(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 357 549**

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

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **06757808 .8**
(96) Fecha de presentación : **23.06.2006**
(97) Número de publicación de la solicitud: **1910562**
(97) Fecha de publicación de la solicitud: **16.04.2008**

(54) Título: **Estrategias para la identificación y detección de alto rendimiento de polimorfismos.**

(30) Prioridad: **23.06.2005 US 693053 P**
16.01.2006 EP 06075104
17.01.2006 US 759034 P

(45) Fecha de publicación de la mención BOPI:
27.04.2011

(45) Fecha de la publicación del folleto de la patente:
27.04.2011

(73) Titular/es: **KEYGENE N.V.**
Agro Business Park 90
6708 PW Wageningen, NL

(72) Inventor/es:
Van Eijk, Michael, Josephus, Theresia y
Van der Poel, Henricus, Johannes, Adam

(74) Agente: **Sugrañes Moliné, Pedro**

DESCRIPCIÓN

Campo técnico

[0001] La presente invención se refiere a los campos de la biología molecular y de la genética. La invención se refiere a la rápida identificación de múltiples polimorfismos en una muestra de ácidos nucleicos. Los polimorfismos identificados pueden utilizarse para el desarrollo de sistemas de cribado de alto rendimiento para polimorfismos en muestras de ensayo.

Antecedentes de la invención

[0002] La exploración del ADN genómico ha sido el deseo de la comunidad científica, en particular de la comunidad médica, desde hace mucho tiempo. El ADN genómico es la clave para la identificación, diagnóstico y tratamiento de enfermedades tales como el cáncer y la enfermedad de Alzheimer. Además de la identificación y tratamiento de enfermedades, la exploración del ADN genómico podría proporcionar ventajas significativas en esfuerzos de cría vegetal y animal, proporcionando respuestas a problemas de alimentación y nutrición en todo el mundo. Es conocido que muchas enfermedades se asocian a componentes genéticos específicos, en particular a polimorfismos en genes específicos. La identificación de polimorfismos en muestras grandes, tales como genomas, en la actualidad es una tarea laboriosa y que requiere mucho tiempo. Sin embargo, esta identificación resulta de gran valor para áreas tales como la investigación biomédica, el desarrollo de productos farmacéuticos, el tipado de tejidos, el genotipado y los estudios poblacionales.

[0003] La patente WO 2004/022758 describe un método para la fragmentación del genoma que utiliza una pluralidad de enzimas de restricción para producir una fragmentación representativa de un ácido nucleico de muestra ligado en un vector con el fin de proporcionar una biblioteca que contenga los fragmentos.

[0004] Nicod *et al.*, Nucleic acids research 31(5):19, 1 de marzo de 2003, describe un método para identificar SNPs basándose en la AFLP utilizando geles radioactivos, la extracción uno a uno de los fragmentos de la AFLP y la evitación de la selección de fragmentos que se encuentran presentes en todos los individuos investigados.

Descripción resumida de la invención

[0005] La presente invención proporciona un método para identificar eficientemente y para detectar fiablemente polimorfismos en una muestra de ácidos nucleicos (por ejemplo ADN o ARN) compleja, por ejemplo muy grande, de un modo rápido y económico utilizando una combinación de métodos de alto rendimiento.

[0006] Dicha integración de métodos de alto rendimiento conjuntamente proporcionan una plataforma que resulta particularmente adecuada para la identificación y detección rápidas y fiables de polimorfismos en muestras de ácidos nucleicos altamente complejas, en las que la identificación y mapeado convencionales de polimorfismos resultaría laboriosa y que requeriría mucho tiempo.

[0007] Una de la cosas que han encontrado los presentes inventores es una solución para identificar polimorfismos, preferentemente polimorfismos de un solo nucleótido, aunque de manera similar para (micro)satélites y/o indels, en particular en genomas grandes. El método es único en su aplicabilidad a genomas tanto grandes como pequeños, aunque proporciona ventajas particulares en genomas grandes, en particular en especies poliploides.

[0008] Para identificar los SNPs (y posteriormente detectar los SNPs identificados) se dispone de varias posibilidades en la técnica. En una primera opción, puede secuenciarse el genoma completo, y ello puede llevarse a cabo en varios individuos. Es un ejercicio en gran parte teórico, al ser engorroso y caro y, a pesar del rápido desarrollo de la tecnología, simplemente no resulta factible su aplicación a cada organismo, especialmente a los que presentan genomas más grandes. La segunda opción es utilizar la información de secuencia disponible (fragmentada), tal como las bibliotecas EST. Esto permite la generación de cebadores de PCR, la resecuenciación y la comparación entre individuos. Nuevamente lo anterior requiere información inicial de secuencia que no se encuentra disponible o que se encuentra disponible sólo en una cantidad limitada. Además, deben desarrollarse ensayos de PCR separados para cada región, lo que supone una adición enorme de costes y tiempo de desarrollo.

[0009] La tercera opción es limitarse a parte del genoma de cada individuo. La dificultad reside en que la parte proporcionada del genoma debe ser igual en diferentes individuos con el fin de proporcionar un resultado comparable para la identificación con éxito de SNPs. Los presentes inventores ahora han resuelto este dilema mediante la integración de métodos altamente reproducibles para seleccionar parte del genoma mediante secuenciación de alto rendimiento para la identificación de los polimorfismos integrada con la preparación de muestras y plataformas de identificación de alto rendimiento. La presente invención acelera el procedimiento de identificación de polimorfismos y utiliza los mismos elementos en el procedimiento posterior para explotar los polimorfismos descubiertos, permitiendo un genotipado de alto rendimiento efectivo y fiable.

[0010] Entre las aplicaciones adicionalmente contempladas por la presente invención se incluyen las bibliotecas de microsatélites enriquecidas mediante cribado, la realización de AFLP-ADNc de perfilado de transcritos (northern digital), la secuenciación de genomas complejos, la secuenciación de bibliotecas de EST (en

ADNc completo o en AFLP-ADNc), la exploración de microARN (secuenciación de bibliotecas de inserciones de pequeño tamaño), la secuenciación de cromosoma artificial bacteriano (BAC) (contig), AFLP/AFLP-ADNc en un enfoque de análisis de segregantes agrupados, la detección rutinaria de fragmentos de la AFLP, por ejemplo para retrocruzamientos asistidos por un marcador (MABC), etc.

Definiciones

[0011] En la descripción y ejemplos posteriormente se utiliza una serie de expresiones. Con el fin de proporcionar una comprensión clara y consistente de la memoria y reivindicaciones, incluyendo el alcance que debe proporcionarse a dichas expresiones, se proporcionan las definiciones siguientes. A menos que se defina de otra manera en la presente memoria, todas las expresiones técnicas y científicas utilizadas presentan los mismos significados comúnmente entendidos por el experto ordinario en la materia a la que pertenece la presente invención.

[0012] Polimorfismo: los polimorfismos se refieren a la presencia de dos o más variantes de una secuencia de nucleótidos en una población. Un polimorfismo puede comprender uno o más cambios de bases, una inserción, una repetición, o una deleción. Un polimorfismo incluye, por ejemplo una repetición de secuencia simple (SSR) y un polimorfismo de un único nucleótidos (SNP), que es una variación que se produce en el caso de que se altere un único nucleótido: adenina (A), timina (T), citosina (C) o guanina (G). Debe producirse generalmente una variación en por lo menos 1% de la población para que se considere un SNP. Los SNPs constituyen 90% de todas las variaciones genéticas humanas y se producen cada 100 a 300 bases a lo largo del genoma humano. Dos de cada tres SNPs sustituyen la citosina (C) por la timina (T). Las variaciones en las secuencias de ADN de, por ejemplo, seres humanos o plantas, pueden afectar a cómo se enfrentan a enfermedades, bacterias, virus, compuestos químicos, fármacos, etc.

[0013] Ácido nucleico: un ácido nucleico según la presente invención puede incluir cualquier polímero u oligómero de bases pirimidina o purina, preferentemente citosina, timina, y uracilo,y adenina y guanina, respectivamente (ver Albert L. Lehninger, Principles of Biochemistry, páginas 793 a 800, Worth Publ, 1982). La presente invención contempla cualquier componente desoxirribonucleótido, ribonucleótido o péptido-ácido nucleico, y cualesquiera variantes químicas de los mismos, tales como formas metiladas, hidroximetiladas o glucosiladas de dichas bases, y similares. Los polímeros u oligómeros pueden ser de composición heterogénea u homogénea, y pueden aislarse a partir de fuentes naturales o producirse artificial o sintéticamente. Además, los ácidos nucleicos pueden ser de ADN o ARN, o una mezcla de los mismos, y pueden existir permanente o transitoriamente en forma de una cadena o de doble cadena, incluyendo estados de homodúplex, heterodúplex e híbridos.

[0014] Reducción de complejidad: la expresión "reducción de la complejidad" se utiliza para referirse a un método en el que la complejidad de una muestra de ácidos nucleicos, tal como ADN genómico, se reduce mediante la generación de un subconjunto de la muestra. Este subconjunto puede ser representativo de la muestra completa (es decir de la muestra compleja) y preferentemente es un subconjunto reproducible. El término "reproducible" se refiere en el presente contexto a que, al reducir la complejidad de la misma muestra utilizando el mismo método, se obtiene el mismo subconjunto o por lo menos uno comparable. El método utilizado para la reducción de la complejidad puede ser cualquier método de reducción de la complejidad conocido de la técnica. Entre los ejemplos de métodos de reducción de la complejidad se incluyen, por ejemplo, AFLP® (Keygene N.V., Países Bajos; ver, por ejemplo, la patente EP 0 534 858), los métodos descritos por Dong (ver, por ejemplo, las patentes WO 03/012118 y 00/24939), la unión indexada (Unrau *et al.*, ver posteriormente), etc. Los métodos de reducción de la complejidad utilizados en la presente invención presentan en común que son reproducibles. Se utiliza término reproducible en el sentido de que se reduce la complejidad de la misma muestra del mismo modo, se obtiene el mismo subconjunto de la muestra, y no en el sentido de la reducción de complejidad más aleatoria, tal como la microdisección o la utilización de ARNm (ADNc), que representa una parte del genoma transcrito en un tejido seleccionado y su reproducibilidad depende de la selección de tejido, momento del aislamiento, etc.

[0015] Etiquetado: el término "etiquetado" se refiere a la adición de una etiqueta a una muestra de ácidos nucleicos con el fin de distinguirla de una segunda o posteriores muestras de ácidos nucleicos. El etiquetado puede llevarse a cabo, por ejemplo, mediante la adición de un identificador de secuencia durante la reducción de complejidad o mediante cualquier otro medio conocido de la técnica. Dicho identificador de secuencia puede ser, por ejemplo, una única secuencia de bases de longitud variable aunque definida, utilizada únicamente para identificar una muestra específica de ácidos nucleicos. Son ejemplos típicos de la misma, por ejemplo, las secuencias ZIP. Mediante la utilización de dicha etiqueta, puede determinarse el origen de una muestra tras el procesamiento adicional. En el caso de que se combinen productos procesados que se originan de diferentes muestras de ácidos nucleicos, deben identificarse las muestras de ácidos nucleicos diferentes utilizando etiquetas diferentes.

[0016] Biblioteca etiquetada: el término "etiquetada" se refiere a una biblioteca de ácidos nucleicos etiquetados.

[0017] Secuenciación: el término "secuenciación" se refiere a determinar el orden de los nucleótidos (secuencias de bases) en una muestra de ácidos nucleicos, por ejemplo ADN o ARN.

[0018] Alinear y alineación: el término "alinear" y "alineación" se refiere a la comparación entre dos o más secuencias de nucleótidos basada en la presencia de segmentos cortos o largos de nucleótidos idénticos o similares. Son conocidos de la técnica varios métodos para alinear secuencias de nucleótidos, tal como se explica adicionalmente después.

[0019] Sondas de detección: la expresión "sondas de detección" se utiliza para referirse a sondas diseñadas para detectar una secuencia específica de nucleótidos, en particular secuencias que contienen uno o más polimorfismos.

[0020] Cribado de alto rendimiento: el cribado de alto rendimiento, con frecuencia abreviado HTS, es un método para la experimentación científica especialmente relevante para los campos de la biología y la química. Mediante una combinación de robótica moderna y otros equipos de laboratorio especializados, permite al investigador cribar eficazmente grandes cantidades de muestras simultáneamente.

[0021] Ácidos nucleicos de muestra de ensayo: la expresión "ácidos nucleicos de muestra de ensayo" se utiliza para indicar una muestra de ácidos nucleicos que se investiga para polimorfismos utilizando el método de la presente invención.

[0022] Endonucleasa de restricción: una endonucleasa de restricción o enzima de restricción es un enzima que reconoce una secuencia específica de nucleótidos (sitio diana) en una molécula de ADN de doble cadena, y corta ambas cadenas de la molécula de ADN en todos los sitios diana.

