento basadas en un modelo de regresión PLS construido usando las relaciones de intensidad seleccionadas del algoritmo genético y tres variables latentes. Este modelo de regresión PLS se usó para predecir el rendimiento para tres cultivos adicionales: PHBE2, PHHB4 y PHB37, hibridados sobre dos sistemas de sondas de 44.000 oligonucleótidos.

Figura 2: muestra las predicciones de rendimiento basadas en un modelo de regresión PLS construido usando las relaciones de intensidad seleccionadas mediante algoritmo genético para todos los nueve cultivos: PHN46, PHR03, PHB73, PHW52, PHK29, PHW61, PHBE2, PHHB4 y PHB37, y relaciones de seis de los cultivos comparados con una medida replicada de PHP38, PHN46, PHR03, PHB73, PHW52, PHK29 y PHW61. El número de variables latentes aumentó hasta cinco y se realizó un autoescalado para contabilizar dicho ruido. Se realizó un centrado promedio sobre los datos de rendimiento.

Figura 3: muestra datos que ilustran la diversidad genética dentro de un genotipo de maíz. Se muestran datos representativos de dos oligos que muestran una variación de número de copia entre plantas.

Figura 4: muestra datos que ilustran la diversidad genética dentro de un grupo heterótico de maíz. Se muestran datos representativos que muestran variación de número de copia entre dos cultivos de maíz stiff stalk.

Figura 5: muestra datos que ilustran la diversidad genética entre dos grupos heteróticos de maíz. Se muestran datos representativos que muestran variación de número de copia entre un cultivo de maíz stiff stalk y un cultivo de maíz no stiff stalk.

Figura 6: muestra datos de predicción de rendimiento a partir de variaciones de número de copia detectadas mediante hibridación genómica comparativa.

Figura 7: muestra datos de predicción de altura de espiga a partir de variaciones de número de copia detectadas mediante hibridación genómica comparativa.

Figura 8: muestra datos de predicción de humedad a partir de variaciones de número de copia detectadas mediante hibridación genómica comparativa.

Figura 9: muestra datos de predicción de altura de planta a partir de variaciones de número de copia detectadas mediante hibridación genómica comparativa.

DESCRIPCIÓN DETALLADA

Los siguientes términos se usarán frecuentemente en la descripción mostrada a continuación. Las siguientes definiciones se proporcionan para facilitar el entendimiento de la descripción.

“Regiones codificadoras” significa regiones de un genoma de un organismo que codifican para proteínas o moléculas de ARN, en donde las regiones codificadoras y/o el ARN pueden incluir intrones, exones, secuencias reguladoras y regiones 5’ y 3’ no traducidas.

“Variación del número de copia” (CNV, en sus siglas en inglés) es un segmento de ADN para el cual se han observado diferencias del número de copia por comparación de dos o más genomas, o por comparación con una secuencia referencia. El término CNV abarca otra terminología para describir variantes que incluyen variantes de número de copia de gran escala (LCV), polimorfismos de número de copia (CNP) y variantes de tamaño intermedio (ISV).

“Variedad de planta híbrida F1” significa la primera generación filial que resulta de cruzar dos líneas parentales distintas.

“Fenotipo relacionado a heterosis” significa un rasgo observable en una planta en donde el fenotipo exhibido en plantas híbridas es más deseable cuando se compara con el correspondiente fenotipo exhibido en plantas parentales homocigotas.

“Intensidad de hibridación” significa una media de la cantidad de ADN genómico hibridado a una molécula sonda de oligonucleótido basada en un marcador cuantificable ligado al ADN genómico preparado. La cantidad de ADN preparado unido a la molécula sonda de oligonucleótido refleja la similitud de secuencia entre el ADN genómico y la molécula sonda de oligonucleótido, así como el número de copia de la región del ADN genómico ligado a la molécula sonda de oligonucleótido.

“Estructura de hibridación” significa una colección de las intensidades de hibridación para cada molécula sonda de oligonucleótidos única en una pluralidad de moléculas sonda de oligonucleótidos después de que las moléculas sonda se han puesto en contacto con una muestra de ADN o ARN.

“Sistema de oligonucleótidos” significa una pluralidad de moléculas sonda de oligonucleótido asociadas de forma estable a un soporte sólido.

“Moléculas sonda de oligonucleótido” significa secuencias cortas de ADN y/o ARN que se hibridarán de forma selectiva con una muestra preparada que contenga ADN y/o ARN.

“Valor-p” significa una medida de probabilidad de que una diferencia observada entre intensidades de hibridación se haya producido por casualidad. Por ejemplo, un valor-p de 0,01 (p = 0,01) significa que existe 1 posibilidad entre 100

de que el resultado se produjera por casualidad. Cuanto menor sea el valor-p, más probable es que la diferencia observada entre intensidades de hibridación esté causada por diferencias reales entre dos muestras.

"ADN genómico preparado" significa ADN de un organismo que ha sido digerido y/o separado y marcado con un marcador detectable. Se puede realizar una manipulación adicional de ADN, que incluye amplificación PCR del ADN antes de que el ADN sea digerido y/o separado, entre la etapa de digestión/separación y la etapa de marcado, o después de la etapa de marcado. También se pueden aplicar técnicas para seleccionar un subconjunto de ADN genómico, tales como, por ejemplo, escrutinio de enzima de restricción sensible a metilo, uso de curvas de fusión y selección basada en la velocidad de replegamiento, uso de Cot ADN, y otras similares. Dichos subconjuntos de ADN genómico se incluyen en esta definición.

"Variación estructural" se refiere a los cambios de estructura genética que se producen en el genoma. Se puede producir un amplio rango de variación estructural en el genoma, incluyendo eliminaciones, inserciones, duplicaciones e inversiones. Estas variaciones oscilan en tamaño, y normalmente se agrupan en tamaños de 1-500 pb (escala fina), 500 pb-100 kb (escala intermedia), y >100 kb (escala grande). Tal como se usa en la presente memoria, la variación estructural no incluye RFLPs.

Se puede usar cualquier método para detectar, cuantificar y/o analizar la variación de número de copia entre dos o más genomas. Por ejemplo, la variación de número de copia puede descubrirse mediante técnicas citogenéticas tales como la hibridación *in situ* fluorescente, la hibridación genómica comparativa, la hibridación genómica comparativa de sistema, la determinación de genotipo SNP a gran escala, el secuenciamiento completo del genoma, el mapeado acabador en par, el resecuenciamiento acabado en clon, análisis *in silico*, o combinaciones de los mismos. Opcionalmente, se pueden usar análisis con ordenador o estadísticos y/o modelización en conjunción con cualquier método CNV.

La detección de la variación del número de copia es distinta de la detección sencilla típica de polimorfismo de nucleótido. Se puede usar la hibridación con oligonucleótidos cortos sobre superficies sólidas para detectar polimorfismos de nucleótido singulares (SNP) (Chee y col. (1996) Science 274: 610-614). En esta aplicación de detección SNP normalmente se usan oligonucleótidos de 20-22 meros para maximizar la capacidad para detectar diferencias singulares entre la sonda y la diana (Lipshutz y col. (1995) Biotechniques 19: 442-447). Oligonucleótidos de mayor longitud, tales como los de 60 meros usados en el Ejemplo 1 para hibridar CGH con afinidad muy similar a dianas perfectamente igualadas y a dianas con una o más diferencias. Por tanto, dichas sondas de oligonucleótido no son adecuadas para la detección SNP. Estas sondas de mayor longitud normalmente son muy sensibles a la presencia o ausencia de la secuencia diana, o a los cambios grandes en la cantidad de secuencia diana, y por tanto son útiles para detectar variaciones de número de copia. En el maíz, existen polimorfismos en regiones codificadoras con una frecuencia global de menos de 1 SNP/100 pb (Ching y col., (2002) BMC Genet 3: 19). La mayoría de las sondas usadas en los Ejemplos 1 – 2 contiene 0-1 diferencias en comparación con el ADN genómico, y se hibridan bien con la diana. En las raras ocasiones de eliminación o multiplicación de la diana en el genoma, es de esperar que dichas sondas muestren relaciones numéricamente grandes de señal de hibridación entre cultivos diferentes, y que se desvíen de las relaciones 1:1 esperadas para dianas que son idénticas o que contienen 1 diferencia. En los Ejemplos 1 – 2 se seleccionaron sondas con una elevada relación de hibridación observada entre diferentes cultivos de maíz, aunque no se realiza una representación específica de las diferencias moleculares subyacentes bajo dichas relaciones de hibridación, excepto que es improbable que sean debidas a la presencia de diferencias de 1-2 pb entre la sonda y la diana.

En un ejemplo, el método descrito en la presente memoria utiliza CGH para predecir el grado de uno o más fenotipos heteróticos en variedades de plantas híbridas. El método descrito permite la selección de líneas parentales cultivadas, a la vez que evita la necesidad de realizar ensayos de cruces que consumen recursos con un número elevado de potenciales líneas parentales. Este método puede usarse con un número de moléculas de sonda de oligonucleótido que oscila entre un número grande y un número inesperadamente pequeño de moléculas sonda de oligonucleótido para la predicción del grado de fenotipos heteróticos. La selección de moléculas sonda de oligonucleótido se puede facilitar mediante el uso de un procedimiento de optimización, un ejemplo del cual se describe en la presente memoria. Adicionalmente, el método CGH descrito proporciona un aumento inesperadamente significativo de la capacidad predictiva sobre técnicas usadas actualmente en la modificación de plantas. El uso de CGH también elimina muchas de las dificultades que se producen por el uso de ARNm para la predicción del grado de uno o más fenotipos heteróticos en plantas, ya que el ADN genómico es el mismo en cada célula somática de la planta (aparte de los gametofitos) independientemente de la etapa de desarrollo, las condiciones ambientales o el tejido analizado. Estos resultandos indican que la CGH es un ensayo fiable para la predicción del grado de uno o más fenotipos heteróticos en plantas.

Una revisión de CGH, que incluye las consideraciones generales y una descripción de la tecnología, puede encontrarse en Pinkel & Albertson, *Nature Genetics* 37, S11-S17 (2005). En la siguiente descripción se asume, por tanto, la familiaridad con la tecnología CGH de los especialistas en la técnica. El uso del método reivindicado para la predicción del grado de uno o más fenotipos heteróticos incluye la selección de una pluralidad de moléculas sonda de oligonucleótido, obtener una muestra de ADN genómico, preparar el ADN genómico, la hibridación de la muestra de ADN con las moléculas sonda de oligonucleótidos, la detección de las intensidades de hibridación resultantes, la comparación de las intensidades detectadas con resultados de una o más muestras diferentes que tengan fenotipos

heteróticos y predecir el fenotipo heterótico de plantas progenie derivadas de las plantas de donde se ha obtenido el ADN genómico.

Un modo de mejorar los métodos descritos es la selección de la pluralidad de moléculas sonda de oligonucleótidos. En un ejemplo, la pluralidad de moléculas sonda de oligonucleótido comprende un sistema de oligonucleótidos. En algunos ejemplos, se puede usar un sistema de oligonucleótidos diseñado para el análisis de ARNm como pluralidad de moléculas sonda de oligonucleótidos. Opcionalmente, el sistema de oligonucleótidos comprende moléculas sonda de oligonucleótidos que cubren el genoma completo de la planta, con muestreo redundante de cada región del genoma, así como controles positivos y negativos. En algunos ejemplos, el sistema de oligonucleótidos comprende moléculas sonda de oligonucleótidos que se sabe predicen el grado de fenotipo heterótico de la planta diana.

Cuando se seleccionan moléculas sonda de oligonucleótido para su uso, se pueden considerar factores tales como el tamaño molecular, la composición molecular y la localización genómica de las moléculas seleccionadas. Respecto al tamaño molecular, las moléculas más pequeñas son menos capaces de hibridarse con secuencias que contengan diferencias, que incluyen inserciones, eliminaciones o sustituciones, pero son menos susceptibles a la formación de estructuras secundarias. Las moléculas sonda de oligonucleótidos de mayor longitud son más capaces de hibridarse con ADN que contenga diferencias, pero también son más susceptibles a la formación de estructuras secundarias.

Las moléculas sonda de oligonucleótido que forman estructuras secundarias son menos capaces de hibridarse con el ADN genómico de la muestra preparada. La predicción de estructuras secundarias en secuencias de oligonucleótidos es bien conocida y existen diferentes paquetes de software que son capaces de predecir formaciones de estructura secundaria y las propiedades termodinámicas, tales como mFOLD (Zuker y col. (1999) Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide in RNA Biochemistry and Biotechnology, Barciszewski & Clark, eds., NATO ASI Series, Kluwer Academic Publishers) y RNAfold (Vienna RNA Package; Hofacker y col. (1994) Monatshefte f. Chemie 125: 167-188; Zuker & Stiegler (1981) Nucl Acids Res 9: 133-148). Usando estas herramientas es posible equilibrar la cobertura de localizaciones genómicas con la probabilidad de formación de estructura secundaria. Cuando se usa un conjunto de moléculas sonda de oligonucleótido exhaustivo, las moléculas sonda de oligonucleótidos pueden seleccionarse de tal modo que se cubra el genoma completo de la planta múltiples veces con sondas que no es probable que formen estructuras secundarias. Cuando se usa un conjunto de moléculas sonda de oligonucleótido más pequeño, se pueden seleccionar las sondas para cubrir las regiones genómicas de interés con una cobertura redundante, a la vez que manteniendo una baja probabilidad de formación de estructuras secundarias.

Las moléculas sonda de oligonucleótido usadas en los métodos generalmente tienen una longitud de entre 20 y 100 nucleótidos. En algunos ejemplos, las moléculas sonda de oligonucleótidos tienen una longitud de 60 nucleótidos. Por supuesto, las moléculas sonda de oligonucleótidos de una pluralidad dada no necesitan ser todas de longitud uniforme, y en algunos ejemplos que el tener moléculas sonda de oligonucleótidos de diferentes longitudes puede utilizar o compensar las características variables de las moléculas sonda de oligonucleótidos de diferentes longitudes descritas anteriormente.

La calidad de los datos producidos mediante el método puede mejorar con la incorporación de más de una molécula sonda de oligonucleótido por gen o región genómica de interés. La inclusión de dichas moléculas sonda de oligonucleótido redundantes proporciona controles internos para determinar si las intensidades de hibridación diferentes son el resultado de una diferencia en el número de copia de un gen o región cromosomal o ruido aleatorio. En algunos ejemplos, se incluye más de una molécula sonda de oligonucleótido por gen o región de ADN de interés en la pluralidad de moléculas sonda de oligonucleótidos. En algunos ejemplos se usan tres moléculas sonda de oligonucleótidos para cada gen o región de interés.

El proceso de creación de sistemas de oligonucleótidos es bien conocido y se dispone de una serie de máquinas comerciales para crear sistemas de oligonucleótidos, tales como BioOdyssey Calligrapher MiniArrayer de BioRad. Adicionalmente, existe una serie de servicios comerciales que crean sistemas de oligonucleótidos a partir de una lista de secuencias de moléculas sonda de oligonucleótidos, tal como el servicio de impresión de sistemas SurePrint de Agilent. La pluralidad de moléculas sonda de oligonucleótido normalmente incluye al menos aproximadamente cien moléculas sonda de oligonucleótidos, pero puede incluir cualquier número de moléculas sonda de oligonucleótidos entre aproximadamente 100 y aproximadamente 80.000 moléculas sonda de oligonucleótidos, o más si se desean mayores intervalos de ensayo. Adicionalmente, la pluralidad de moléculas sonda de oligonucleótidos puede diseñarse para incluir cualquier número de controles positivos o negativos para asegurar la validez de los datos adquiridos mediante el uso de la pluralidad de moléculas sonda de oligonucleótidos.

Otro aspecto del método reivindicado es la preparación de ADN genómico antes de poner en contacto con la pluralidad de moléculas sonda de oligonucleótido. La preparación y el etiquetado de ADN genómico son bien conocidos, y se dispone de kits para la preparación de ADN genómico para CGH, tales como “Genomic DNA Labeling Kit PLUS” (Agilent). Se aísla el ADN genómico de cada línea paterna y se etiqueta individualmente. Normalmente, se usan cantidades aproximadamente iguales de ADN de cada línea paterna, si no la precisión de los resultados en relación a las diferencias de número de copia puede verse resentida, y por tanto ser potencialmente menos efectiva como predicción del grado de un fenotipo relacionado a heterosis de interés. La cantidad de ADN

genómico aislado requerido depende de una serie de factores, que incluyen el tamaño del sistema de oligonucleótidos y los protocolos usados. Cuando se usa un sistema de oligonucleótidos de tamaño medio (entre aproximadamente 40.000 y 100.000 moléculas sonda de oligonucleótido) siguiendo protocolos estándares, la cantidad de ADN genómico usado normalmente se encuentra entre 0,2 y 3,0 µg. Cuando la muestra no contiene suficiente ADN genómico para una hibridación directa, se puede usar cualquier técnica de amplificación conocida (por ejemplo, amplificación mediante PCR) para aumentar la cantidad de ADN genómico preparada.

Normalmente, una vez que se dispone de una cantidad suficiente de ADN genómico, el ADN genómico se fragmenta usando técnicas estándares tales como digestión con al menos una endonucleasa de restricción, separación mecánica, o una combinación de ambas, para proporcionar fragmentos de ADN genómico de longitud relativamente uniforme. La muestra de ADN genómico fragmentado a continuación puede purificarse, cuantificarse y concentrarse usando técnicas estándares. Los fragmentos de ADN genómico concentrado resultantes pueden etiquetarse en una reacción PCR usando cebadores aleatorios y moléculas dUTP marcadas teniendo cada línea paterna una marca fluorescente única. Si se usan diferentes sistemas de oligonucleótidos para cada línea paterna, entonces es posible usar la misma etiqueta con ambos padres, aunque normalmente ambas muestras son analizadas en un único sistema. Opcionalmente, también es posible usar más de dos etiquetas para padres potenciales adicionales.

Generalmente, se extrae el ADN genómico de muestras de tejido que son frescas o que están congeladas. Se puede usar cualquier método de almacenamiento de tejidos, siendo el objetivo reducir la degradación del ADN genómico. Adicionalmente, se puede mejorar la fortaleza de la señal eliminando el ADN de baja complejidad usando técnicas estándares tales como escrutinios de genoma con enzima de restricción sensible a metilo, el uso de curvas de fusión con selección basada en la velocidad de replegamiento, y el uso de Cot ADN para precipitar secuencias de baja complejidad.

Tras la preparación del ADN, el ADN preparado se pone en contacto con la pluralidad de moléculas sonda de oligonucleótido. El ADN genómico preparado y etiquetado normalmente se pone en contacto con un sistema de oligonucleótidos en condiciones de hibridación estrictas. Las técnicas y condiciones requeridas para la hibridación de una muestra de ADN con sistemas de oligonucleótidos son conocidas, y se encuentran disponibles comercialmente kits que contienen las disoluciones y tampones requeridos, tales como el kit Oligo aCGH/ChIP-on-chip Hybridization Kit (Agilent, Santa Clara, CA, EE.UU.). El ADN genómico preparado a partir de los padres normalmente se hibrida con el mismo sistema de oligonucleótidos. Alternativamente, el ADN genómico preparado y etiquetado de cada padre puede hibridarse con diferentes sistemas a diferentes tiempos, siempre que los diferentes sistemas contengan al menos algún subconjunto de moléculas sonda de oligonucleótidos comunes. En algunos ejemplos, el ADN de cada padre se hibrida con dos sistemas separados pero idénticos de moléculas sonda de oligonucleótido en las mismas condiciones de hibridación.

Tras poner en contacto el ADN preparado con la pluralidad de moléculas sonda de oligonucleótidos, se detectan las intensidades de hibridación generadas mediante la hibridación del ADN genómico con las moléculas sonda de oligonucleótidos. Opcionalmente, se usa un escáner de microsistema (tal como el Agilent DNA Microarray Scanner) para detectar las intensidades de hibridación. Las intensidades de hibridación detectadas normalmente se presentan en un software asociado al escáner, y pueden exportarse opcionalmente en una serie de formatos de archivo para un procesado avanzado. El software de análisis de datos puede generar estadísticas basadas en las intensidades de hibridación detectadas. Esto permite al investigador determinar el número de sondas que muestran diferentes intensidades de hibridación y el grado de las diferencias de intensidad. En algunos ejemplos, el software se usa para determinar el número de diferencias, la diferencia de plegamiento, o ambas, de moléculas sonda de oligonucleótido que presentan una diferencia superior a 1,5 veces en la intensidad de hibridación. Opcionalmente, el software puede usarse para determinar el número de moléculas sonda de oligonucleótido que presentan al menos una diferencia de 2 veces en la intensidad de hibridación. En algunos ejemplos, el software puede usarse para determinar el número de moléculas sonda de oligonucleótido que presentan más de tres veces, pero menos de diez veces, de diferencia en la intensidad de hibridación. Por supuesto, se pueden usar otros valores para la mínima diferencia y/o para la máxima diferencia, si uno quiere estrechar o ampliar el grupo de intensidades de hibridación relevantes. Por ejemplo, las diferencias mínimas pueden incluir cualquier valor entre 1,5 y 10 veces de diferencia, y la diferencia máxima puede incluir cualquier valor entre 1,5 y 50 veces de diferencia. Estos valores de corte mínimo y máximo pueden usarse independientemente (por ejemplo, todas las moléculas sonda de oligonucleótidos que presenten una diferencia en la intensidad de hibridación superior a 1,7) o en conjunto (por ejemplo, todas las moléculas sonda de oligonucleótido que presenten una diferencia superior a 2,1 pero inferior a 11,4 veces) para proporcionar conjuntos de datos para un posterior procesado.