[0023] Fragmentos de restricción: las moléculas de ADN producidas mediante digestión con una endonucleasa de restricción se denominan fragmentos de restricción. Se digiere cualquier genoma dado (o ácido nucleico, con independencia de su origen) mediante una endonucleasa de restricción particular en un conjunto discreto de fragmentos de restricción. Los fragmentos de ADN que resultan del corte con endonucleasa de restricción pueden utilizarse adicionalmente en una diversidad de técnicas y pueden detectarse mediante, por ejemplo, electroforesis en gel.

[0024] Electroforesis en gel: con el fin de detectar fragmentos de restricción, puede resultar necesario un método analítico para fraccionar moléculas de ADN de doble cadena basándose en el tamaño. La técnica utilizada más comúnmente para conseguir dicho fraccionamiento es la electroforesis (capilar) en gel. La tasa a la que se desplazan los fragmentos de ADN en dichos geles depende de su peso molecular; de esta manera, se reducen las distancias recorridas a medida que se incrementa la longitud del fragmento. Los fragmentos de ADN fraccionados mediante electroforesis en gel pueden visualizarse directamente mediante un procedimiento de tinción, por ejemplo tinción con plata o tinción utilizando bromuro de etidio, en el caso de que el número de fragmentos incluido en el patrón sea suficientemente reducido. Alternativamente, el tratamiento adicional de los fragmentos de ADN puede incorporar marcajes detectables en los fragmentos, tales como fluoróforos o marcajes radioactivos.

[0025] Ligación: la reacción enzimática catalizada por un enzima ligasa en el que se unen covalentemente entre sí dos moléculas de ADN de doble cadena se denomina ligación. En general, ambas cadenas de ADN se unen covalentemente entre sí, aunque también resulta posible evitar la ligación de una de las dos cadenas mediante modificación química o enzimática de uno de los extremos de las cadenas. En este caso, se producirá la unión covalente en únicamente una de las dos cadenas de ADN.

[0026] Oligonucleótido sintético: las moléculas de ADN de una cadena que presentan preferentemente entre aproximadamente 10 y aproximadamente 50 bases, que pueden sintetizarse químicamente se denominan oligonucleótidos sintéticos. En general, estas moléculas de ADN sintético se diseñan para que presenten una secuencia de nucleótidos única o deseada, aunque resulta posible sintetizar familias de moléculas que presenten secuencias relacionadas y que presenten composiciones de nucleótidos diferentes en posiciones específicas dentro de la secuencia de nucleótidos. La expresión oligonucleótido sintético se utiliza para referirse a moléculas de ADN que presentan una secuencia de nucleótidos diseñada o deseada.

[0027] Adaptadores: moléculas cortas de ADN de doble cadena con un número limitado de pares de bases, por ejemplo una longitud de entre aproximadamente 10 y aproximadamente 30 pares de bases, que se diseñan de manera que puedan ligarse a los extremos de fragmentos de restricción. Los adaptadores están compuestos generalmente de dos oligonucleótidos sintéticos que presentan secuencias de nucleótidos que son parcialmente complementarias entre sí. Al mezclar los dos oligonucleótidos sintéticos en solución bajo condiciones apropiadas, se aparean entre sí formando una estructura de doble cadena. Tras la hibridación, un extremo de la molécula adaptadora se diseña de manera que sea compatible con el extremo de un fragmento de restricción y pueda ligarse al mismo; el otro extremo del adaptador puede diseñarse de manera que no pueda ligarse, aunque éste no es necesariamente el caso (adaptadores doblemente ligados).

[0028] Fragmentos de restricción ligados a adaptador: fragmentos de restricción a los que se han añadido caperuzas de adaptadores.

[0029] Cebadores: en general, el término cebadores se refiere a una cadena de ADN que puede cebar la síntesis del ADN. La ADN polimerasa no puede sintetizar ADN *de novo* sin cebadores: únicamente puede

extender una cadena de ADN existente en una reacción en la que la cadena complementaria se utiliza como molde para dirigir el orden de nucleótidos que deben ensamblarse. Se hace referencia a las moléculas oligonucleótidas sintéticas que se utilizan en una reacción en cadena de la polimerasa (PCR) como cebadores.

[0030] Amplificación de ADN: el término amplificación de ADN típicamente se utiliza para referirse a la síntesis *in vitro* de moléculas de ADN de doble cadena utilizando la PCR. Se indica que existen otros métodos de amplificación y que pueden utilizarse en la presente invención sin apartarse de la esencia de la misma.

Descripción detallada de la invención

[0031] La presente invención proporciona un método para identificar uno o más polimorfismos, comprendiendo dicho método las etapas de:

a) proporcionar una primera muestra de ácidos nucleicos de interés;

b) llevar a cabo una reducción de complejidad de la primera muestra de ácidos nucleico de interés, proporcionando una primera biblioteca de la primera muestra de ácidos nucleicos;

c) llevar acabo consecutiva o simultáneamente las etapas a) y b) con una segunda o posteriores muestras de ácidos nucleicos de interés, obteniendo una segunda o posteriores bibliotecas de la segunda o posteriores muestras de ácidos nucleicos de interés;

d) secuenciar por lo menos una parte de la primera biblioteca y de la segunda o posteriores bibliotecas, en la que la secuenciación se lleva a cabo en un soporte sólido, tal como una perla;

e) alinear las secuencias obtenidas en la etapa d);

f) determinar uno o más polimorfismos entre la primera muestra de ácidos nucleicos y la segunda o posteriores muestras de ácidos nucleicos en la alineación de la etapa e);

g) utilizar el polimorfismo o polimorfismos determinados en la etapa f) para diseñar sondas de detección;

h) proporcionar una muestra de ensayo de ácido nucleico de interés;

i) llevar a cabo la reducción de complejidad de la etapa b) en la muestra de ensayo de ácido nucleico de interés, proporcionando una biblioteca de ensayo de la muestra de ensayo de ácidos nucleicos;

j) someter la biblioteca de ensayo a cribado de alto rendimiento para identificar la presencia, ausencia o cantidad de los polimorfismos determinados en la etapa f) utilizando las sondas de detección diseñadas en la etapa g);

y en la que la reducción de complejidad de la etapa (b) se lleva a cabo mediante:

- digestión de la muestra de ácidos nucleicos con por lo menos una endonucleasa de restricción para fragmentarla en fragmentos de restricción;
- ligación de los fragmentos de restricción obtenidos con por lo menos un adaptador oligonucleótido sintético de doble cadena que presenta un extremo compatible con uno o ambos extremos de los fragmentos de restricción para producir fragmentos de restricción ligados con adaptadores;
- poner en contacto dichos fragmentos de restricción ligados con adaptadores con uno o más cebadores oligonucleótidos en condiciones de hibridación; y
- amplificar dichos fragmentos de restricción ligados con adaptadores mediante alargamiento de uno o más cebadores oligonucleótidos,
- en el que por lo menos uno o más de los cebadores oligonucleótidos incluye una secuencia de nucleótidos que presenta la misma secuenciad de nucleótidos que las partes terminales de las cadenas en los extremos de dichos fragmentos de restricción ligados con adaptadores, incluyendo los nucleótidos implicados en la formación de la secuencia diana de dicha endonucleasa de restricción e incluyendo por lo menos parte de los nucleótidos presentes en los adaptadores, en el que, opcionalmente, por lo menos uno de dichos cebadores incluye en su extremo 3' una secuencia seleccionada que comprende por lo menos un nucleótido situado inmediatamente contiguo a los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción.

[0032] En la etapa b), se lleva a cabo una reducción de la complejidad de la primera muestra de ácidos nucleicos de interés, proporcionando una primera biblioteca de la primera muestra de ácidos nucleicos.

[0033] En una realización de la invención, la etapa de reducción de la complejidad de la muestra de ácidos nucleicos comprende cortar enzimáticamente la muestra de ácidos nucleicos en fragmentos de restricción, separar los fragmentos de restricción y seleccionar un grupo particular de fragmentos de restricción.

Opcionalmente, los fragmentos seleccionados seguidamente se ligan a secuencias adaptadores que contienen moldes/secuencias ligantes de cebador de PCR.

[0034] En una realización de reducción de complejidad, se utiliza una endonucleasa de tipo II para digerir la muestra de ácidos nucleicos y los fragmentos de restricción se ligan selectivamente a secuencias adaptadoras. Las secuencias adaptadoras pueden contener diversos nucleótidos en el extremo protuberante que debe ligarse y únicamente el adaptador con el conjunto correspondiente de nucleótidos en el extremo protuberante se liga con el fragmento y se amplifica posteriormente. Esta tecnología se describe en la técnica como "línkers de indexación". Pueden encontrarse ejemplos de este principio en, entre otros, Unrau P. y Deugau K.V., Gene 145:163-169, 1994.

[0035] En otra realización, el método de reducción de la complejidad utiliza dos endonucleasas de restricción que presentan diferentes sitios diana y frecuencias y dos secuencias adaptadoras diferentes.

[0036] En otra realización de la invención, la etapa de reducción de la complejidad comprende llevar a cabo una PCR arbitrariamente cebada en la muestra.

[0037] En todavía otra realización de la invención, la etapa de reducción de la complejidad comprende eliminar secuencias repetidas mediante desnaturalización y rehibridación del ADN y posterior eliminación de los dúplex de doble cadena.

[0038] En otra realización de la invención, la etapa de reducción de la complejidad comprende hibridar la muestra de ácidos nucleicos con una perla magnética que se une a una sonda oligonucleótida que contiene una secuencia deseada. Esta realización puede comprender además exponer la muestra hibridada a una ADN nucleasa de una cadena para eliminar el ADN de una cadena y ligar una secuencia adaptadora que contiene un enzima de restricción de clase II para liberar la perla magnética. Esta realización puede comprender o no la amplificación de la secuencia de ADN aislada. Además, la secuencia adaptadora puede utilizarse o no como molde para el cebador oligonucleótido de PCR. En esta realización, la secuencia adaptadora puede contener o no un identificador o etiqueta de secuencia.

[0039] En otra realización, el método de reducción de la complejidad comprende exponer la muestra de ADN a una proteína de unión con rror de apareamiento y digerir la muestra con una exonucleasa 3' a 5' y después con una nucleasa de cadenas individuales. Esta realización puede incluir o no la utilización de una perla magnética unida a la proteína ligante con error de apareamiento.

[0040] En otra realización de la presente invención, la reducción de la complejidad comprende el método CHIP tal como se describe posteriormente en la presente memoria, o el diseño de cebadores de PCR dirigidos contra motivos conservados, tales como SSRs, regiones NBS (regiones ligantes de nucleótidos), secuencias de promotores/intensificadores, secuencias de consenso de telómeros, genes de caja MADS, familias génicas de ATPasa y otras familias génicas.

[0041] En la etapa c), se llevan a cabo las etapas a) y b) consecutiva o simultáneamente con un segunda o posterior muestra de ácidos nucleicos de interés, obteniendo una segunda o posterior biblioteca de la segunda o posterior muestra de ácidos nucleicos de interés. Dicha segunda o posterior muestra de ácidos nucleicos de interés preferentemente también es una muestra compleja de ácidos nucleicos, tal como ADN genómico total. Resulta preferente que la muestra compleja de ácidos nucleicos sea ADN genómico total. También resulta preferente que dicha segunda o posterior muestra de ácidos nucleicos esté relacionada con la primera muestra de ácidos nucleicos. La primera muestra de ácidos nucleicos y el segundo o posterior ácido nucleico pueden ser, por ejemplo, diferentes líneas de una planta, tal como diferentes líneas de la planta del pimiento, o diferentes variedades. Las etapas a) y b) pueden llevarse a cabo para meramente una segunda muestra de ácidos nucleicos de interés, aunque también pueden llevarse a cabo adicionalmente para una tercera, cuarta, quinta, etc., muestra de ácidos nucleicos de interés.

[0042] Debe indicarse que el método según la presente invención resultará más útil al llevar a cabo reducción de la complejidad utilizando el mismo método y bajo condiciones sustancialmente iguales, preferentemente idénticas, para la primera muestra de ácidos nucleicos y la segunda o posterior muestras de ácidos nucleicos. Bajo dichas condiciones, se obtienen fracciones similares (comparables) de las muestras (complejas) de ácidos nucleicos.

[0043] En la etapa d), se secuencia por lo menos una parte de la primer biblioteca y de la segunda o posterior bibliotecas. Resulta preferente que la cantidad de solapamiento de los fragmentos secuenciados de la primera biblioteca y segunda o posterior bibliotecas sea de por lo menos 50%, más preferentemente de por lo menos 60%, todavía más preferentemente de por lo menos 70%, todavía más preferentemente de por lo menos 80%, todavía más preferentemente de por lo menos 90% y todavía más preferentemente de por lo menos 95%.