En otro ejemplo, se pueden usar métodos de secuenciamiento de genoma completo para detectar la variación del número de copia. En 1979 (Staden (1979) Nucl Acids Res 6: 2601-2610) se usó el secuenciamiento de escopeta de genoma completo para genomas pequeños (de 4.000 a 7.000 pb). La metodología ha evolucionado para permitir el secuenciamiento de genomas de mayor tamaño y complicación, incluyendo el genoma de la mosca de la fruta y el genoma humano. En general, el ADN de elevado peso molecular es separado en fragmentos aleatorios, se selecciona por tamaño (normalmente 2, 10, 50 y 150 kb) y se clona en un vector apropiado. Los clones son secuenciados desde los dos extremos, normalmente usando un método de terminación de cadena para producir dos secuencias cortas. Cada secuencia se denomina una lectura-final o lectura, y dos lecturas del mismo clon reciben el nombre de pares compañeros. El método de terminación de cadena produce típicamente lecturas de

aproximadamente 500-1.000 bases, y por tanto los pares compañeros raramente se solapan. La secuencia original se reconstruye a partir de todas las lecturas usando un software de montaje de secuencias. Las lecturas que solapan se recogen en secuencias compuestas de mayor longitud conocidas con contiguos. Los contiguos pueden estar unidos unos a otros en estructuras siguiendo las conexiones entre pares compañeros. La distancia entre contiguos puede deducirse de las posiciones de pares compañeros si se conoce la longitud de fragmento promedio de la biblioteca y si tiene una ventana de desviación estrecha. Se dispone de muchas tecnologías de secuenciamiento que usan métodos de gel, métodos capilares, métodos de partículas o métodos de sistemas. Las tecnologías de secuenciamiento que avanzan rápidamente incluyen el secuenciamiento por síntesis, los sistemas de partículas paralelos, los microchips electrónicos, los biochips, los microchips paralelos, el secuenciamiento mediante ligación, el secuenciamiento de molécula de ADN sencilla y el nanoporo-secuenciamiento. En este ejemplo, las eliminaciones/inserciones se detectarían mediante la alineación de las secuencias respecto a un genoma de referencia. Se detectarían las CNVs contando el número de veces que se observa una etiqueta/secuencia y a continuación comparar el número con otra muestra o genoma de referencia.

Se utilizó una estrategia *in silico* para comparar dos genomas humanos a nivel de secuencia de ADN (Tuzun y col. (2005) Nat Genet 37: 727-732). Se adoptó como genoma de referencia la secuencia del genoma humano en NCBI. Aproximadamente el 67% de esta secuencia de referencia procedía de una única biblioteca de ADN (la biblioteca RPCI-11 BAC) de un único individuo. El segundo genoma comprendía pares de lecturas de secuencia-finales de >500.000 clones fósmidos de la biblioteca de ADN G248. Esta biblioteca de ADN fue derivada de una mujer anónima de Norte América con ancestros europeos. Puesto que los tamaños de los clones fósmidos están estrechamente regulados en aproximadamente 40 kb, se esperaba que los pares de secuencias finales para cualquier clon fósmido dado se alinearan con la secuencia de referencia con aproximadamente un espaciado de 40 kb. Una desviación significativa del espaciado de alineamiento (es decir, <32 kb ó >48 kb) sugería la presencia de una CNV en dicha localización. Usando este criterio se identificaron 241 CNVs, estando la mayoría en el intervalo de 8 kb a 40 kb, y el 80% de las mismas no habían sido identificadas previamente. Asimismo, la mayoría de dichas CNVs estaba por debajo de la resolución esperada de las plataformas del sistema usadas en estudios de CNV anteriores. Una ventaja sobre los métodos basados en sistemas es que la estrategia *in silico* también detecta otras variantes genómicas estructurales, por ejemplo inversiones. Estas variantes estructurales pueden detectarse a través de discrepancias consistentes en la orientación alineada de secuencias múltiples de extremo emparejadas.

La quimiométrica es la aplicación de métodos matemáticos o estadísticos para el diseño experimental y/o el análisis de datos. Se puede usar la quimiométrica para identificar información adicional a partir de dichos datos usando varios métodos que incluyen estadística, reconocimiento de patrones, modelización, estimaciones de relaciones estructura-propiedad, o combinaciones de los mismos. Por ejemplo, los datos pueden ser datos de hibridación, datos normalizados, datos de secuenciamiento, productos de análisis de secuencia tales como contiguos, alineamientos, puntuaciones de similitud, puntuaciones de valor esperado, valores-p, indelos u otros datos generados mediante un método para detectar variaciones estructurales genómicas.

En algunos ejemplos, se usa el software de análisis de datos para calcular valores-p en base a las diferencias medidas en la intensidad de hibridación. Estos valores pueden usarse como sustitutos, o además, de las diferencias de plegamiento en la intensidad entre moléculas sonda de oligonucleótidos. Cuando se usa el valor-p en lugar de la diferencia de plegamiento uno puede aumentar la restricción disminuyendo el valor-p máximo considerado. Por ejemplo, un investigador puede desear aplicar un valor de corte de baja restricción seleccionando todas las moléculas sonda de oligonucleótido en las que la diferencia de intensidad de hibridación dio lugar a un valor-p inferior a 0,1. La restricción se puede aumentar disminuyendo el valor-p máximo a 0,05, 0,01, 0,001 o a cualquier valor en el intervalo entre 0,01 y 0,001.

Una vez que se han tomado los datos, se puede predecir el grado de un fenotipo heterótico en base a los resultados obtenidos. Dicha predicción se realiza comparando el número de sondas que cumplen el nivel definido por el usuario durante el análisis al número de sondas que muestran los mismos criterios en otras hibridaciones que implican líneas parentales en las que se conoce el fenotipo heterótico en la progenie híbrida F1 resultante. Adicionalmente, se pueden usar técnicas estadísticas comunes, tales como regresión lineal, para llevar a cabo la predicción.

Opcionalmente, el grado predicho de uno o más tipos heteróticos se puede usar para seleccionar líneas parentales para el desarrollo de líneas de plantas híbridas F1 como parte de un programa de cultivo de plantas. Los programas modernos de cultivo de plantas consideran una amplia variedad de factores cuando seleccionan plantas para el cultivo. En otro ejemplo, el grado predicho de un fenotipo heterótico se incluye entre los factores y forma parte al menos del razonamiento para seleccionar dos líneas parentales para el cultivo en un programa comercial o en otro programa de cultivo de plantas.

Se pueden usar métodos para desarrollar una pluralidad de moléculas sonda de oligonucleótidos especializadas para la predicción del grado de uno o más fenotipos heteróticos. La identificación de moléculas sonda de oligonucleótidos que son predictivas de fenotipos heteróticos en una planta diana se puede realizar a través del uso de una estrategia empírica. En un ejemplo, se crea una serie de líneas de plantas híbridas F1 y se cultivan en condiciones controladas y se mide el fenotipo heterótico de interés. Usando un sistema de oligonucleótidos, normalmente uno que cubra una mayor cantidad del genoma de la planta, se lleva a cabo la CGH de las líneas parentales. Las intensidades de hibridación resultantes se analizan para determinar las moléculas sonda de

oligonucleótido que demuestran una mejor capacidad para predecir el grado de fenotipo heterótico en las líneas de plantas híbridas F1 medidas. Las moléculas sonda de oligonucleótido que son mejores predictores se usan entonces en un sistema de oligonucleótidos mejorado para predecir el grado de un fenotipo heterótico, tanto en sustitución como además de un sistema de oligonucleótidos exhaustivo, tal como se ha descrito antes.

En algunos ejemplos, el análisis de las intensidades de hibridación se lleva a cabo usando una estrategia computacional evolucionaria iterada. En dicha estrategia, el software forma subgrupos arbitrarios de las moléculas sonda de oligonucleótido y usa análisis por regresión para determinar la capacidad predictiva de los subconjuntos sonda. La regresión puede acoplarse con un método de aprendizaje mecánico y usarse para seleccionar los subgrupos de moléculas sonda de oligonucleótido que demuestran mejores propiedades de predicción de fenotipos heteróticos. Los tipos de análisis de regresión que pueden usarse incluyen, por ejemplo, regresión de componente principal, cuadrados mínimos clásicos, cuadrados mínimos inversos y cuadrados mínimos parciales. Los métodos de aprendizaje mecánicos que pueden usarse incluyen, por ejemplo, máquinas de vector soporte y redes neurales. La regresión y el aprendizaje mecánico se pueden usar individualmente o en combinación para llevar a cabo el análisis. La selección de predictor de intensidad de hibridación en el análisis de regresión sólo se puede hacer como se muestra en algunos ejemplos usando la variable de proyección de importancia dentro del espacio de representación PLS. El proceso de formación de subgrupos y de selección de mejores predictores a través del uso de regresión y aprendizaje mecánico también puede repetirse hasta un punto definido por el usuario. En algunos ejemplos, el proceso se itera hasta que sólo se producen ligeros aumentos de la capacidad predictiva de los subconjuntos. En otros ejemplos, el proceso se itera hasta que no hay ningún incremento en la capacidad predictiva de los subconjuntos.

Opcionalmente, se crea un sistema de oligonucleótidos que comprende las moléculas sonda de oligonucleótido identificadas. En algunos ejemplos, el sistema de oligonucleótidos creado es parte de un kit para la predicción del grado de uno o más fenotipos heteróticos en una planta que está disponible para su venta comercial o para su uso interno.