[0044] La secuenciación puede llevarse a cabo, en principio, por cualquier medio conocido de la técnica, tal como el método de terminación de cadena dideoxi. Sin embargo, resulta preferente que la secuenciación se lleve a cabo utilizando métodos de secuenciación de alto rendimiento, tales como los métodos dados a conocer en las patentes WO 03/004690, 03/054142, 2004/069849, 2004/070005, 2004/070007 y 2005/003375 (todas a nombre

de 454 Corporation), en Seo *et al.*, Proc. Natl. Acad. Sci. USA 101:5488-93, 2004, y las técnicas de Helios, Solexa, US Genomics, etc. Resulta más preferente que la secuenciación se lleve a cabo utilizando el aparato y/o método dado a conocer en las patentes WO 03/004690, 03/054142, 2004/069849, 2004/070005, 2004/070007 y 2005/003375 (todas a nombre de 454 Corporation). La tecnología descrita permite la secuenciación de 40 millones de bases en una única operación y es 100 veces más rápida y económica que la tecnología competidora. La tecnología de secuenciación consiste en términos generales de 4 etapas: 1) fragmentación del ADN y ligación de adaptadores específicos a una biblioteca de ADN de una cadena (ADNss), 2) hibridación de ADNss a perlas y emulsificación de las perlas en microrreactores de agua en aceite, 3) deposición de las perlas que portan ADN en una placa PicoTiterPlate®, y 4) secuenciación simultánea en 100.000 pocillos mediante generación de una señal lumínica del pirofosfato. El método se explica en mayor detalle posteriormente.

[0045] En la etapa e), las secuencias obtenidas en la etapa d) se alinean proporcionando una alineación. Los métodos de alineación de secuencias con fines de comparación son bien conocidos de la técnica. Se describen diversos programas y algoritmos de alineación en: Smith y Waterman, Adv. Appl. Math. 2:482, 1981; Needleman y Wunsch, J. Mol. Biol. 48:443, 1970; Pearson y Lipman, Proc. Natl. Acad. Sci. USA 85:2444, 1988; Higgins y Sharp, Gene 73:237-244, 1988; Higgins y Sharp, CABIOS 5:151-153, 1989; Corpet *et al.*, Nucl. Acids Res. 16:10881-90, 1988; Huang *et al.*, Computer Appl. in the Biosci. 8:155-65, 1992, y Pearson *et al.*, Meth. Mol. Biol. 24:307-31, 1994. Altschul *et al.*, Nature Genet. 6:119-29, 1994, presentan una consideración detallada de los métodos de alineación de secuencias y de los cálculos de homologías.

[0046] La herramienta Basic Local Alignment Search Tool (BLAST) del NCBI (Altschul *et al.*, 1990) se encuentra disponible de varias fuentes, incluyendo del National Center for Biological Information (NCBI, Bethesda, Md.) y en internet, para la utilización con los programas de análisis de secuencias blastp, blastn, blastx, tblastn y tblastx. Se puede obtener acceso a los mismos en la dirección <http://www.ncbi.nlm.nih.gov/BLAST/>. Una descripción de como determinar la identidad de la secuencia usando este programa está disponible en la dirección <http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html>. Una aplicación adicional podría ser la exploración de microsatélites (ver Varshney *et al.*, Trends in Biotechn. 23(1):48-55, 2005).

[0047] Típicamente la alineación se lleva a cabo en datos de secuencias que han sido recortados para los adaptadores/cebadores y/o identificadores, es decir, utilizando únicamente los datos de secuencia de los fragmentos que se originan de la muestra de ácidos nucleicos. Típicamente, los datos de secuencia obtenidos se utilizan para identificar el origen del fragmento (es decir, la muestra de procedencia), se eliminan de los datos las secuencias derivadas de adaptador y/o identificador y se lleva a cabo la alineación en este conjunto recortado.

[0048] En la etapa f), se determina uno o más polimorfismos entre la primera muestra de ácidos nucleicos y la segunda o posterior muestra de ácidos nucleicos en la alineación. La alineación puede realizarse de manera que las secuencias derivadas de la primera muestra de ácidos nucleicos y la segunda o posterior muestras de ácidos nucleicos puedan compararse. A continuación, pueden identificarse las diferencias que reflejen polimorfismos.

[0049] En la etapa g), el polimorfismo o polimorfismos determinados en la etapa g) se utilizan para diseñar sondas de detección, por ejemplo para la detección mediante hibridación en chips de ADN o en una plataforma de detección basada en perlas. Las sondas de detección se diseñan de manera que se refleje un polimorfismo en las mismas. En el caso de los polimorfismos de un solo nucleótido (SNPs), las sondas de detección típicamente contienen los alelos variantes de SNPs en la posición central de manera que se maximice la discriminación de alelos. Dichas sondas pueden utilizarse ventajosamente para cribar muestras de ensayo que presentan un determinado polimorfismo. Las sondas pueden sintetizarse utilizando cualquier método conocido de la técnica. Las sondas típicamente se diseñan de manera que resulten adecuadas para métodos de cribado de alto rendimiento.

[0050] En la etapa h), se proporciona una muestra de ensayo de ácidos nucleicos de interés. La muestra de ensayo de ácidos nucleicos puede ser cualquier muestra, aunque preferentemente es otra línea o variedad que debe mapearse para identificar polimorfismos. Comúnmente se utiliza una colección de muestras de ensayo que representa el plasma germinal de los organismos estudiados con el fin de validar experimentalmente que el polimorfismo (SN) es genuino y detectable, y para calcular las frecuencias alélicas de los alelos observados. Opcionalmente se incluyen muestras de una población de mapeado genético en la etapa de validación con el fin de determinar también la posición en el mapa genético del polimorfismo.

[0051] En la etapa i), se lleva a cabo la reducción de complejidad de la etapa b) en la muestra de ensayo de ácidos nucleicos de interés, proporcionando una biblioteca de ensayo de la muestra de ensayo de ácidos nucleicos. Resulta altamente preferente que durante todo el método según la presente invención, se utilice el mismo método para la reducción de complejidad, utilizando condiciones sustancialmente iguales, preferentemente idénticas, cubriendo de esta manera una fracción similar de la muestra. Sin embargo, no resulta necesario obtener una biblioteca de ensayo etiquetada, aunque puede encontrarse presente una etiqueta en los fragmentos en la biblioteca de ensayo.

[0052] En la etapa j), la biblioteca de ensayo se somete a cribado de alto rendimiento para identificar la presencia, ausencia o cantidad de los polimorfismos determinados en la etapa f) utilizando las sondas de

detección diseñadas en la etapa g). El experto en la materia conoce varios métodos para el cribado de alto rendimiento utilizando sondas. Resulta preferente que una o más sondas diseñadas utilizando la información obtenida en la etapa g) se inmovilicen en una matriz, tal como un chip de ADN, y que dicha matriz posteriormente se ponga en contacto con la biblioteca de ensayo bajo condiciones de hibridación. Los fragmentos de ADN en la biblioteca de ensayo que sean complementarios a una o más sondas en la matriz se hibridarán bajo dichas condiciones con dichas sondas, y podrán detectarse de esta manera. También se encuentran contemplados otros métodos de cribado de alto rendimiento dentro del alcance de la presente invención, tales como la inmovilización de la biblioteca de ensayo obtenida en la etapa j) y la puesta en contacto de dicha biblioteca de ensayo inmovilizada con las sondas diseñadas en la etapa h) bajo condiciones de hibridación.

[0053] Affymetrix, entre otros, proporciona otra técnica de cribado mediante secuenciación de alto rendimiento que utiliza la detección basada en un chip de SNPs y tecnología de perlas proporcionada por Illumina.

[0054] En una realización ventajosa, la etapa b) en el método según la presente invención comprende además la etapa de etiquetar la biblioteca para obtener una biblioteca etiquetada, y dicho método comprende además la etapa c1) de combinar la primera biblioteca etiquetada y una segunda o posteriores bibliotecas etiquetadas para obtener una biblioteca combinada.

[0055] Resulta preferente que el etiquetado se lleve a cabo durante la etapa de reducción de la complejidad para reducir el número de etapas requerido para obtener la primera biblioteca etiquetada de la primera muestra de ácidos nucleicos. Dicho etiquetado simultáneo puede conseguirse mediante, por ejemplo, AFLP, utilizando adaptadores que comprenden un identificador (nucleótido) único para cada muestra.

[0056] El etiquetado pretende distinguir entre muestras de origen diferente, por ejemplo obtenidas de diferentes líneas vegetales, en el caso de que se combinen bibliotecas de dos o más muestras de ácidos nucleicos para obtener una biblioteca de combinación. De esta manera, preferentemente se utilizan etiquetas diferentes para preparar las bibliotecas etiquetadas de la primera muestra de ácidos nucleicos y la segunda o posteriores muestras de ácidos nucleicos. En el caso de que se utilicen, por ejemplo, cinco muestras de ácidos nucleicos, se pretenden obtener cinco bibliotecas etiquetadas diferentemente, indicando las cinco etiquetas diferentes las muestras originales respectivas.

[0057] La etiqueta puede ser cualquier etiqueta conocida de la técnica para distinguir muestras de ácidos nucleicos, aunque preferentemente es una secuencia identificadora corta. Dicha secuencia identificadora puede ser, por ejemplo, una secuencia de bases única de longitud variable utilizada para indicar el origen de la biblioteca obtenida mediante reducción de complejidad.

[0058] En una realización preferente, el etiquetado de la primera biblioteca y de la segunda o posteriores bibliotecas se lleva a cabo utilizando etiquetas diferentes. Tal como se ha comentado anteriormente, resulta preferente que cada biblioteca de una muestra de ácidos nucleicos se identifique con su propia etiqueta. La muestra de ensayo de ácidos nucleicos no requiere ser etiquetada.

[0059] En una realización preferente de la invención, se lleva a cabo la reducción de la complejidad por medio de AFLP® (Keygene N.V., Países Bajos, ver, por ejemplo, la patente EP 0 534 858 y Vos *et al.*, AFPL: a new technique for DNA fingerprinting, Nucleic Acids Research 23(21):4407-4414, 1995).

[0060] La AFLP es un método para la amplificación selectiva de fragmentos de restricción. La AFLP no requiere información de secuencia previa y puede llevarse a cabo en cualquier ADN de partida. En general, la AFLP comprende las etapas de:

(a) digestión de un ácido nucleico, en particular un ADN o ADNc, con una o más endonucleasas de restricción específicas, para fragmentar el ADN en una serie correspondiente de fragmentos de restricción;

(b) ligación de los fragmentos de restricción obtenidos de esta manera con un adaptador oligonucleótido sintético de doble cadena, un extremo del cual es compatible con uno o ambos extremos de los fragmentos de restricción, produciendo de esta manera fragmentos de restricción ligados con adaptadores, preferentemente etiquetados, del ADN de partida;

(c) puesta en contacto de los fragmentos de restricción ligados con adapta dor, preferentemente etiquetados, bajo condiciones de hibridación con por lo menos un cebador oligonucleótido que contenga por lo menos un nucleótido selectivo en su extremo 3';

(d) amplificación de los fragmentos de restricción ligados con adaptadores, preferentemente etiquetados, hibridados con los cebadores mediante PCR o una técnica similar de manera que se provoca el alargamiento adicional de los cebadores hibridados a lo largo de los fragmentos de restricción del ADN de partida con el que se hibridaron los cebadores; y (e) detección, identificación o recuperación del fragmento de ADN amplificado o alargado obtenido de esta manera.

[0061] La AFLP proporciona de esta manera un subconjunto reproducible de fragmentos ligados con adaptadores. Otros métodos adecuados para la reducción de la complejidad son la inmunoprecipitación de la

cromatina (ChiP). Lo anterior se refiere al aislamiento del ADN nuclear, mientras que las proteínas tales como factores de transcripción se unen al ADN. Con ChiP en primer lugar se utiliza un anticuerpo contra la proteína, resultando en un complejo de proteína Ab-ADN. Mediante la purificación de este complejo y su precipitación, se selecciona el ADN al que se une dicha proteína. A continuación, puede utilizarse el ADN para la construcción y secuenciación de la biblioteca, es decir, éste es un método para llevar a cabo una reducción de la complejidad de un modo no aleatorio dirigido a áreas funcionales específicas; en el presente ejemplo, factores de transcripción específicos.

[0062] Una variante útil de la tecnología AFLP utiliza nucleótidos no selectivos (es decir, cebadores +0/+0) y en ocasiones se denomina PCR de línkers. También proporciona una reducción de complejidad muy adecuada.

[0063] Para una descripción adicional de la AFLP, sus ventajas, realizaciones, así como las técnicas, enzimas, adaptadores, cebadores y compuestos adicionales y herramientas utilizados en las mismas, se hace referencia a las patentes US 6.045.994, EP-B-0 534 858, EP 976835 y EP 974672, WO 01/88189, y Vos *et al.*, Nucleic Acids Research 23:4407-4414, 1995.

[0064] De esta manera, en una realización preferente del método de la presente invención, se lleva a cabo la reducción de la complejidad mediante:

- digestión de la muestra de ácidos nucleicos con por lo menos una endonucleasa de restricción para fragmentarla en fragmentos de restricción;
- ligación de los fragmentos de restricción obtenidos con por lo menos un adaptador oligonucleótido sintético de doble cadena que presenta un extremo compatible con uno o ambos extremos de los fragmentos de restricción para producir fragmentos de restricción ligados con adaptadores;
- poner en contacto dichos fragmentos de restricción ligados con adaptadores con uno o más cebadores oligonucleótidos bajo condiciones de hibridación; y
- amplificación de dichos fragmentos de restricción ligados con adaptadores mediante alargamiento de uno o más cebadores oligonucleótidos,

en la que por lo menos uno o más cebadores oligonucleótidos incluye una secuencia de nucleótidos que presenta la misma secuencia de nucleótidos que las partes terminales de las cadenas en los extremos de dichos fragmentos de restricción ligados con adaptadores, incluyendo los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción, e incluyendo por lo menos parte de los nucleótidos presentes en los adaptadores, en la que, opcionalmente, por lo menos uno de dichos cebadores incluye en su extremo 3' una secuencia seleccionada que comprende por lo menos un nucleótido situado inmediatamente contiguo a los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción.

[0065] La AFLP es un método altamente reproducible para la reducción de la complejidad y por lo tanto resulta particularmente adecuado para el método según la presente invención.

[0066] En una realización preferente del método según la presente invención, el adaptador o el cebador comprende una etiqueta. Éste es particularmente el caso para la identificación real de los polimorfismos, en donde resulta importante distinguir entre secuencias derivadas de bibliotecas separadas. La incorporación de una etiqueta oligonucleótida en un adaptador o cebador resulta muy conveniente debido a que no resultan necesarias etapas adicionales para etiquetar una biblioteca.

[0067] En otra realización, la etiqueta es una secuencia identificadora. Tal como se ha comentado anteriormente, dicha secuencia identificadora puede ser de longitud variable dependiendo del número de muestras de ácidos nucleicos que debe compararse. Resulta suficiente una longitud de aproximadamente 4 bases (4^4=256 secuencias de etiqueta diferentes posibles) para distinguir entre el origen de un número limitado de muestras (como máximo 256), aunque resulta preferente que las secuencias de etiqueta difieran en no más de una base entre las muestras que deben distinguirse. Según resulte necesario, puede ajustarse la longitud de las secuencias de etiqueta.

[0068] En una realización, la secuenciación se lleva a cabo en un soporte sólido, tal como una perla (ver, por ejemplo, las patentes WO 03/004690, 03/054142, 2004/069849, 2004/070005, 2004/070007 y 2005/003375 (todas a nombre de 454 Corporation).

[0069] Dicho método de secuenciación resulta particularmente adecuado para la secuenciación económica y eficiente de muchas muestras simultáneamente.

[0070] En una realización preferente, la secuenciación comprende las etapas de:

- unir con perlas fragmentos ligados con adaptadores, estando unida cada perla a un único fragmento ligado con adaptador;

- emulsionar las perlas en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una única perla;

- cargar las perlas en pocillos, comprendiendo cada pocillo una única perla; y

- generar una señal de pirofosfato.

[0071] En la primera etapa, los adaptadores de secuenciación se ligan a fragmentos en la biblioteca de combinación. Dicho adaptador de secuenciación incluye por lo menos una región "clave" para la unión a una perla, una región de cebador de secuenciación y una región de cebador de PCR. De esta manera se obtienen fragmentos ligados a adaptadores.

[0072] En una etapa adicional, se unen fragmentos ligados con adaptadores a perlas, uniéndose cada perla a un único fragmento ligado a adaptador. Al grupo de fragmentos ligados con adaptadores se añaden perlas en exceso para garantizar la unión de un solo fragmento ligado a adaptador por perla para la mayoría de las perlas (distribución de Poisson).

[0073] En la etapa siguiente, se emulsionan las perlas en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una única perla. Los reactivos de PCR presentes en los microrreactores de agua en aceite permiten que tenga lugar una reacción de PCR en los microrreactores. A continuación, se rompen los microrreactores y se enriquece para las perlas que comprenden ADN (perlas positivas para ADN).

[0074] En una etapa posterior, se cargas las perlas en pocillos, comprendiendo cada pocillo una única perla. Los pocillos preferentemente son parte de una placa PicoTiter TM que permite la secuenciación simultánea de una gran cantidad de fragmentos.

[0075] Tras la adición de perlas que portan enzima, se determina la secuencia de los fragmentos mediante pirosecuenciación. En etapas sucesivas, la placa picotiter y las perlas, así como las perlas con enzima en la misma se someten a diferentes desoxirribonucleótidos en presencia de reactivos de secuenciación convencionales, y tras la incorporación de un desoxirribonucleótido, se genera una señal lumínica que se registra. La incorporación del nucleótido correcto genera una señal de pirosecuenciación que puede detectarse.

[0076] La pirosecuenciación misma es conocida de la técnica y se describe, en otros, en www.biotagebio.com, www.pyrosequencing.com/tab technology. La tecnología se aplica además en, por ejemplo, las patentes WO 03/004690, 03/054142, 2004/069849, 2004/070005, 2004/070007 y 2005/003375 (todas a nombre de 454 Corporation).

[0077] El cribado de alto rendimiento de la etapa k) preferentemente se lleva a cabo mediante inmovilización de las sondas diseñadas en la etapa h) en una matriz, seguido de la puesta en contacto de la matriz que comprende las sondas con una biblioteca de ensayo bajo condiciones de hibridación. Preferentemente, la etapa de puesta en contacto se lleva a cabo bajo condiciones de hibridación astringentes (ver Kennedy *et al.*, Nat. Biotech., publicado en internet el 7 de septiembre de 2003, páginas 1 a 5). El experto en la materia es consciente de la existencia de métodos adecuados para la inmovilización de sondas en una matriz y de métodos de puesta en contacto bajo condiciones de hibridación. La tecnología típica que resulta adecuada para este fin se revisa en Kennedy *et al.*, Nat. Biotech., publicado en internet el 7 de septiembre de 2003, páginas 1 a 5).

[0078] Una aplicación ventajosa particular es el cultivo de especies poliploides. Mediante la secuenciación de cultivos poliploides con una elevada cobertura, la identificación de SNPs y los diversos alelos, y el desarrollo de sondas para la amplificación específica de alelo, pueden realizarse avances significativos en el cultivo de especies poliploides.

[0079] Como parte de la invención, se ha encontrado que la combinación de generar subconjuntos seleccionados aleatoriamente mediante amplificación selectiva para una pluralidad de muestras y la tecnología de secuenciación de alto rendimiento presenta ciertos problemas complejos que debían resolverse para mejorar adicionalmente el método descrito en la presente memoria para la identificación eficiente y de alto rendimiento de los polimorfismos. Más en detalle, se ha encontrado que al combinar múltiples muestras (es decir la primera y la segunda o posteriores) en un grupo tras realizar una reducción de complejidad, se produce el problema de que muchos fragmentos aparentemente proceden de dos muestras o, en otras palabras, se identificaron muchos fragmentos que no podían asignarse únicamente a una muestra y que de esta manera no pudieron utilizarse en el procedimiento de identificación de polimorfismos. Ello condujo a una reducción de la fiabilidad del método y a polimorfismos (SNPs, indels, SSRs) que no pudieron identificarse adecuadamente.

[0080] Tras el análisis cuidadoso y detallado de la secuencia de nucleótidos completa de los fragmentos que no pudieron localizarse, se encontró que aquellos fragmentos contenían dos adaptadores diferentes que comprendían etiquetas y que probablemente se formaban entre la generación de las muestras de complejidad reducida y la ligación de los adaptadores de secuenciación. El fenómeno se describe como "etiquetado mixto". El fenómeno descrito como "etiquetado mixto", tal como se utiliza en la presente memoria, se refiere de esta manera a fragmentos que contienen una etiqueta que relaciona el fragmento con una muestra por un lado, mientras que el lado opuesto del fragmento contiene una etiqueta que relaciona el fragmento con otra muestra. De esta manera, un fragmento aparentemente se deriva de dos muestras (*quod non*). Esto conduce a una

identificación errónea de polimorfismos y por lo tanto no resulta deseable.

[0081] Se ha teorizado con que la formación de fragmentos heterodúplex entre dos muestras se encuentra en la raíz de dicha anomalía.

[0082] Se ha encontrado la solución a dicho problema en un rediseño de la estrategia para la conversión de muestras de las que se ha reducido la complejidad en fragmentos unidos a perlas que pueden amplificarse antes de la secuenciación de alto rendimiento. En la presente realización, cada muestra se somete a reducción de complejidad y purificación opcional. A continuación, se generan extremos romos en cada muestra (pulido de extremos) seguido de ligación del adaptador de secuenciación que es capaz de unirse a la perla. Los fragmentos ligados a adaptadores de secuenciación de las muestras seguidamente se combinan y se ligan a las perlas para la polimerización en emulsión y la posterior secuenciación de alto rendimiento.

[0083] A modo de parte adicional de la presente invención, se ha encontrado que la formación de concatámeros dificultó la identificación correcta de polimorfismos. Se han identificado concatámeros como fragmentos que se forman tras "formar extremos romos" o "pulir" los productos de reducción de complejidad, por ejemplo con la ADN polimerasa de T4, y en lugar de ligarlos a los adaptadores que permiten la unión a las perlas, se ligan entre sí, creando de esta manera concatámeros, es decir, un concatámero es el resultado de la dimerización de fragmentos de extremos romos.

[0084] La solución a este problema se encontró en la utilización de determinados adaptadores modificados específicamente. Los fragmentos amplificados obtenidos de la reducción de complejidad típicamente contienen un extremo protuberante 3'-A debido a las características de determinadas polimerasas preferentes, que no presentan actividad correctora de errores de exonucleasas 3'-5'. La presencia de dicho extremo protuberante 3'-A también es el motivo de que se formen extremos romos en los fragmentos antes de la ligación de adaptadores. Al proporcionar un adaptador que podía unirse a una perla en el que el adaptador contiene un extremo protuberante 3'-T, se encontró que podía resolverse en una etapa tanto el problema de las "etiquetas mixtas" como el de los concatámeros. Una ventaja adicional de utilizar dichos adaptadores modificados es que la etapa convencional de "formación de extremos romos" y la etapa de fosforilación posteriores podían omitirse.

[0085] De esta manera, en una realización preferente adicional, tras la etapa de reducción de complejidad de cada muestra, se lleva a cabo una etapa en los fragmentos de restricción amplificados ligados con adaptadores que se han obtenido de la etapa de reducción de complejidad, de manera que a estos fragmentos se ligan adaptadores de secuenciación, los cuales contienen un extremo protuberante 3'-T y son capaces de unirse a las perlas.

[0086] Se ha encontrado además que, al fosforilar los cebadores utilizados en la etapa de reducción de complejidad, la etapa de pulido de extremos (formación de extremos romos) y la fosforilación de intermediarios previa a la ligación pueden evitarse.