Los siguientes ejemplos ilustran con más detalle la presente invención y no pretenden limitar las reivindicaciones en modo alguno. La presente invención puede llevarse a la práctica usando muchas variaciones diferentes y se ha demostrado a través de ejemplos ilustrativos. La invención no se limita a las realizaciones descritas, sino que también incluye todas las modificaciones, equivalentes y alternativas que caigan dentro del espíritu y del alcance de la invención según se establece en las reivindicaciones.

Ejemplo 1: Hibridación de Genoma Comparativa (CGH) en Maíz

ADN genómico:

Se obtuvo ADN genómico de los siguientes cultivos de maíz: PHP38, PHK29, PHW61, PHR03, PHW52, PHN46, PHHB4, PHBE2, PHB37, PH1FA, PHT11 y PHB47. Se aisló el ADN celular total a partir de muestras de hojas congeladas frescas usando kits DNeasy Plant Mini (Qiagen) que incluyen una incubación con ARNasa siguiendo las instrucciones del fabricante. Las muestras fueron cuantificadas con un espectrofotómetro y se aplicaron a un gel de agarosa para comprobar su integridad.

aCGH:

Para cada hibridación CGH, se digirieron 2 µg de ADN genómico con enzimas de restricción AluI y RsaI (Promega). Después de una incubación de dos horas, las muestras fueron calentadas a 65ºC durante 20 minutos para desactivar las enzimas. El ADN fragmentado fue marcado mediante una reacción de etiquetado imprimada aleatoriamente (Agilent Oligonucleotide Array-Based CGH for Genomic DNA analysis, v4.0) que incorporó Cy3-UTP al producto. El ADN marcado se filtró con una columna Microcon YM-30 (Millipore) para eliminar los nucleótidos no incorporados. Las muestras fueron cuantificadas con un espectrofotómetro Hitachi para medir el rendimiento y las tasas de incorporación de colorante. Se añadieron tampones de hibridación y de bloqueo (Agilent Technologies) a las muestras antes de ser desnaturalizadas a 95ºC durante 3 minutos y se incubaron a 37ºC durante 30 minutos. Se hibrido cada muestra con un sistema durante 40 horas a 65ºC a la vez que se rotaba a 10 rpm. Los sistemas fueron desmontados y lavados en tampón de lavado Oligo aCGH 1 (Agilent Technologies) a temperatura ambiente durante 5 minutos. Se llevó a cabo un segundo lavado en tampón de lavado Oligo aCGH 2 (Agilent Technologies) durante 1 minuto a 37ºC. Se sumergieron láminas en acetonitrilo y se secaron al aire. Se utilizó un escáner de microsistema de ADN Agilent G2505B para capturar imágenes TIF.

Microsistemas de oligonucleótidos:

Se utilizaron microsistemas Custom 44K (Agilent Technologies) que contienen 82.272 oligómeros únicos de 60 meros que expanden dos microsistemas dirigidos a secuencias expresadas del genoma del maíz para la hibridación de los siguientes cultivos: PHP38, PHK29, PHW61, PHR03, PHW52, PHN46, PHHB4, PHBE2 y PHB37. Adicionalmente, se utilizó un microsistema custom 2x105K (Agilent Technologies) que contiene 102.349 oligómeros únicos de 60 meros, de los cuales 82.272 oligómeros estaban representados en los anteriores sistemas de 44K, para la hibridación de los siguientes cultivos: PHP38, PHK29, PHW61, PHR03, PHW52, PHN46, PHHB4, PHBE2, PHB37, PH1FA, PHT11 y PHB47.

Análisis de imágenes y datos:

Las imágenes de microsistemas fueron inspeccionadas visualmente en búsqueda de artefactos de imagen. Se extrajeron las intensidades características, se filtraron y se normalizaron con el software Feature Extraction de Agilent (versión 9.5.1). Se realizó un control de calidad adicional utilizando herramientas de análisis de datos de la base de datos Rosetta's Resolver.

Nebulización frente a digestión RE

Las muestras fueron separadas aleatoriamente mediante nebulización. Se mezclaron de 4 a 6 µg de muestras de ADN purificadas, en un volumen total de 50 µL, en el nebulizador con 700 µL de tampón de nebulización (25% de glicerol, Tris-HCl 50 mM, $MgCl_2$ 15 mM). El nebulizador se enfrió en hielo y se conectó a una fuente de aire comprimido. Se suministró aire a una presión de 2,2 bar (32 psi) durante 6 minutos. El nebulizador fue girado hacia abajo y se recuperó la disolución de ADN. El ADN fue purificado en una columna de purificación QIAquick® PCR (Qiagen) y eluído en 30 µL de Tris-HCl 10 mM de pH 8,5. Se usaron 0,5 µg de ADN separado aleatoriamente para las etapas de etiquetado e hibridación descritas previamente.

Después de la hibridación, se compararon los datos del digesto de enzima de restricción (RE en sus siglas en inglés) y de muestras separadas aleatoriamente para determinar si había alguna diferencia con la metodología de preparación de muestra. La comparación de muestras nebulizadas frente a muestras digeridas con RE demostró una alta correlación de cambios de plegamiento (R cuadrado = 0,89). Por lo tanto, no existen diferencias importantes con los datos cuando se utilizan diferentes métodos de preparación de muestra.

Ejemplo 2: Análisis de regresión

Las relaciones de intensidad de sistema CGH, los valores, los números de acceso y las secuencias sonda de oligonucleótido fueron exportados a formato de texto ASCII usando el Rosetta Resolver 6.0 (Rosetta Biosoftware, Seattle, WA). Las intensidades CGH fueron importadas y alineadas para cada cultivo y sistema en el entorno de computación técnica de Matlab (versión 7.4.0, Mathworks, Natick, MA) usando tanto números de acceso como secuencias de oligonucleótidos. Se llevó a cabo la selección de relación de intensidad de algoritmo genético usando análisis de regresión de cuadrados mínimos parciales usando el PLSToolbox 4.0 (Eigenvector Research, Wenatchee, WA) en el espacio de trabajo de Matlab. Todos los cálculos se realizaron en un Dell Latitude D620 con un procesador Intel duo core de 1,8 GHz usando modo multitarea.

Los valores de relación de intensidad procedentes de los dos sistemas de sondas de 44.000 oligonucleótidos descritos anteriormente fueron montados para los cultivos PHB73, PHW61, PHR03, PHK29, PHW52 y PHN46. Para el método del ejemplo mostrado aquí se usaron valores-p inferiores a 0,01 para reducir el número de relaciones de intensidad de candidato predictivas de algoritmo genético desde 82435 a 2786. Todas las intensidades y las intensidades seleccionadas mediante criterios de cambio de plegamiento también han sido usadas como entradas del algoritmo genético.

El algoritmo genético aplicado a la selección de relación de intensidad predictiva fue la función gaselctr.m de PLSToolbox. El algoritmo se aplicó a un tamaño de población inicial de 256 conjuntos de relación de intensidad únicos con un 10% de las 2786 relaciones seleccionadas en cada individuo. Se llevó a cabo la regresión de cuadrados mínimos parciales (PLS, de sus siglas en inglés) del rendimiento respecto de las relaciones de intensidad seleccionadas para realizar la predicción de rendimiento. Los conjuntos de relaciones de intensidad fueron clasificados por su error de predicción de rendimiento en PLS. Se realizaron diez veces cien generaciones de combinación de cruce doble usando los mejores 128 conjuntos de relación de intensidad individual. El número de variables latentes en la regresión PLS se fijó a un máximo de tres. Las 201 relaciones de intensidad seleccionadas mediante dicho método de selección de variable de algoritmo genético predijeron el rendimiento con el menor error de raíz cuadrada en validación de cruce dejando uno fuera entre los 100.000 conjuntos de relaciones de intensidad evaluados mediante el algoritmo genético y un modelo de regresión construido con todas las relaciones de intensidad.

Se construyó un modelo de regresión PLS usando las relaciones de intensidad seleccionadas del algoritmo genético y tres variables latentes. Este modelo de regresión PLS se usó para predecir el rendimiento para tres cultivos adicionales, PHBE2, PHHB4 y PHB37, hibridados sobre dos sistemas de sonda de 44.000 oligonucleótidos. Estas predicciones de rendimiento sirvieron como validación del modelo y del método de selección de relación de intensidad. Las predicciones se muestran en la **Figura 1**. Las comparaciones de predicción indicadas con un triángulo son para los cultivos que son una parte del modelo de regresión. Los asteriscos indican la predicción de las muestras de calibrado.

A continuación se construyó un modelo de regresión PLS usando las relaciones de intensidades seleccionadas mediante el algoritmo genético para los nueve cultivos, PHN46, PHR03, PHB73, PHW52, PHK29, PHW61, PHBE2, PHHB4 y PHB37, y las relaciones de seis de los cultivos comparadas con una medida replicada de PHP38, PHN46, PHR03, PHB73, PHW52, PHK29 y PHW61. Las réplicas contribuyeron al ruido de intensidad de relación en la construcción del modelo. El número de variables latentes aumentó hasta cinco y se llevó a cabo un autoescalado para la relación de intensidad a fin de contabilizar dicho ruido. Se realizó un centrado en la media con los datos de

rendimiento. Las predicciones se muestran en la **Figura 2** para las relaciones de intensidad derivadas de los sistemas con 20.000 oligonucleótidos adicionales procedentes de regiones codificadoras del genoma. Los nuevos sistemas se hibridaron para los nueve cultivos mencionados y para tres nuevos cultivos PH1FA, PHT11 y PHB47. Las comparaciones de rendimiento predicho y de rendimiento medido para los nuevos cultivos se indican con asteriscos. Las muestras de calibración del modelo de regresión PLS se indican mediante triángulos. El error de raíz cuadrada media de la predicción para los nuevos cultivos fue de 9 bu/ac.

Los valores de heterosis predichos serán una aproximación del cambio de rendimiento (bu/ac). Este método puede usarse como escrutinio preliminar de germoplasma, particularmente de nuevo germoplasma, y se puede usar para seleccionar un conjunto más pequeño para medidas experimentales de heterosis. En la presente solicitud, el método proporciona una reducción del número de líneas a evaluar en el campo.

Este método fue validado usando un conjunto mayor de muestras, y con más diversidad de genotipos. Se generaron datos de CGH esencialmente como se ha descrito en el Ejemplo 1, mediante hibridación contra sistemas de CGH de maíz en formato 2X105K. Las muestras para CGH se tomaron de 14 experimentos R2 que contenían plantas procedentes de 3 grupos de madurez relativa, representando 181 genotipos (91 stiff stalk y 90 no stiff stalk) que produjeron 914 híbridos. Los datos se analizaron para identificar oligonucleótidos asociados con heterosis usando datos fenotípicos que incluyen el rendimiento, la altura de espiga, la humedad, el peso de ensayo, el verdor, la altura de planta, la capacidad de almacenamiento y las raíces. Los datos de este análisis se cruzaron y validaron con datos de mapeado cuando se disponía de ellos. Para el cultivo A stiff-stalk frente a 36 cultivos no stiff-stalk, se identificaron conjuntos de oligonucleótidos predictivos putativos para el rendimiento, la altura de espiga, la humedad y la altura de planta usando el método de proyección de importancia de variable descrito en el Ejemplo 5, y se muestran en las **Figuras 6-9**.