[0087] De esta manera, en una realización altamente preferente de la invención, la invención se refiere a un método para identificar uno o más polimorfismos, comprendiendo dicho método las etapas de:

a) proporcionar una pluralidad de muestras de ácidos nucleicos de interés,

b) llevar a cabo una reducción de complejidad de cada una de las muestras, proporcionando una pluralidad de bibliotecas de las muestras de ácidos nucleicos, en las que la reducción de complejidad se lleva a cabo mediante:

- digestión de cada muestra de ácidos nucleicos con por lo menos una endonucleasa de restricción para fragmentarla en fragmentos de restricción;
- ligación de los fragmentos de restricción obtenidos con por lo menos un adaptador oligonucleótido sintético de doble cadena que presenta un extremo compatible con uno o ambos extremos de los fragmentos de restricción para producir fragmentos de restricción ligados con adaptadores;
- puesta en contacto de dichos fragmentos de restricción ligados con adaptadores con uno o más cebadores oligonucleótidos fosforilados bajo condiciones de hibridación; y
- amplificación de dichos fragmentos de restricción ligados con adaptadores mediante alargamiento de uno o más cebadores oligonucleótidos, en la que por lo menos uno de entre uno o más cebadores oligonucleótidos incluye una secuencia de nucleótidos que presenta la misma secuencia de nucleótidos que las partes terminales de las cadenas en los extremos de dichos fragmentos de restricción ligados con adaptadores, incluyendo los nucleótidos implicados en la formación de la secuencia diana de dicha endonucleasa de restricción, e incluyendo por lo menos parte de los nucleótidos presentes en los adaptadores, en la que, opcionalmente, por lo menos uno de dichos cebadores incluye en su extremo 3' una secuencia seleccionada que comprende por lo menos un nucleótido situado inmediatamente contiguo a los nucleótidos implicados en la formación de la secuencia diana de dicha endonucleasa de restricción y en la que el adaptador y/o el cebador contienen una etiqueta;

c) combinar dichas bibliotecas en una biblioteca combinada;

d) ligar adaptadores de secuenciación capaces de unirse a perlas a los fragmentos amplificados con caperuzas de adaptadores en la biblioteca combinada, utilizando un adaptador de secuenciación que porta un extremo protuberante 3'-T y someter los fragmentos unidos a perla a polimerización en emulsión;

e) secuenciar por lo menos una parte de la biblioteca combinada;

f) alinear la secuencias de cada muestra obtenida en la etapa e);

g) determinar uno o más polimorfismos entre la pluralidad de muestras de ácidos nucleicos en la alineación de la etapa f);

h) utilizar uno o más polimorfismos determinados en la etapa g) para diseñar sondas de detección;

i) proporcionar una muestra de ensayo de ácidos nucleicos de interés;

j) realizar la reducción de complejidad de la etapa b) en la muestra de ensayo de ácidos nucleicos de interés, proporcionando una biblioteca de ensayo de la muestra de ensayo de ácidos nucleicos;

k) someter la biblioteca de ensayo a cribado de alto rendimiento para identificar la presencia, ausencia o cantidad de polimorfismos determinados en la etapa g) utilizando las sondas de detección diseñadas en la etapa h).

Breve descripción de los dibujos

[0088]

La **figura 1A** muestra un fragmento según la presente invención unido a una perla ("perla de 454") y la secuencia del cebador utilizado para la preamplificación de las dos líneas de planta del pimiento. La expresión "fragmento de ADN" se refiere al fragmento obtenido tras la digestión con una endonucleasa de restricción, "adaptador Keygene" se refiere a un adaptador que proporciona un sitio de unión para los cebadores oligonucleótidos (fosforilados) utilizados para generar una biblioteca, "KRS" se refiere a una secuencia (etiqueta) identificadora, "adaptador SEC. de 454" se refiere a un adaptador de secuenciación, y "adaptador de PCR de 454" se refiere a un adaptador que permite la amplificación en emulsión del fragmento de ADN. El adaptador de PCR permite la unión a la perla y la amplificación, y puede contener un extremo protuberante 3'-T.

La **figura 1B** muestra un cebador esquemático utilizado en la etapa de reducción de la complejidad. Dicho cebador generalmente comprende una región de sitio de reconocimiento indicado como (2), una región constante que puede incluir una sección de etiqueta indicada como (1) y uno o más nucleótidos selectivos en una región selectiva indicada como (3) en el extre mo 3' de los mismos).

La **figura 2** muestra la estimación de concentración de ADN utilizando electroforesis en gel de agarosa al 2%. S1 se refiere a PSP11; S2 se refiere a PI201234. 50, 100, 250 y 500 ng se refieren, respectivamente, a 50 ng, 100 ng, 250 ng y 500 ng para estimar las cantidades de ADN de S1 y de S2. Las figs. 2C y 2D muestran la determinación de la concentración de ADN utili zando espectrofotometría NanoDrop.

La **figura 3** muestra los resultados de las evaluaciones de calidad de intermediarios del

Ejemplo 3.

La **figura 4** muestra gráficos de flujo del procesamiento de datos de secuencia, es decir, las etapas entre la generación de los datos de secuenciación y la identificación de los SNPs, SSRs e indels putativos, mediante etapas de eliminación de información de secuencia conocida en recorte y etiquetado, resultando en datos de secuencia ajustados que se agrupan y se ensamblan para proporcionar contigs y singletons (fragmentos que no pueden ensamblarse para formar un contig), después de lo cual pueden identificarse y evaluarse polimorfismos putativos. La figura 4B proporciona detalles adicionales del procedimiento de exploración de polimorfismos.

La **figura 5** se refiere al problema de las etiquetas mixtas y proporciona en el panel 1 un ejemplo de una etiqueta mixta que incluye etiquetas asociadas a la muestra 1 (MS1) y a la muestra 2 (MS2). El panel 2 proporciona una explicación esquemática del fenómeno. Los fragmentos de restricción de la AFLP derivados de la muestra 1 (S1) y de la muestra 2 (S2) se ligan utilizando adaptadores ("adaptador de Keygene") en ambos extremos que portan etiquetas específicas de las muestras S1 y S2. Tras la amplificación y secuenciación, los fragmentos esperados presentan las etiquetas S1-S2 y las etiquetas S2-S2. Además, inesperadamente se observaron fragmentos que portaban etiquetas S1-S2 ó S2-S1. El panel 3 explica la causa hipotética de que se generasen etiquetas mixtas, por la que se forman productos heterodúplex a partir de fragmentos de las muestras 1 y 2. Los heterodúplex posteriormente se liberan, debido a la actividad exonucleasa 3'-5' de la ADN polimerasa de T4 o Klenow, de los extremos 3'-protuberantes. Durante la polimerización, se rellenan los huecos con nucleótidos y se introduce la eti queta incorrecta. Esto funciona para heterodúplex de aproximadamente la misma longitud (panel superior), aunque también para heterodúplex de longitud más variable. El panel

4 proporciona en la parte derecha el protocolo convencional que conduce a la formación de etiquetas mixtas y, en la parte derecha, el protocolo modificado.

La **figura 6** se refiere al problema de la formación de concatámeros, en la cual en el panel 1 se proporciona un ejemplo típico de concatámero, en el que las diversas secciones de adaptador y de etiqueta se encuentran subra yadas y con su origen (es decir, MS1, MS2, ES1 y ES2, correspondiendo respetivamente a un adaptador-sitio de restricción MseI de la muestra 1, adaptador-sitio de restricción MseI de la muestra 2, adaptador-sitio de restricción EcorI de la muestra 1, adaptador-sitio de restricción EcoRI de la muestra 2). El panel 2 muestra los fragmentos esperados que portan las etiquetas S1-S1 y S2-S2 y el observado aunque inesperado S1-S1-S2-S2, que es un concatámero de fragmentos de las muestras 1 y 2. El panel 3 pro porciona la solución para evitar la generación de concatámeros, así como de etiquetas mixtas, mediante la introducción de un extremo protuberante en los adaptadores de AFLP, adaptadores de secuenciación modificados y la omisión de la etapa de pulido de extremos al ligar los adaptadores de secuenciación. No se observó formación de concatámeros debido a que los fragmentos de ALP no pueden ligarse entre sí y no se producen fragmentos mixtos debido a que se omite la etapa de pulido de extremos. El panel 4 proporciona el protocolo modificado, que utiliza adaptadores modificados para evitar la formación de concatámeros, así como de etiquetas mixtas.

Figura 7. Alineación múltiple "10037_CL989contig2" de secuencias de fragmentos de la AFLP de la planta del pimiento, que contiene un polimorfismo de un solo nucleótido (SNP) putativo. Observar que el SNP (indicado por una flecha negra) se encuentra definido por un alelo A presente en ambas lecturas de la muestra 1 (PSP11), indicadas por la presencia de la etiqueta MS1 en el nombre de las dos lecturas superiores y un alelo G presente en la muestra 2 (PI201234), indicado por la presencia de la etiqueta MS2 en el nombre de las dos lecturas de la parte inferior. Los nombres de lectura se muestran en la parte izquierda. La secuencia de consenso de esta alineación múltiple es (5'-3'):

TAACACGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACA[A/G]TGTTGGTTTT

GGTGCTAACTTCAACCCCACTACTGTTTTGCTCTATTTTTG.

Figura 8A. Representación esquemática de la estrategia de enriquecimiento de repeticiones de secuencia simple de direccionamiento (SSRs) en combinación con la secuenciación de alto rendimiento para la identificación *de novo* de SSRs.

Figura 8B. Validación de un SNP G/A en la planta del pimiento utilizando la detección SNPWave. P1 = PSP11, P2 = PI201234. Se indican los ocho descendientes RIL con los números 1 a 8.

Ejemplos

Ejemplo 1

[0089] Se generó una mezcla de ligación de restricción EcoRI/MseI (1) a partir de ADN genómico de las líneas de la planta del pimiento PSP-11 y PI20234. La mezcla de ligación de restricción se diluyó 10 veces y se preamplificaron 5 microlitros de cada muestra (2) con cebadores EcoRI +1(A) y MseI +1(C) (conjunto I). Tras la amplificación, se comprobó la calidad del producto de preamplificación de las dos muestras de pimiento en un gel de agarosa al 1%. Los productos de preamplificación se diluyeron 20 veces, seguido de una preamplificación mediante AFLP con KRSEcoRI +1(A) y KRSMseI +2(CA). Los segmentos (identificadores) KRS se encuentran subrayados y los nucleótidos selectivos se encuentran en negrita, en el extremo 3' de las secuencias de cebadores SEC ID 1 a 4, más abajo. Tras la amplificación, se comprobó la calidad del producto de preamplificación de las dos muestras de pimiento en un gel de agarosa al 1% y mediante la técnica de la huella genética utilizando AFLP (4) con EcoRI +3(A) y MseI + 3(C) (3). Los productos de preamplificación de las dos líneas del pimiento se purificaron separadamente en una columna de PCR de Qiagen (5). Se midió la concentración de las muestras en el NanoDrop. Se mezcló y se secuenció un total de 5.006,4 ng de PSP-11 y 5.006,4 ng de PI20234.

Conjunto I de cebadores utilizado para la preamplificación de PSP-11:

E01LKRS1 5'-CGTCAGACTGCGTACCAATTC**A**-3' [SEC ID 1]

M15KKRS1 5'-TGGTGATGAGTCCTGAGTAAC**A**-3' [SEC ID 2]

Conjunto II de cebadores utilizado para la preamplificación de PI20234:

E01LKRS2 5'-CAAGAGACTGCGTACCAATTC**A**-3' [SEC ID 3]

M15KKRS2 5'-AGCCGATGAGTCCTGAGTAAC**A**-3' [SEC ID 4]

(1) Mezcla de ligación de restricción EcoRI/MseI

[0090]

Mezcla de restricción (40 µl/muestra)

ADN	6 µl (±300 ng)
EcoRI (5 U)	0,1 µl
MseI (2U)	0,05 µl
5xRL	8 µl
MQ	25,85 µl
Total	40 µl

[0091] Incubación durante 1 hora a 37ºC.

[0092] Adición de:

Mezcla de ligación (10 µl/muestra)

ATP 10 mM	1 µl
ADN ligasa de T4	1 µl
Adapt. EcoRI (5 pmol/µl)	1 µl
Adapt. MseI (50 pmol/µl)	1 µl
5xRL	2 µl
MQ	4 µl
Total	10 µl

[0093] Incubación durante 3 horas a 37ºC.

Adaptador EcoRI

[0094]

91M35/91M36: *-CTCGTAGACTGCGTACC :91M35 [SEC ID 5]

± bio CATCTGACGCATGGTTAA :91M36 [SEC ID 6]

Adaptador MseI

[0095]

92A18/92A19: 5-GACGATGAGTCCTGAG-3' :92A18 [SEC ID 7]

3-TACTCAGGACTCAT-5 :92A19 [SEC ID 8]

(2) Pre-amplificación

Preamplificación (A/C):

[0096]

Mezcla RL (10x)	5 µl
EcoRI-pr E01L (50 ng/µl)	0,6 µl
MseI-pr M02K (50 ng/µl)	0,6 µl

dNTPs (25 mM)	0,16 µl
Pol. Taq (5 U)	0,08 µl
10XPCR	2,0 µl
MQ	11,56 µl
Total	20 µl/reacción

[0097] Perfil térmico de la preamplificación.

[0098] Se llevó a cabo la preamplificación selectiva en un volumen de reacción de 50 µl. Se llevó a cabo la PCR en un sistema GeneAmp 9700 de PE y se inició un perfil de 20 ciclos con una etapa de desnaturalización a 94ºC durante 30 segundos, seguido de una etapa de hibridación a 56ºC durante 60 segundos y una etapa de extensión a 72ºC durante 60 segundos.

EcoRI +1(A)[1]

E01 L 92R11: 5-AGACTGCGTACCAATTCA-3 [SEC ID 9]

MseI +1(C)[1]

M02k 93E42: 5-GATGAGTCCTGAGTAAC-3 [SEC ID 10]

Preamplificación A/CA:

[0099]

Mezcla PA+1/+1 (20x) :	5 µl
EcoRI-pr :	1,5 µl
MseI-pr :	1 µl
dNTPs (25 mM) :	0,4 µl
Pol. Taq (5 U) :	0,2 µl
10XPCR :	5 µl
MQ :	36,3 µl
Total :	50 µl

[0100] Se realizó la preamplificación selectiva en un volumen de reacción de 50 µl. Se llevó a cabo la PCR en un sistema GeneAmp 9700 de PE y un perfil de 30 ciclos que se inició con una etapa de desnaturalización a 94ºC durante 30 segundos, seguido de una etapa de hibridación a 56ºC durante 60 segundos y una etapa de extensión a 72ºC durante 60 segundos.

(3) KRSEcoRI +1(A) y KRSMsel + 2(CA)[2]

[0101]

05F212 E01LKRS1 CGTCAGACTGCGTACCAATTC**A**-3' [SEC ID 11]

05F213 E01LKRS2 CAAGAGACTGCGTACCAATTC**A**-3' [SEC ID 12]

05F214 M15KKRS1 TGGTGATGAGTCCTGAGTAAC**A**-3' [SEC ID 13]

05F215 M15KKRS2 AGCCGATGAGTCCTGAGTAAC**A**-3' [SEC ID 14]

nucleótidos selectivos en negrita y etiquetas (KRS) subrayadas

Muestra PSP11 : E01LKR**S1**/M15KKR**S1**

Muestra PI120234 : E01LKR**S2**/M15KKR**S2**

(4) Procolo de AFLP

[0102] Se realizó una amplificación selectiva en un volumen de reacción de 20 µl. Se llevó a cabo una PCR en un sistema GeneAmp 9700 de PE. Se inició un perfil de 13 ciclos con una etapa de desnaturalización a 94ºC durante 30 segundos, seguido de una etapa de hibridación a 65ºC durante 30 segundos, con una etapa de reducción en la que se redujo la temperatura de hibridación en 0,7ºC en cada ciclo, y una etapa de extensión a 72ºC durante 60 segundos. A este perfil siguió un perfil de 23 ciclos con una etapa de desnaturalización a 94ºC durante 30 segundos, seguido de una etapa de hibridación a 56ºC durante 30 segundos y una etapa de extensión a 72ºC durante 60 segundos.

EcoRI +3 (**AAC**) y MseI +3 (**CAG**)

E32 92S02: 5-GACTGCGTACCAATTC**AAC**-3 [SEC ID 15]

M49 92G23: 5'-GATGAGTCCTGAGTAA**CAG**-3 [SEC ID 16]

(5) Columna de Qiagen

[0103] Se llevó a cabo una purificación Qiagen siguiendo las instrucciones del fabricante: manual de QIAquick® Spin

(http://www.qiagen.com/literature/handbooks/PDF/DNACleanupAndConcentration/QQ_Spin/1021422_HBQQSpin07002WW.pdf).

Ejemplo 2: PIMIENTO

[0104] Se utilizó ADN procedente de las líneas del pimiento PSP-11 y PI20234 para generar producto de AFLP mediante la utilización de cebadores específicos de sitio de reconocimiento de *Keygene* para AFLP (estos cebadores de AFLP son esencialmente iguales a los cebadores convencionales de AFLP, por ejemplo los descritos en la patente EP 0 534 858, y generalmente contienen una región de sitio de reconocimiento, una región constante y uno o más nucleótidos selectivos en una región selectiva.

Procedentes de las líneas del pimiento PSP-11 ó PI20234 se digirieron 150 ng de ADN con las endonucleasas de restricción *EcoRI* (5 U/reacción) y *MseI* (2 U/reacción) durante 1 hora a 37ºC, seguido de la inactivación durante 10 minutos a 80ºC. Los fragmentos de restricción obtenidos se ligaron con adaptador oligonucleótido sintético de doble cadena, un extremo del cual era compatible con uno o ambos extremos de los fragmentos de restricción EcoRI y/o MseI. Se llevaron a cabo reacciones de preamplificación de AFLP (20 µl/reacción) con los cebadores de AFLP +1/+1 en mezcla de restricción-dilución diluida 10 veces. Perfil de PCR: 20*(30 segundos a 94ºC + 60 segundos a 56ºC + 120 segundos a 72ºC). Se llevaron a cabo reacciones de AFLP adicionales (50 µl/reacción) con diferentes cebadores específicos de sitio de reconocimiento de Keygene para AFLP +1 *EcoRI* y +2 *MseI* (Tabla, más abajo; las etiquetas se indican en negrita, los nucleótidos selectivos se han subrayado) en producto de preamplificación de AFLP EcoRI/MseI +1/+1 diluido 20 veces: 30*(30 segundos a 94ºC + 60 segundos a 56ºC + 120 segundos a 72ºC). El producto de AFLP se purificó mediante la utilización del kit de purificación por PCR QIAquick (QIAGEN) siguiendo el manual 07/2002 de QIAquick® Spin, página 18, y la concentración se midió con un espectrofotómetro NanoDrop® ND-1000. Se combinó un total de 5 µg de producto de AFLP PSP-11 +1/+2 y 5 µg de producto de AFLP PI20234 +1/+2 y se resolvió en 23,3 µl de TE. Finalmente, se obtuvo una mezcla con una concentración de 430 ng/µl +1/+2 de producto de AFLP.

Tabla

SEC ID	Cebador de PCR	Cebador -3'	Pimiento	Reacción de AFLP
[SEC ID 17]	05F21	**CGTC**AGACTGCGTACCAATTCA	PSP-	1
[SEC ID 18]	05F21	**TGGT**GATGAGTCCTGAGTAACA	PSP-	1
[SEC ID 19]	05F21	**CAAG**AGACTGCGTACCAATTCA	PI2023	2
[SEC ID 20]	05F21	**AGCC**GATGAGTCCTGAGTAACA	PI2023	2

Ejemplo 3: maíz

[0105] Se utilizó ADN de las líneas de maíz B73 y M017 para generar producto de AFLP mediante la utilización de cebadores específicos de *sitio de reconocimiento de Keygene* para AFLP (estos cebadores de AFLP son esencialmente los mismos que los cebadores convencionales de AFLP, por ejemplo descritos en la patente EP 0 534 858, y generalmente contienen una región de sitio de reconocimiento, una región constante y

uno o más nucleótidos selectivos en el extremo 3' de los mismos).

[0106] Se digirió ADN de las líneas del pimiento B73 o M017 con las endonucleasas de restricción *TaqI* (5 U/reacción) durante 1 hora a 65ºC y *MseI* (2 U/reacción) durante 1 hora a 37ºC, seguido de la inactivación durante 10 minutos a 80ºC. Los fragmentos de restricción obtenidos se ligaron con adaptador oligonucleótido sintético de doble cadena, un extremo del cual es compatible con uno o ambos extremos de los fragmentos de restricción TaqI y/o MseI.

[0107] Se llevaron a cabo reacciones de preamplificación de AFLP (20 µl/reacción) con cebadores de AFLP +1/+1 en mezcla de restricción-ligación diluida 10 veces. Perfil de PCR: 20*(30 segundos a 94ºC + 60 segundos a 56ºC + 120 segundos a 72ºC). Se llevaron a cabo reacciones de AFLP adicionales (50 µl/reacción) con diferentes cebadores de sitio de reconocimiento de Keygene para FLP *TaqI* y *MseI* +2 (Tabla más abajo; las etiquetas se muestran en negrita, los nucleótidos selectivos se han subrayado) en producto de preamplificación de AFLP TaqI/MseI +1/+1 diluido 20 veces. Perfil de PCR: 30*(30 segundos a 94ºC + 60 segundos a 56ºC + 120 segundos a 72ºC). El producto de AFLP se purificó mediante la utilización del kit de purificación PCR QIAquick (QIAGEN) siguiendo el manual 07/2002 del QIAquick® Spin, página 18, y se midió la concentración con un espectrofotómetro NanoDrop® ND-1000. Se combinó un total de 1,25 µg de cada producto diferente de AFLP B73 +2/+2 y 1,25 µg de cada producto diferente de AFLP M017 +2/+2 y se resolvió en 30 µl de TE. Finalmente, se obtuvo una mezcla con una concentración de 333 ng/µl de producto de AFLP +2/+2.

Tabla

SEC ID	Cebador de PCR	Cebador -3'	Pimiento	Reacción de AFLP
[SEC ID 21]	05G360	**ACGT**GTAGACTGCGTACCGAAA	B73	1
[SEC ID 22]	05G368	**ACGT**GATGAGTCCTGAGTAACA	B73	1
[SEC ID 23]	05G362	**CGTA**GTAGACTGCGTACCGAAC	B73	2
[SEC ID 24]	05G370	**CGTA**GATGAGTCCTGAGTAACA	B73	2
[SEC ID 25]	05G364	**GTAC**GTAGACTGCGTACCGAAG	B73	3
[SEC ID 26]	05G372	**GTAC**GATGAGTCCTGAGTAACA	B73	3
[SEC ID 27]	05G366	**TACG**GTAGACTGCGTACCGAAT	B73	4
[SEC ID 28]	05G374	**TACG**GATGAGTCCTGAGTAACA	B73	4
[SEC ID 29]	05G361	**AGTC**GTAGACTGCGTACCGAAA	M017	5
[SEC ID 30]	05G369	**AGTC**GATGAGTCCTGAGTAACA	M017	5
[SEC ID 31]	05G363	**CATG**GTAGACTGCGTACCGAAC	M017	6
[SEC ID 32]	05G371	**CATG**GATGAGTCCTGAGTAACA	M017	6
[SEC ID 33]	05G365	**GAGC**GTAGACTGCGTACCGAAG	M017	7
[SEC ID 34]	05G373	**GAGC**GATGAGTCCTGAGTAACA	M017	7
[SEC ID 35]	05G367	**TGAT**GTAGACTGCGTACCGAAT	M017	8
[SEC ID 36]	05G375	**TGAT**GATGAGTCCTGAGTAACA	M017	8

[0108] Finalmente, se agruparon y se concentraron las 4 muestras P1 y las 4 muestras P2. Se obtuvo una cantidad total de 25 µl de producto de ADN y una concentración final de 400 ng/µl (total de 10 µg). Se proporcionan las evaluaciones de calidad de intermediarios en la figura 3.

SECUENCIACIÓN POR 454

[0109] Unas muestras de fragmento de AFLP de pimiento y de maíz tal como se ha descrito anteriormente en la presente memoria fueron procesadas por 454 Life Sciences tal como se ha descrito (Margulies *et al.*, Genome

sequencing in microfabricated high-density picolitre reactors, Nature 435(7057):376-80, publicado electrónicamente el 31 de julio de 2005).

PROCESAMIENTO DE LOS DATOS

Línea de procesamiento:

Datos de entrada

[0110] Se recibieron los datos crudos de secuencias para cada análisis:

- 200.000 a 400.000 lecturas
- puntuaciones de calidad de asignaciones de bases

Recorte y etiquetado

[0111] Estos datos de secuencias se analizaron para la presencia de sitios de reconocimiento de Keygene (KRS) al inicio y final de la lectura. Estas secuencias KRS consisten de secuencias de adaptador de AFLP y de marcaje de muestra y son específicas de una determinada combinación de cebadores de AFLP en una muestra determinada. Las secuencias de KRS fueron identificadas por BLAST y recortadas, y se restituyeron los sitios de restricción. Las lecturas se marcaron con una etiqueta para la identificación del origen de KRS. Las secuencias recortadas se seleccionaron a partir de la longitud (mínimo de 33 nt) para participar en el procesamiento posterior.

Agrupamiento y ensamblaje

[0112] Se llevó a cabo un análisis *MegaBlast* de todas las lecturas recortadas y seleccionadas según tamaño para obtener agrupaciones de secuencias homólogas. Consecutivamente se ensamblaron todos los grupos con *CAP3*, resultando en contigs ensamblados. Tras ambas etapas se habían identificado lecturas de secuencia única que no se correspondían con ninguna otra lectura. Estas lecturas se señalan como singletons.

La línea de procesamiento seguida para llevar a cabo las etapas descritas en la presente memoria se muestra en la figura 4A.

Exploración de polimorfismos y evaluación de la calidad

[0113] Los contigs resultantes del análisis de ensamblaje formaron la base para la detección de polimorfismos. Cada "apareamiento incorrecto" en la alineación de cada agrupación es un polimorfismo potencial. Se definieron criterios de selección para obtener una puntuación de calidad:

- número de lecturas en cada contig
- frecuencia de "alelos" en cada muestra
- aparición de secuencia de homopolímero
- aparición de polimorfismos contiguos

[0114] Los SNPs e indels con una puntuación de calidad superior al umbral se identificaron como polimorfismos putativos. Para la exploración de SSRs se utilizó la herramienta MISA (identificación de microsatélites) (http://pgrc.ipk-gatersleben.de/misa). Esta herramienta identifica los motivos dinucleótido, trinucleótido, tetranucleótido y motivos SSR del compuesto aplicando criterios predefinidos y resumen las apariciones de estos SSRs.