Ejemplo 3: Comparación de métodos de preparación de ADN genómico

ADN genómico:

Se obtuvo ADN genómico a partir de los siguientes cultivos de maíz: PHP38, PHK29, PHW61, PHR03, PHW52, PHN46, PHHB4, PHBE2, PHB37, PH1FA, PHT11 y PHB47. Se aisló el ADN celular total de muestras de hojas congeladas frescas usando kits DNeasy Plant Mini Kits (Qiagen) que incluyen una incubación con RNAsaA siguiendo las instrucciones del fabricante. Las muestras fueron cuantificadas con un espectrofotómetro y se analizaron en un gel de agarosa para comprobar la integridad.

aCGH:

Para cada hibridación CGH, se digirieron 2 µg de ADN genómico con enzimas de restricción AluI y RsaI (Promega). Tras una incubación de dos horas, las muestras fueron calentadas hasta 65ºC durante 20 minutos para desactivar las enzimas. El ADN fragmentado se marcó mediante una reacción de etiquetado imprimado aleatorio (Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis, v 4.0) que incorporó Cy3-UTP en el producto. El ADN marcado se filtró con una columna Microcon YM-30 (Millipore) para eliminar nucleótidos no incorporados. Las muestras se cuantificaron con un espectrofotómetro Hitachi para medir las tasas de rendimiento y de incorporación de colorante. Se añadieron tampones de hibridación y de bloqueo (Agilent Technologies) a las muestras antes de ser desnaturalizadas a 95ºC durante 3 minutos e incubadas a 37ºC durante 30 minutos. Se hibridó cada muestra con un sistema durante 40 horas a 65ºC a la vez que se rotaba a 10 rpm. Los sistemas fueron desmontados y lavados en tampón Oligo aCGH Wash Buffer 1 (Agilent Technologies) a temperatura ambiente durante 5 minutos. Se llevó a cabo un segundo lavado en tampón Oligo aCGH Wash Buffer 2 (Agilent Technologies) durante 1 minuto a 37ºC. A continuación se sumergieron láminas en acetonitrilo y se secaron al aire. Se utilizó un escáner de microsistemas de ADN Agilent G2505B para capturar las imágenes TIF.

Microsistemas de oligonucleótidos:

Se utilizaron microsistemas Custom 44K (Agilent Technologies) que contienen 82.272 oligonucleótidos únicos de 60 meros que expanden dos microsistemas dirigidos a secuencias expresadas del genoma del maíz para la hibridación de los siguientes cultivos: PHP38, PHK29, PHW61, PHR03, PHW52, PHN46, PHHB4, PHBE2 y PHB37. Adicionalmente, se utilizó un microsistema custom 2x105K (Agilent Technologies) que contiene 102.349 oligómeros únicos de 60 meros, de los cuales 82.272 oligómeros estaban representados en los anteriores sistemas de 44K, para la hibridación de los siguientes cultivos: PHP38, PHK29, PHW61, PHR03, PHW52, PHN46, PHHB4, PHBE2, PHB37, PH1FA, PHT11 y PHB47.

Análisis de imágenes y datos:

Las imágenes de microsistemas fueron inspeccionadas visualmente en búsqueda de artefactos de imagen. Se extrajeron las intensidades características, se filtraron y se normalizaron con el software Feature Extraction de Agilent (versión 9.5.1). Se realizó un control de calidad adicional utilizando herramientas de análisis de datos de la base de datos Rosetta's Resolver.

Nebulización frente a digestión RE

Las muestras fueron separadas aleatoriamente mediante nebulización. Se mezclaron de 4 a 6 µg de muestras de ADN purificadas, en un volumen total de 50 µL, en el nebulizador con 700 µL de tampón de nebulización (25% de glicerol, Tris-HCl 50 mM, $MgCl_2$ 15 mM). El nebulizador se enfrió en hielo y se conectó a una fuente de aire comprimido. Se suministró aire a una presión de 2,2 bar (32 psi) durante 6 minutos. El nebulizador fue girado hacia abajo y se recuperó la disolución de ADN. El ADN fue purificado en una columna de purificación QIAquick® PCR (Qiagen) y eluído en 30 µL de Tris-HCl 10 mM de pH 8,5. Se usaron 0,5 µg de ADN separado aleatoriamente para las etapas de etiquetado e hibridación descritas previamente.

Después de la hibridación, se compararon los datos del digesto de enzima de restricción (RE) y de muestras separadas aleatoriamente para determinar si había alguna diferencia con la metodología de preparación de muestra. La comparación de muestras nebulizadas frente a muestras digeridas con RE demostró una alta correlación de cambios de plegamiento (R cuadrado = 0,89). Por lo tanto, no existen diferencias importantes con los datos cuando se utilizan diferentes métodos de preparación de muestra.

Ejemplo 4: Diversidad genética

La metodología descrita en los Ejemplos 1-3 se usó para generar estimaciones de diversidad genética de variación del número de copia en genotipos de maíz seleccionados. Tal como se muestra en bibliografía, la investigación en humanos ha demostrado que existe variación del número de copia entre gemelos monocigóticos (Bruder y col. (2008) Am J Hum Genetic 82: 763-771).

A. Variación de planta

El ADN procedente de diez plantas de maíz del mismo genotipo se sometió a una hibridación de genoma comparativa y se analizó esencialmente como se ha descrito en los Ejemplos 1-3 para identificar CNVs putativas entre las plantas individuales. La variación observada entre plantas oscila entre 0,09% y 0,38%. También se determinó la variación técnica, y se estimó que era del 0,08%. Los datos representativos correspondientes a dos CNVs putativas que muestran Logaritmo de intensidad frente a número de planta se muestran en la **Figura 3**.

B. Variación dentro de un grupo heterótico

Con el objetivo de estimar la diversidad dentro de un grupo heterótico de maíz, se analizó ADN aislado procedente de dos cultivos del grupo heterótico stiff-stalk como se ha descrito en los Ejemplos 1-3 para identificar variaciones de número de copia. La variación observada se representó como una relación logarítmica de los dos genotipos para cada cromosoma individual, tal como se muestra en la **Figura 4**.

C. Variación entre grupos heteróticos

Con el fin de estimar la diversidad entre dos grupos heteróticos de maíz, se usó ADN aislado de dos cultivos, un cultivo stiff stalk y un cultivo no stiff stalk. El ADN fue analizado tal como se ha descrito en los Ejemplos 1-3 para identificar variaciones del número de copia. La variación observada se representó como una relación logarítmica de los dos genotipos para cada cromosoma individual, tal como se muestra en la **Figura 5**.

Ejemplo 5: Quimiométrica

Se ha aplicado quimiométrica a los datos de hibridación para identificar los oligómeros con probabilidad de ser predictivos de al menos un fenotipo heterótico. Los análisis descritos en el Ejemplo 2 también son métodos quimiométricos que pueden aplicarse a datos de variación estructural genómica.

En general, el objetivo de los análisis de quimiométrica fue predecir las propiedades de la planta en base a los datos de intensidad de CGH. Los análisis fueron optimizados mediante la selección de variables, que incluye algoritmos basados en el pre-procesado y en la predicción. El análisis fue validado usando uno o más ensayos que incluyen un ensayo de calibrado “deja uno fuera”, la predicción para una nueva muestra del grupo heterótico, y/o la comparación de oligonucleótidos seleccionados con marcadores conocidos o datos de mapeado. El pre-procesado incluye etapas tales como la clasificación de datos basada en la intensidad de hibridación: no variación en la intensidad CGH de referencia; un cambio de menos de 10 veces en la intensidad; y un cambio de más de 2 veces en la intensidad. La selección de variable basada en la predicción incluye el uso de un algoritmo genético (GA), que es un método más lento pero más exhaustivo, o el uso de proyección de importancia de variable (VIP), que es una determinación rápida y temprana que usa una clasificación predictiva.

Los datos de CGH se generaron esencialmente como se ha descrito en el Ejemplo 1, mediante hibridación contra sistemas de CGH de maíz en formato 2X105K. Las muestras para CGH fueron tomadas de 14 experimentos R2 que contenían plantas procedentes de 3 grupos de madurez relativa, que representan 181 genotipos (91 cultivos stiff stalk, 90 no stiff stalk) que produjeron 914 híbridos. Los datos fueron analizados para identificar oligonucleótidos asociados a heterosis usando datos fenotípicos que incluyen el rendimiento, la altura de espiga, la humedad, el peso de ensayo, el verdor, la altura de la planta, la capacidad de almacenamiento y las raíces. Los datos del análisis

fueron cruzados y validados con datos de mapeado cuando se disponía de ellos. Para el cultivo stiff stalk A frente a 36 cultivos stiff stalk, se identificaron conjuntos de oligómeros predictivos putativos para el rendimiento, la altura de espiga, la humedad y la altura de planta.

En este experimento, se adoptaron cambios en la estrategia para incluir un método adicional más rápido de selección de variables. Las intensidades de CGH fueron incluidas en la regresión multivariable si no se producía variación en el conjunto de datos de hibridación de referencia, la intensidad relativa de cada uno de los oligómeros para cada cultivo era inferior a diez para todos los oligómeros pero mayor de 2 para al menos un cuarto de los cultivos. Para el conjunto de ensayo "Cultivo A", 34.541 de los 103.250 oligómeros disponibles se ajustaron a estos criterios de selección de pre-procesado. Se construyó un modelo de regresión PLS para cada uno de los rasgos fenotípicos, rendimiento, altura de espiga, altura de planta y humedad, usando una variable latente. La importancia de variable en la proyección (puntuación VIP) se calculó entonces y se usó para seleccionar oligómeros para un modelo adicional. El umbral VIP para la inclusión en el modelo se fijó al menos superior a 1 y como mucho 10. Entonces se construyó un segundo modelo con el número reducido de variables y se llevó a cabo una segunda selección VIP con dichas variables usando criterios similares al primero. Después de la segunda iteración de selección de variables se llevó a cabo la validación cruzada deja-uno-fuera para estimar el error de predicción de cada cultivo. Los rasgos predichos se comparan con los rasgos medidos en las Figuras 6-9. Los datos del análisis quimiométrico correspondientes al Cultivo A frente a 36 cultivos no stiff stalk se resumen en la Tabla 1 mostrada a continuación. Dentro de estos conjuntos de oligómeros predictivos, se encontraron 2 oligómeros comunes en los conjuntos de predicción de rendimiento y de altura de planta, lo que indica que algunos rasgos pueden ser correlacionados. Una regresión de la altura de la planta frente a los datos de rendimiento dio un valor de R^2 de 0,310.