[0115] El procedimiento de exploración de polimorfismos y asignación de calidad se muestra en la figura 4B.

RESULTADOS

[0116] La Tabla a continuación resume los resultados del análisis combinado de secuencias obtenido a partir de 2 análisis de secuenciación de 454 para las muestras combinadas de pimiento y 2 análisis para las muestras combinadas de maíz.

	Pimiento	Maíz
Número total de lecturas	457.178	492.145
Número de lecturas recortadas	399.623	411.008
Número de singletons	105.253	313.280

Número de contigs	31.863	14.588
Número de lecturas en contigs	294.370	97.728
Número total de secuencias que contienen SSRs	611	202
Número de secuencias diferentes que contienen SSR	104	65
Número de motivos SSR diferentes (di, tri, tetra y compuesto)	49	40
Número de SNPs con puntuación Q ≥ 0,3*	1.636	782
Número de indels*	4.090	943
* ambos seleccionando contra SNPs contiguos, por lo menos 12 pb de secuencia flanqueante y no presentes en secuencias de homopolímero mayores de 3 nucleótidos.		

Ejemplo 4. Identificación de polimorfismos de un nucleótido (SNP) en el pimiento.

Aislamiento del ADN

[0117] Se aisló el ADN genómico de las dos líneas parentales de una población recombinante consaguínea (RIL) de pimiento y 10 descendientes de RIL. Las líneas parentales eran PSP11 y PI201234. Se aisló el ADN genómico a partir de material foliar de plántulas individuales utilizando un procedimiento CTAB modificado descrito por Stuart y Via (Stuart C.N. Jr. y Via L.E., A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications, Biotechniques 14:748-750, 1993). Se diluyeron las muestras de ADN hasta una concentración de 100 ng/µl en TE (Tris-HCl 10 mM, pH 8,0, EDTA 1 mM) y se almacenaron a -20ºC.

Preparación de molde para AFLP utilizando cebadores de AFLP etiquetados

[0118] Se prepararon moldes para AFLP de las líneas parentales del pimiento PSP11 y PI201234 utilizando la combinación de endonucleasas de restricción EcoRI/MseI tal como describen Zabeau y Vos, Selective restriction fragment amplification; a general method for DNA fingerprinting, 1993, patente EP 0534858-A1, B1; patente US 6045994) y Vos *et al.* (Vos P., Hogers R., Bleeker M., Reijans M., van de Lee T., Holmes M. Frijters A., Pot J., Peleman J., Kuiper M. *et al.*, AFLP: a new technique for DNA fingerprinting, Nucl. Acids Res. 21:4407-4414, 1995).

[0119] Específicamente, se llevó a cabo la restricción del ADN genómico con EcoRI y MseI de la manera siguiente:

Restricción de ADN

ADN	100 a 500 ng
EcoRI	5 unidades
MseI	2 unidades
5x tampón RL	8 µl
Agua MilliQ	Hasta 40 µl

[0120] La incubación se realizó durante 1 hora a 37ºC. Tras la restricción enzimática, los enzimas se inactivaron mediante incubación durante 10 minutos a 80ºC.

Ligación de adaptadores

ATP 10 mM	1 µl
ADN ligasa de T4	1 µl
Adaptador EcoRI (5 pmoles/µl)	1 µl
Adaptador MseI (50 pmoles/µl)	1 µl

5x tampón RL	2 µl
Agua MilliQ hasta	40 µl

La incubación se realizó durante 3 horas a 37ºC

Amplificación selectiva mediante AFLP

[0121] Tras la restricción-ligación, la reacción de restricción/ligación se diluyó 10 veces con $T_{10}E_{0,1}$ y se utilizaron 5 µl de mezcla diluida como molde en una etapa de amplificación selectiva. Observar que debido a que se pretendía una amplificación selectiva +1/+2, en primer lugar se llevó a cabo una etapa de preamplificación selectiva +1/+1 (con cebadores de AFLP estándares). Las condiciones de reacción de la amplificación +1/+1 (+A/+C) fueron las siguientes.

Mezcla de restricción-ligación (diluida 10 veces)	5 µl
Cebador-EcoRI +1 (50 ng/µl)	0,6 µl
Cebador-MseI +1 (50 ng/µl)	0,6 µl
dNTPs (20 mM)	0,2 µl
Polimerasa Taq (5 U/µl Amplitaq, PE)	0,08 µl
10x tampón de PCR	2,0 µl
Agua MilliQ hasta	20 µl

[0122] Las secuencias de los cebadores eran:

EcoRI+1: 5'-AGACTGCGTACCAATTCA-3' [SEC ID 9] y

MseI+1: 5'-GATGAGTCCTGAGTAAC-3' [SEC ID 10]

[0123] Las amplificaciones mediante PCR se llevaron a cabo utilizando un PE9700 con un bloque de oro o plata utilizando las condiciones siguientes: 20 ciclos (30 segundos a 94ºC, 60 segundos a 56ºC y 120 segundos a 72ºC).

[0124] Se comprobó la calidad de los productos de preamplificación +1/+1 generados en un gel de agarosa al 1% utilizando una escalera de 100 pares de bases y una escalera de 1 Kb para comprobar la distribución de las longitudes de los fragmentos. Tras la amplificación selectiva +1/+1, la reacción se diluyó 20 veces con $T_{10}E_{0,1}$ y se utilizaron 5 µl de mezcla diluida como molde en la etapa de amplificación selectiva +1/+2 utilizando cebadores de AFLP etiquetados. Finalmente, se llevaron a cabo amplificaciones selectivas mediante AFLP +1/+2 (A/+CA):

producto de amplificación selectiva +1/+1 (diluido 20 veces): 5,0 µl.

KRS cebador-EcoRI +A (50 ng/µl)	1,5 µl
KRS cebador-MseI + CA (50 ng/µl)	1,5 µl
dNTPs (20 mM)	0,5 µl
Polimerasa Taq (5 U/µl Amplitaq, Perkin Elmer)	0,2 µl
10X tampón para PCR	5,0 µl
MQ hasta	50 µl

[0125] Las secuencias de los cebadores de AFLP etiquetados eran:

PSP11:

05F212: EcoRI+1: 5'-CGTCAGACTGCGTACCAATTCA-3' [SEC ID 1] y

05F214: MseI+2: 5'-TGGTGATGAGTCCTGAGTAACA-3' [SEC ID 2]

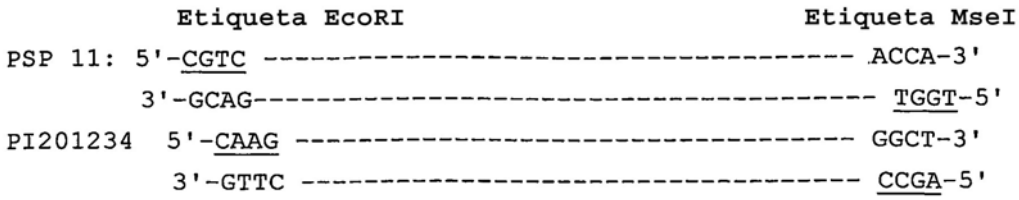
PI201234:

05F213: EcoRI+1: 5'-CAAGAGACTGCGTACCAATTCA-3' [SEC ID 3] y

05F215: MseI+1: 5'-AGGCGATGAGTCCTGAGTAACA-3' [SEC ID 4]

[0126] Observar que dichos cebadores contienen etiquetas de 4 pb (subrayadas anteriormente) en sus extremos 5 prima para distinguir los productos de amplificación originados de las líneas de pimiento respectivas al final del procedimiento de secuenciación.

[0127] Representación esquemática de los productos de amplificación del pimiento de AFLP +1/+2 tras la amplificación con cebadores de AFLP que contenían secuencias de etiqueta 5 prima de 4 pb.



[0128] Se llevaron a cabo amplificaciones de PCR (24 por muestra) utilizando un PE7900 con un bloque de oro o de plata bajo las condiciones siguientes: 30 ciclos (30 segundos a 94ºC + 60 segundos a 56ºC + 120 segundos a 72ºC).

[0129] Se comprobó la calidad de los productos de amplificación generados en un gel de agarosa al 1% utilizando una escalera de 100 pares de bases y una escalera de 1 Kb para comprobar la distribución de las longitudes de los fragmentos.

Purificación y cuantificación de la reacción de AFLP

[0130] Tras agrupar dos reacciones de AFLP selectiva +1/+2 de 50 microlitros por cada muestra de pimiento, los 12 productos resultantes de 100 µl de reacción de AFLP se purificaron utilizando el kit de purificación por PCR QIAquick (QIAGEN) siguiendo el manual de QIAquick® Spin (página 18). En cada columna se cargó un máximo de 100 µl de producto. Los productos amplificados se eluyeron en $T_{10}E_{0,1}$. Se comprobó la calidad de los productos purificados en un gel de agarosa al 1% y se midieron las concentraciones en el NanoDrop (figura 2).

[0131] Se utilizaron las mediciones de concentración en NanoDrop para ajustar la concentración final de cada producto de PCR purificado a 300 nanogramos por microlitro. Se mezclaron cinco microgramos de producto amplificado purificado de PSP11 y 5 microgramos de PI201234 para generar 10 microgramos de material de molde para la preparación de la biblioteca de secuenciación de 454.

Preparación de biblioteca de secuencias y secuenciación de alto rendimiento

[0132] Los productos de amplificación mixtos procedentes de ambas líneas de pimiento se sometieron a secuenciación de alto rendimiento utilizando la tecnología de secuenciación de 454 Life Sciences, tal como se describe en Margulies *et al.* (Margulies *et al.*, Nature 437:376-380 y Online Supplements). Específicamente, en primer lugar se pulieron los extremos de los productos de PCR AFLP y después se ligaron con adaptadores para facilitar la amplificación por PCR en emulsión y posterior secuenciación de los fragmentos tal como describen Margulies y colaboradores. Las secuencias de adaptadores de 454, los cebadores de PCR en emulsión, los cebadores de secuencia y las condiciones operativas de la secuenciación fueron todas las indicadas por Margulies y colaboradores. El orden lineal de elementos funcionales en un fragmento de PCR en emulsión amplificado sobre perlas de sefarosa en el procedimiento de secuenciación de 454 fue el siguiente, tal como se ejemplifica en la figura 1A:

[0133] Adaptador de PCR de 454 - adaptador de secuencia de 454 - etiqueta 1 de 4 pb de cebador de AFLP - secuencia 1 de cebador de AFLP que incluye el nucleótido o nucleótidos selectivos - secuencia interna de fragmento de AFLP - secuencia 2 de cebador de AFLP que incluye uno o más nucleótidos selectivos, etiqueta 2 de 4 pb de cebadores de AFLP - adaptador de secuencia de 454 - adaptador de PCR de 454 - perla de sefarosa.

[0134] Dos análisis de secuenciación de alto rendimiento de 454 fueron llevados a cabo por 454 Life Sciences (Branford, CT; Estados Unidos).

Procesamiento de datos de operación de secuenciación de 454

[0135] Los datos de secuencia resultantes de 2 análisis de secuenciación de 454 se procesaron utilizando un procedimiento bioinformático (Keygene N.V.). Específicamente, se convirtieron en formato FASTA lecturas de secuencia no procesadas con asignación de bases obtenidas de 454 y se inspeccionaron para la presencia de secuencias adaptadoras de AFLP etiquetadas utilizando un algoritmo de BLAST. Tras las correspondencias de alta confianza con las secuencias de cebador de AFLP etiquetadas conocidas, las secuencias se recortaron, se

restituyeron los sitios de endonucleasa de restricción y se asignaron las etiquetas apropiadas (muestra 1 EcoRI (ES1), muestra 1 MseI (MS1), muestra 2 EcoRI (ES2) o muestra 2 MseI (MS2), respectivamente). A continuación, todas las secuencias recortadas mayores de 33 bases se agruparon utilizando un procedimiento megaBLAST basado en homologías globales de secuencia. A continuación, se ensamblaron las agrupaciones en uno o más contigs y/o singletons por agrupación utilizando un algoritmo CAP3 de alineación múltiple. Los contigs que contenían más de una secuencia se inspeccionaron para apareamientos incorrectos de secuencias, representativos de polimorfismos putativos. Se asignaron puntuaciones de calidad a los apareamientos incorrectos de secuencia basándose en los criterios siguientes:

* número de lecturas en un contig

* la distribución observada de alelos

[0136] Los dos criterios anteriormente indicados forman la base para la denominada puntuación Q asignada a cada SNP/indel putativo. Las puntuaciones Q se encuentran comprendidas entre 0 y 1; una puntuación Q de 0,3 sólo puede alcanzarse en el caso de que ambos alelos se observen por lo menos dos veces.