TABLA1

Predicción de rasgo	**Nº de oligómeros**	**R^2**	**Validación**
Rendimiento	18	0,7811	4 oligómeros mapeados a la región asociada con el rendimiento
Altura de espiga	8	0,5838	5 oligómeros mapeados a la región asociada a la altura de espiga
Humedad	18	0,6991	8 oligómeros mapeados a la región asociada a la humedad
Altura de planta	32	0,6362	11 oligómeros mapeados a la región asociada a la altura de planta

Ejemplo 6: Secuenciamiento completo del genoma

Otros métodos que pueden usarse para la detección de variantes estructurales genómicas, tal como las variaciones del número de copia, las inserciones, las eliminaciones y los polimorfismos de nucleótido (SNPs), incluyen métodos para el secuenciamiento de ADN comparativo directo de genomas. El secuenciamiento comparativo directo puede llevarse a cabo de varias maneras conocidas por los especialistas en la técnica, que incluyen las estrategias mostradas a continuación, aunque sin limitarse a ellas.

Por ejemplo, el secuenciamiento de escopeta de genoma completo y el montaje usando secuenciamiento de didesoxinucleótidos fluorescentes se pueden usar para detectar y caracterizar diferencias estructurales. Los genomas de las líneas vegetales individuales que difieren en sus genotipos, determinado mediante análisis de marcador genético o análisis de pedigrí, son secuenciadas y a continuación se comparan unas con otras usando herramientas de software bioinformático disponibles. Cualquier diferencia se cataloga mediante el tipo y la localización genómica, y sus números en cada categoría se presentan al análisis, por ejemplo como se ha descrito en los Ejemplos 2 y/o 5.

El secuenciamiento de escopeta de genoma completo que usa tecnologías de ultra-alto rendimiento, tales como el sistema proporcionado por Illumina, Inc. (www.illumina.com), se puede usar para producir una pluralidad de secuencias de genomas de líneas vegetales individuales. Las lecturas de secuenciamiento producidas mediante esta estrategia se ensamblan y se analizan como se ha indicado antes. Opcionalmente o adicionalmente, se prepara el catálogo de fragmentos de secuencia obtenidos, o de sub-secuencias dentro de ellos (k-meros), y se pueden comparar los dos catálogos de dos individuos diferentes. Las diferencias en el número de fragmentos en cada categoría se anotan y se realiza el análisis estadístico para estimar los intervalos de confianza para dichas diferencias abundantes. El catálogo de las diferencias que cumplen los criterios estadísticos de confianza se envía al análisis descrito en los Ejemplos 2 y 5, o a métodos equivalentes de la técnica.

Alternativamente, se pueden secuenciar subconjuntos de cada genoma. Por ejemplo, un subconjunto puede ser cromosomas individuales obtenidos mediante clasificación de cromosomas, segmentos de genoma seleccionados

mediante hibridación y posterior elución desde microsistemas, o un subconjunto generado por cualquier otro método conocido por los especialistas en la técnica. El catálogo de diferencias para los subconjuntos de cada genoma que cumplen los criterios estadísticos de confianza de genoma se envía a analizar tal como se ha descrito en los Ejemplos 2 y/o 5, o mediante otros métodos equivalentes. En algunos ejemplos, también se pueden usar métodos alternativos de secuenciamiento de genoma completo o parcial, siempre que los métodos puedan producir un catálogo de diferencias en las secuencias de los genomas que se están comparando.

En un ejemplo, la secuencia completa directa de genoma implica las siguientes etapas:

1) aislar ADN genómico;

2) preparar ADN genómico para el secuenciamiento, opcionalmente marcar la(s) secuencia(s);

3) secuenciar el ADN genómico de la etapa 1 (el método de secuenciamiento puede marcar polinucleótidos);

4) mapear secuencias al genoma y contabilizar la aparición de etiquetas o marcas;

5) tras normalizar los datos, comparar las etiquetas entre muestras para determinar CNVs;

6) aplicar métodos de análisis de datos (por ejemplo, Ejemplo 2 y/o Ejemplo 5) para relacionar las CNVs observadas con al menos un fenotipo heterótico.

Opcionalmente, el ADN genómico aislado procedente de la etapa 1 o de la etapa 2 podría ser procesado para eliminar secuencias repetitivas, o para reducir la complejidad de la muestra antes del secuenciamiento. Por ejemplo, se podrían sintetizar oligómeros de las regiones repetitivas y marcarse con una molécula de biotina. Los oligómeros biotinilados se añaden al ADN, y la muestra se aplica a una columna se estreptavidina. La muestra que atraviesa la columna de ADN no repetitivo se recoge para un análisis posterior. En otro ejemplo, se crea un microsistema dirigido a las regiones repetitivas. La muestra de ADN se hibrida al sistema de tal modo que los fragmentos no ligados se recogen y se usan para el secuenciamiento. En otro método, el ADN genómico podría ser digerido usando una enzima de restricción, y a continuación iniciar el secuenciamiento a partir del sitio de RE.

Todas las publicaciones y las solicitudes de patente mencionadas en la especificación son indicativas del nivel de los especialistas en la técnica a los que va dirigida esta invención. Aunque la invención precedente ha sido descrita con algo de detalle a modo de ilustración y ejemplo con fines de claridad de entendimiento, será obvio que se pueden realizar determinados cambios y modificaciones dentro del alcance de las reivindicaciones anexas. Tal como se usa en la presente memoria y en las reivindicaciones anexas, las formas singulares “uno”, “una” y “el”, “la” incluyen referencias plurales a menos que el contexto dicte claramente lo contrario. Por lo tanto, por ejemplo, la referencia a “una planta” incluye una pluralidad de plantas; la referencia a “una célula” incluye una o más células y equivalentes de las mismas conocidas por los especialistas en la técnica, etc.

REIVINDICACIONES

1.- Un método que comprende:

(a) seleccionar una primera planta y una segunda planta, en donde la primera planta y la segunda planta pueden cruzarse para producir una planta progenie fértil que presente un fenotipo relacionado con heterosis en comparación con las plantas parentales;

(b) detectar variaciones estructurales de ADN entre un genoma de la primera planta y un genoma de la segunda planta; y,

(c) relacionar las variaciones estructurales con el fenotipo relacionado a heterosis usando una estrategia computacional evolucionaria iterada,

identificando con ello las variaciones estructurales que predicen el grado esperado del fenotipo relacionado a heterosis en la planta progenie;

en donde, en la etapa (b),

(i) las variaciones estructurales son detectadas usando un método de hibridación genómica comparativo; o

(ii) las variaciones estructurales son variaciones del número de copia.

2.- El método de la reivindicación 1, en el que la planta progenie es maíz.

3.- El método de la reivindicación 1, en el que el método de hibridación genómica comparativa comprende:

(a) poner en contacto ADN genómico de una primera planta con una primera pluralidad de moléculas sonda de oligonucleótido;

(b) detectar las intensidades de hibridación correspondientes a al menos un subconjunto de moléculas sonda de oligonucleótido de la primera pluralidad de moléculas sonda de oligonucleótido;

(c) poner en contacto ADN genómico procedente de una segunda planta con una segunda pluralidad de moléculas sonda de oligonucleótido, en donde dichas primera y segunda pluralidades de moléculas sonda de oligonucleótido tienen al menos un subconjunto de moléculas sonda de oligonucleótido en común;

(d) detectar las intensidades de hibridación correspondientes a al menos un subconjunto de moléculas sonda de oligonucleótidos en la segunda pluralidad de moléculas sonda de oligonucleótido;

(e) determinar medidas relativas de intensidad de hibridación correspondientes a una pluralidad de moléculas sonda de oligonucleótido individuales en dicho subconjunto común de moléculas sonda de oligonucleótido;

(f) usar dichas intensidades de hibridación relativas para predecir el grado de un fenotipo relacionado a heterosis correspondiente a una planta progenie derivada de dichas primera y segunda plantas.

4.- El método de la reivindicación 3, en el que:

(a) al menos una de dichas primera y segunda plantas comprende una variedad de planta cultivada; o

(b) la pluralidad de moléculas sonda de oligonucleótido comprende un sistema de oligonucleótidos; o

(c) dicho método comprende además seleccionar dichas primera y segunda plantas para el desarrollo de una variedad de planta híbrida F1 en base al menos en parte a dicha predicción de un fenotipo relacionado a heterosis; o

(d) las intensidades de hibridación relativas comprenden una medida de las variaciones de número de copia entre dicha primera planta y dicha segunda planta.

5.- Un método para desarrollar un sistema de oligonucleótidos para la predicción de un fenotipo relacionado a heterosis en una planta que comprende:

(a) seleccionar una pluralidad de líneas parentales en las que se ha cuantificado el fenotipo relacionado a heterosis en una pluralidad de los cruces F1 de dichas líneas parentales;

(b) poner en contacto ADN genómico de cada una de la pluralidad de dichas líneas parentales con una pluralidad de moléculas sonda de oligonucleótido, en donde dichas pluralidades de moléculas sonda de oligonucleótido tienen al menos un subconjunto de moléculas sonda de oligonucleótido en común;

(c) detectar las intensidades de hibridación correspondientes a moléculas sonda de oligonucleótido individuales

en las pluralidades de moléculas sonda de oligonucleótido;

(d) determinar medidas relativas de intensidad de hibridación para una pluralidad de las moléculas sonda de oligonucleótido individuales en dicho subconjunto de moléculas sonda de oligonucleótido;

(e) seleccionar moléculas sonda de oligonucleótido que muestren intensidades de hibridación diferentes entre dichas líneas parentales;

(f) relacionar dichas intensidades de hibridación con dichas moléculas sonda de oligonucleótido seleccionadas con un fenotipo relacionado a heterosis de las plantas progenie; y

(g) crear un sistema de oligonucleótidos especializado para la predicción de un fenotipo relacionado a heterosis que comprende dichas moléculas sonda de oligonucleótido seleccionadas que se relacionan con un fenotipo relacionado a heterosis.

6.- El método de una cualquiera de las reivindicaciones 3-5, en el que la planta es maíz.

7.- El método de una cualquiera de las reivindicaciones 3-6, en el que el fenotipo relacionado a heterosis es el rendimiento.

8.- El método de una cualquiera de las reivindicaciones 3-7, en el que el ADN genómico comprende ADN genómico preparado.

9.- El método de una cualquiera de las reivindicaciones 3-8, en el que dichas pluralidades de moléculas sonda de oligonucleótido comprenden al menos un 50% de moléculas sonda de oligonucleótido que se hibridan con regiones codificadoras o con otras secuencias de ADN genómico no repetitivas.

10.-El método de una cualquiera de las reivindicaciones 3-9, en el que dicho subconjunto de moléculas sonda de oligonucleótido común comprende al menos 100 moléculas sonda de oligonucleótido.

11.-El método de una cualquiera de las reivindicaciones 3-10, en el que dicho subconjunto de moléculas sonda de oligonucleótido no contiene más de 150 moléculas sonda de oligonucleótido.

12.-El método de una cualquiera de las reivindicaciones 3-11, en el que dichas moléculas sonda de oligonucleótido tiene una longitud de al menos 20 nucleótidos, pero no más de 100.

13.-El método de una cualquiera de las reivindicaciones 3-12, en el que dicha intensidad de hibridación relativa comprende una relación de intensidad de hibridación.

14.-El método de una cualquiera de las reivindicaciones 3-13, en el que se seleccionan las moléculas sonda de oligonucleótido que exhiben al menos una diferencia de tres veces, pero menos de una diferencia de diez veces, en la intensidad de hibridación.

15.-El método de la reivindicación 5, en el que la etapa de selección de moléculas sonda de oligonucleótido comprende además una estrategia computacional evolucionaria iterada que comprende:

(a) formar subconjuntos de moléculas sonda de oligonucleótidos seleccionados aleatoriamente a partir de dicho subconjunto común de moléculas sonda de oligonucleótidos;

(b) determinar la capacidad de un subconjunto para predecir dicho fenotipo relacionado a heterosis en base a las intensidades relativas de los subconjuntos de oligonucleótidos;

(c) seleccionar los subconjuntos que se determine que son mejores predictores de dicho fenotipo relacionado a heterosis;

(d) formar nuevos subconjuntos combinando segmentos de los subconjuntos de intensidad predictivos mediante la adición aleatoria de nuevos oligonucleótidos procedentes del conjunto común de sondas;

(e) repetir las etapas (b) a (d) hasta que sólo se produzcan incrementos pequeños de la capacidad predictiva de los subconjuntos o de la convergencia en la población de subconjunto predictivo;

preferiblemente en donde la capacidad de un subconjunto para predecir se analiza a través de análisis de regresión, o mediante un método de aprendizaje mecánico.

16.-El método de una cualquiera de las reivindicaciones 5-15, en el que dicho sistema de oligonucleótidos no contiene más de 150 moléculas sonda de oligonucleótido.

FIGURA 1

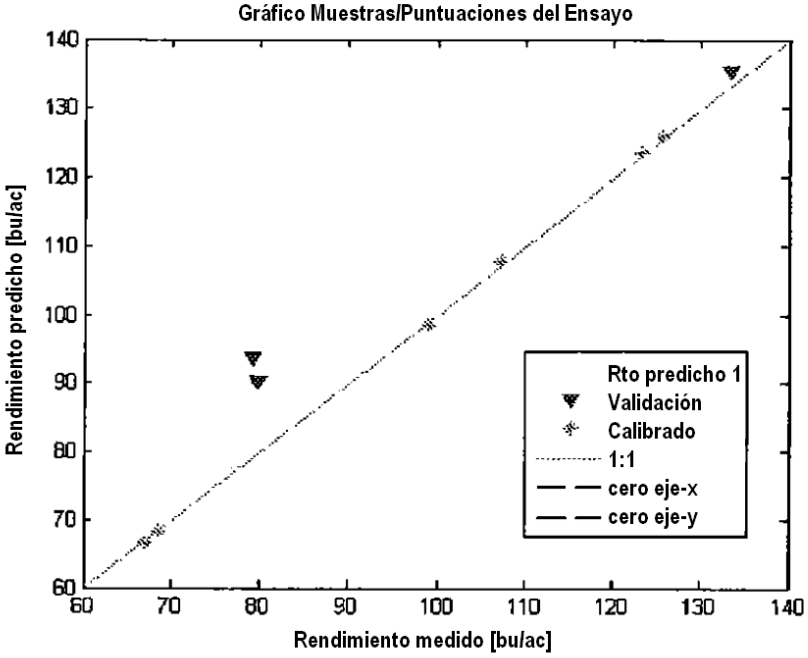

FIGURA 2

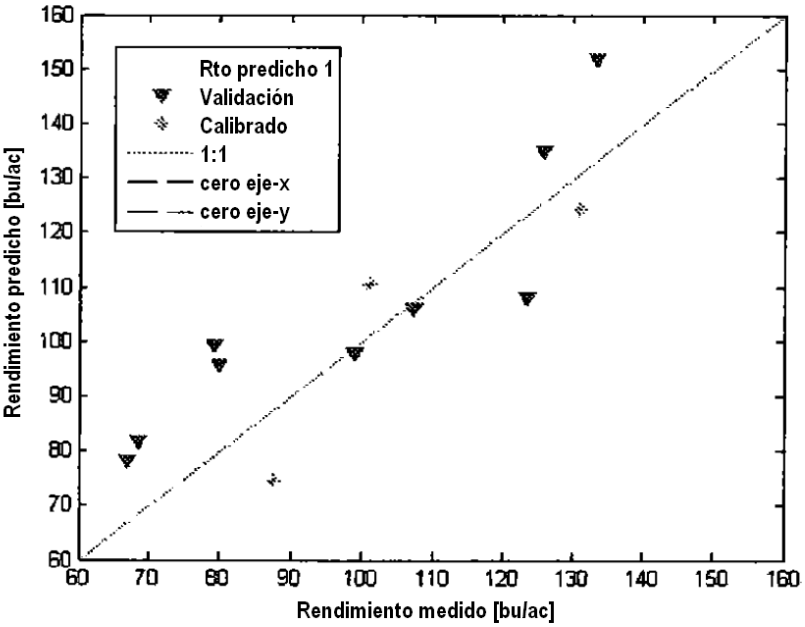

FIGURA 3

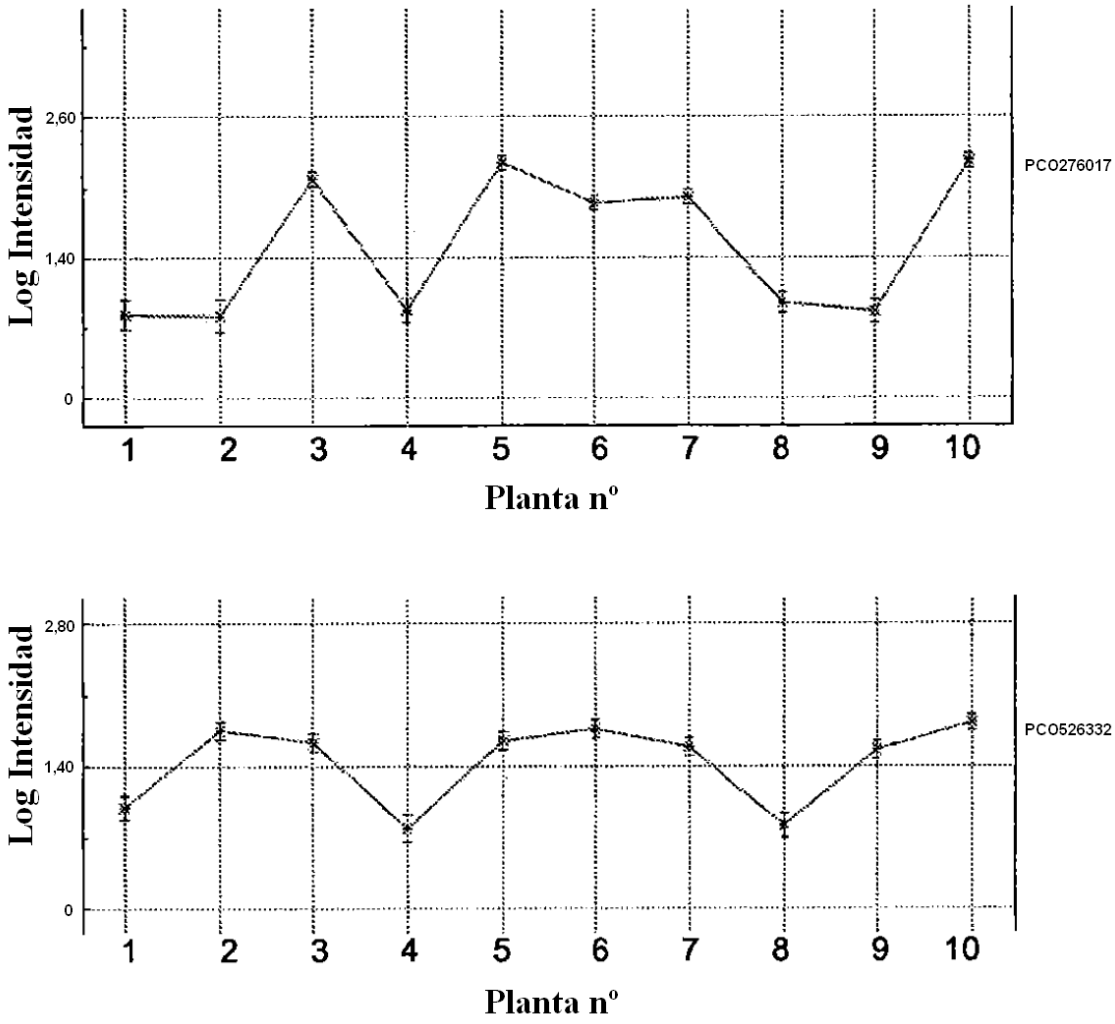
Log Intensidad
2,60
1,40
0
1
2
3
4
5
6
7
8
9
10
Planta nº
PCO276017
Log Intensidad
2,80
1,40
0
1
2
3
4
5
6
7
8
9
10
Planta nº
PCO526332

FIGURA 4

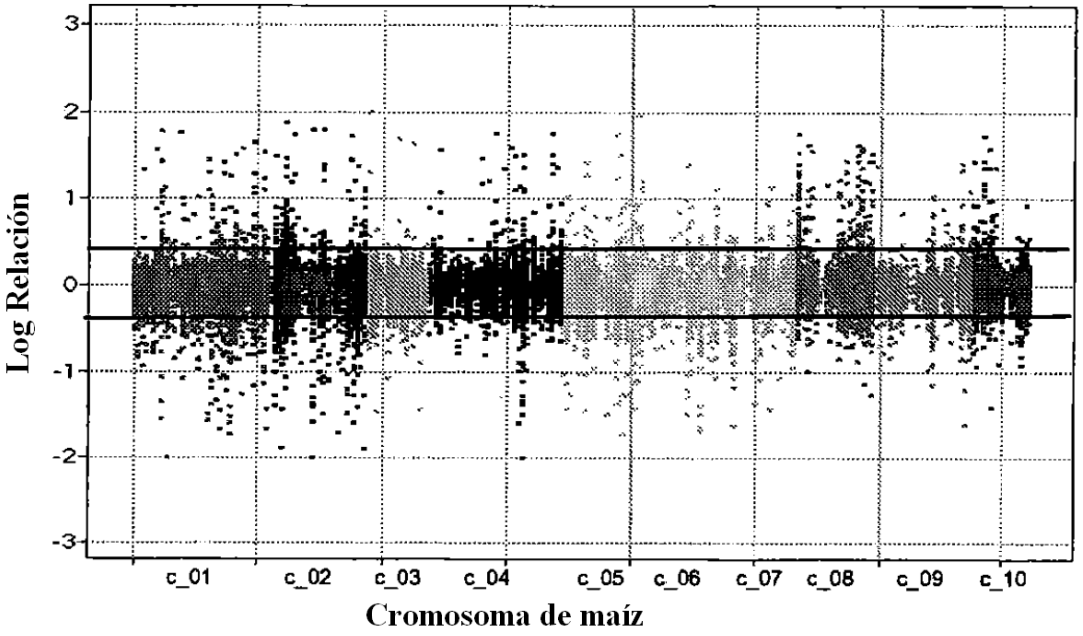

FIGURA 5

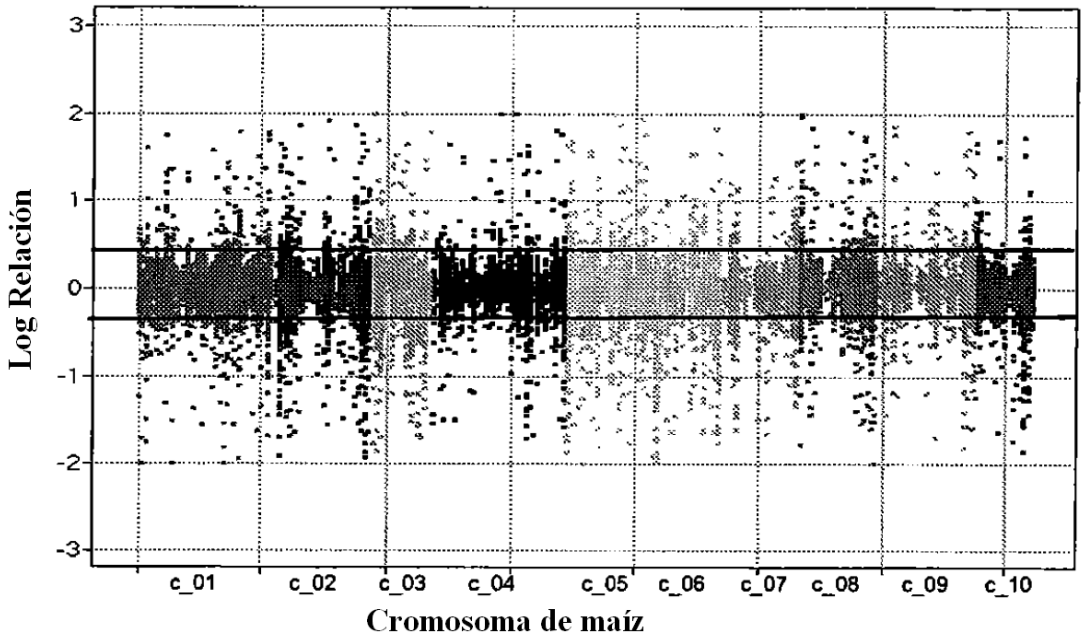

FIGURA 6

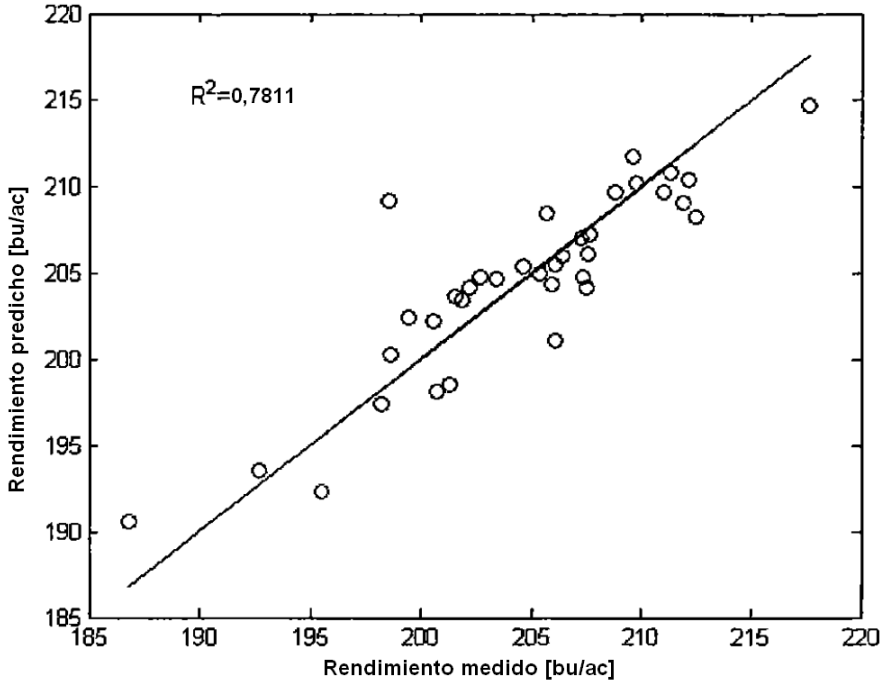

FIGURA 7

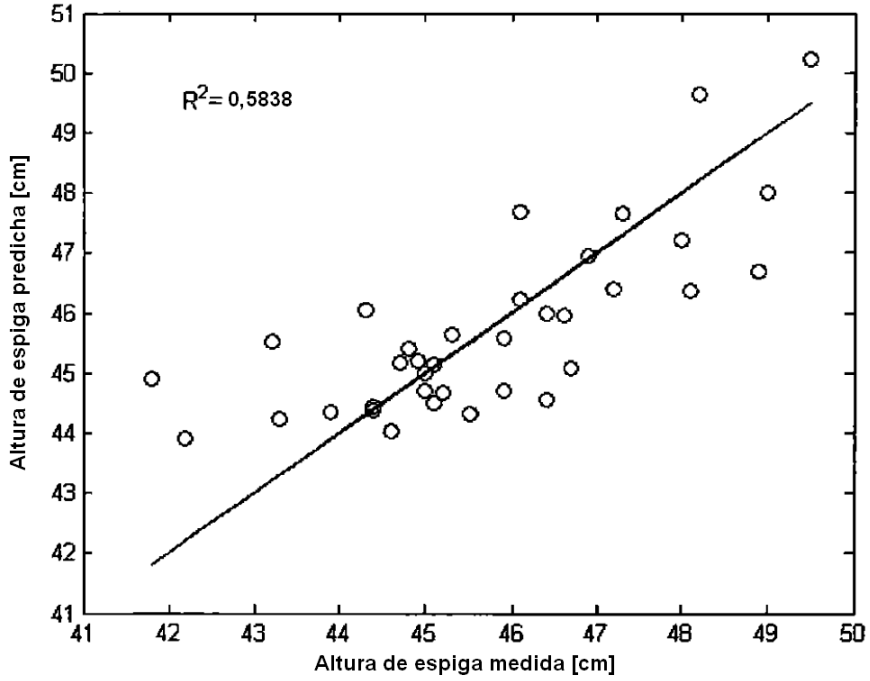

FIGURA 8

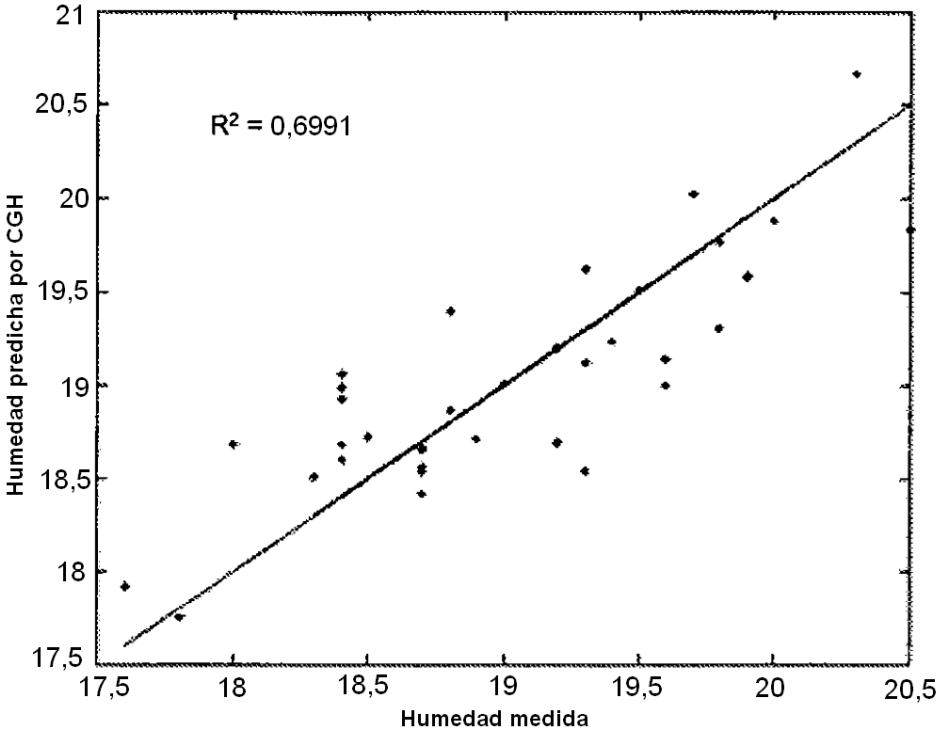

FIGURA 9