* localización en homopolímeros de una determinada longitud (ajustable; valor por defecto para evitar polimorfismos localizados en homopolímeros de 3 bases o más largos).

* número de contigs en una agrupación.

* distancia a los apareamientos incorrectos de secuencia contigua más próximos (ajustable; importante para determinados tipos de ensayos de genotipado de sondeo de secuencias flanqueantes)

* el nivel de asociación de los alelos observados con la muestra 1 o con la muestra 2; en el caso de una asociación perfecta consistente entre los alelos de un polimorfismo putativo y las muestras 1 y 2, el polimorfismo (SNP) se indica en forma de putativo polimorfismo (SNP) "de elite". Se considera que un polimorfismo de élite presenta una elevada probabilidad de encontrarse localizado en una secuencia genómica única o de bajo número de copia, en el caso de que se hayan utilizado dos líneas homocigóticas en el procedimiento de exploración. A la inversa, una asociación débil de un polimorfismo con el origen de la muestra presenta un riesgo elevado de que se hayan descubierto polimorfismos falsos surgidos de la alineación de secuencias no alélicas en un contig.

[0137] Las secuencias que contienen motivos de SSR se identificaron utilizando la herramienta de búsqueda MISA (herramienta de identificación de microsatélites; disponible en http://pgrc.ipk-gatersleben.de/misa/).

[0138] Se muestran las estadísticas globales de la operación en la Tabla a continuación.

Tabla. Estadísticas globales de un análisis de secuenciación de 454

para la identificación de SNPs en el pimiento.

Combinación de enzimas	Operación
Recorte	
Todas las lecturas	254.308
Incorrectos	5.293 (2%)
Correctos	249.015 (98%)
Concatámeros	2.156 (8,5%)
Etiquetas mixtas	1.120 (0,4%)
Lecturas correctas	
Un extremo recortado	240.817 (97%)
Ambos extremos recortados	8.198 (3%)
Número de lecturas muestra 1	136.990 (55%)
Número de lecturas muestra 2	112.025 (45%)

Agrupaciones	
Número de contigs	21.918
Lecturas en contigs	190.861
Número medio de lecturas por contig	8,7
Exploración para SNPs	
Recorte	
SNPs con puntuación Q ≥ 0,3*	1.483
Indels con puntuación Q ≥ 0,3*	3.300
Exploración de SSRs	
Número total de motivos SSR identificados	359
Número de lecturas que contienen uno o más motivos SSR	353
Número de motivos SSR con tamaño unitario 1	0
(homopolímero)	
Número de motivos SSR con tamaño unitario 2	102
Número de motivos SSR con tamaño unitario 3	240
Número de motivos SSR con tamaño unitario 4	17
*Los criterios de exploración de SNP/indels fueron los siguientes:	

[0139] No se encontraron polimorfismos contiguos con una puntuación Q superior a 0,1 a menos de 12 bases en cada lado, en homopolímeros de 3 ó más bases. Los criterios de exploración no consideraron la asociación consistente con las muestras 1 y 2, es decir, los SNPs e indels no son necesariamente putativos SNPs/indels de élite.

[0140] En la figura 7 se muestra un ejemplo de una alineación múltiple que contiene un polimorfismo putativo de único nucleótido de élite.

Ejemplo 5. Validación de SNPs mediante amplificación por PCR y secuenciación de Sanger

[0141] Con el fin de validar el SNP A/G putativo que se identifica en el Ejemplo 1, se diseñó un ensayo de sitio de secuencia etiquetada (STS) para este SNP utilizando cebadores de PCR flanqueantes. Las secuencias de los cebadores de PCR eran las siguientes:

Cebador_1.2f: 5'-AAACCCAAACTCCCCCAATC-3' [SEC ID 37] y

Cebador_1.2r:5'- AGCGGATAACAATTTCACACAGGACATCAGTAGTCACACTGGTA CAAAAATAGAGCAAAACAGTAGTG -3' [SEC ID 38]

[0142] Observar que el cebador 1.2r contenía un sitio de unión de cebador de secuencia de M13 y un fragmento de relleno en su extremo 5 prima. Se llevó a cabo la amplificación por PCR utilizando los productos de amplificación por AFLP +A/+CA de PSP11 y PI210234 preparados tal como se describe en el Ejemplo 4 a modo de molde. Las condiciones de PCR fueron las siguientes:

Para 1 reacción de PCR se mezclaron los componentes siguientes:

5 µl de mezcla para AFLP diluida 1/10 (aprox. 10 ng/µl)

5 µl Ipmol/µl de cebador 1.2f (diluido directamente a partir de una solución madre 500 µM)

5 µl Ipmol/µl de cebador 1.2r (diluido directamente a partir de una solución madre 500 µM)

5 µl de mezcla para PCR	- 2 µl de 10 x tampón para PCR
	- 1 µl de dNTPs 5 mM
	- 1,5 µl de $MgCl_2$ 25 mM
	- 0,5 µl de H_2O
5 µl de mezcla de enzimas	- 0,5 µl de 10 x tampón para PCR (Applied Biosystems
	- 0,1 µl de ADN polimerasa AmpliTaq 5 U/µl (Applied Biosystems)
	- 4,4 µl de H_2O

Se utilizó el perfil de PCR siguiente:

Ciclo 1	2';	94ºC
Ciclos 2 a 34	20";	94ºC
	30";	56ºC
	2'30";	72ºC
Ciclo 35	7';	72ºC
	∞;	4ºC

[0143] Los productos de PCR se clonaron en el vector pCR2.1 (kit de clonación TA, Invitrogen) utilizando el método de clonación TA y se transformaron en células *E. coli* competentes INVαF'. Los transformantes se sometieron a cribado azul/blanco. Se seleccionaron tres transformantes blancos independientes de cada uno de PSP11 y PI-201234 y se cultivaron O/N en medio selectivo líquido para el aislamiento de plásmidos.

[0144] Los plásmidos se aislaron utilizando el kit miniprep QIAprep Spin (QIAGEN). A continuación, se secuenciaron las inserciones de estos plásmidos siguiendo el protocolo indicado posteriormente y se resolvieron en el MegaBACE 1000 (Amersham). Las secuencias obtenidas se inspeccionaron en presencia del alelo SNP. Dos plásmidos independientes que contenían la inserción PI-201234 y 1 plásmido que contenía la inserción PSP11 contenían la secuencia de consenso esperada flanqueantes del SNP. La secuencia derivada del fragmento de PSP11 contenía el alelo A esperado (subrayado) y la secuencia derivada del fragmento PI-201234 contenía el alelo G esperado (doble subrayado):

PSP11 (secuencia 1): (5'-3')

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA
CCCCACTACTGTTTTGCTCTATTTTTGT **[SEC ID 39]**

PI-201234 (secuencia 1): (5'-3')

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAGTGTTGGTTTTGGTGCTAACTTCAA
CCCCACTACTGTTTTGCTCTATTTTTG **[SEC ID 40]**

PI-201234 (secuencia 2): (5'-3')

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAGTGTTGGTTTTGGTGCTAACTTCAA
CCCCACTACTGTTTTGCTCTATTTTTG **[SEC ID 41]**

[0145] Este resultado indica que el putativo SNP A/G del pimiento representa un polimorfismo genético verdadero detectable utilizando el ensayo STS diseñado.

Ejemplo 6: validación de SNPs mediante detección con SNPWave

[0146] Con el fin de validar el putativo SNP A/G identificado en el Ejemplo 1, se definieron conjuntos de sondas de ligación SNPWave para ambos alelos de dicho SNP utilizando la secuencia de consenso. Las secuencias de las sondas de ligación eran las siguientes:

secuencias de sonda SNPWave (5'-3'):

06A162 GATGAGTCCTGAGTAACCCAATCGATTTCAAACCTAGAACAA (42 bases) [SEC ID 42]

06A163 GATGAGTCCTGAGTAACCACCAATCGATTTCAAACCTAGAACAG (44 bases) [SEQ ID 43]

06A164 Fosfato-TGTTGGTTTTGGTGCTAACTTCAACCAACATCTGGAATTGGTACGCAGTC (52 bases) [SEC ID 44]

[0147] Observar que las sondas específicas de alelo 06A162 y 06A163 para los alelos A y G, respectivamente, difieren en tamaño en 2 bases, de manera que, tras la ligación a la sonda común específica de locus 06A164, resultan tamaños de producto de ligación de 94 (42+54) y 96 (44+52) bases.

[0148] Las reacciones de ligación SNPWave y de PCR se llevaron a cabo tal como describen Van Eijk y colaboradores (M.J.T van Eijk, J.L.N. Broekhof, H.J.A. van der Poel, R.C.J. Hogers, H. Schneiders, J. Kamerbeek, E. Verstege, J.W. van Aart, H. Geerlings, J.B. Buntjer, A.J. van Oeveren y P. Vos, SNPWaveTM: a flexible multiplexed SNP genotyping technology, Nucleic Acids Research 32:e47, 2004), utilizando 100 ng de ADN genómico de las líneas de pimiento PSP11 y PI201234 y 8 descendientes RIL como materiales de partida. Las secuencias de los cebadores de PCR eran:

93L01FAM (E00k): 5-GACTGCGTACCAATTC-3' [SEC ID 45]

93E40 (M00k): 5-GATGAGTCCTGAGTAA-3' [SEC ID 46]

[0149] Tras la amplificación por PCR, la purificación y detección del producto de PCR en el MegaBACE1000 fue tal como ha sido descrita por van Eijk y colaboradores (ver anteriormente). En la figura 8B se muestra una pseudoimagen en gel de los productos de amplificación obtenidos de PSP11, PI201234 y de 8 descendientes RIL.

[0150] Los resultados del SNPWave demuestran claramente que el SNP A/G se detecta mediante el ensayo SNPWave, resultando en productos de 92 pb (=genotipo homocigótico AA) para P1 (PSP11) y los descendientes RIL 1, 2, 3, 4, 6 y 7) y en productos de 94 pb (genotipo homocigótico GG) para P2 (PI201233) y los descendientes RIL 5 y 8.

Ejemplo 7. Estrategias para el enriquecimiento de bibliotecas de fragmentos de AFLP en secuencias de bajo número de copia.

[0151] El presente ejemplo describe varios métodos de enriquecimiento centrados en secuencias genómicas únicas o de bajo número de copia con el fin de incrementar el rendimiento de polimorfismos de elite tal como se describe en el Ejemplo 4. Los métodos pueden clasificarse en cuatro categorías:

1) Métodos destinados a preparar ADN genómico de alta calidad, excluyendo secuencias de cloroplastos

[0152] Se propone preparar ADN nuclear en lugar de ADN genómico total tal como se describe en el Ejemplo 4, para excluir el coaislamiento de abundante ADN de cloroplastos, lo que podría resultar en un número reducido de secuencias de ADN genómico de la planta, dependiendo de las endonucleasas de restricción y los cebadores de AFLP selectiva utilizados durante el procedimiento de preparación de la biblioteca de fragmentos. Se ha descrito un protocolo de aislamiento de ADN nuclear altamente puro del tomate en Peterson D.G., Boehm K.S. y Stack S.M., Isolation of Milligram Quantities of Nuclear DNA From Tomato (*Lycopersicon esculentum*), A Plant Containing High Levels of Polyphenolic Compounds, Plant Molecular Biology Reporter 15(2):148-153, 1997.

2) Métodos destinados a utilizar endonucleasas de restricción durante el procedimiento de preparación de moldes para AFLP que se espera que rindan niveles elevados de secuencias de bajo número de copia.

[0153] Se propone utilizar determinadas endonucleasas de restricción durante el procedimiento de

preparación de moldes para AFLP, que se espera que presenten diana en secuencias genómicas de bajo número de copia o únicas, resultando en bibliotecas de fragmentos enriquecidas en polimorfismos con una capacidad incrementada de ser convertibles en ensayos de genotipado. Un ejemplo de una endonucleasa de restricción con diana en una secuencia de bajo número de copia en genomas vegetales es PstI. Otras endonucleasas de restricción sensibles a la metilación también pueden presentar diana preferentemente en secuencias genómicas de bajo número de copia o únicas.

3) Métodos destinados a eliminar selectivamente secuencias altamente duplicadas basándose en cinética de re-apareamiento de secuencias repetidas frente a secuencias de bajo número de copia.

[0154] Se propone eliminar selectivamente secuencias altamente duplicadas (repetidas) de la muestra de ADN genómico total o del material de molde de AFLP (ADNc) antes de la amplificación selectiva.

[0155] 3a) La preparación de ADN de C_0t elevado es una técnica utilizada comúnmente para enriquecer secuencias de bajo número de copia de apareamiento lento a partir de una mezcla compleja de ADN genómico vegetal (Yuan *et al.*, High-C_0t sequence analysis of the maize genome, Plant J. 34:249-255, 2003). Se sugiere utilizar ADN de C_0t elevado en lugar de ADN genómico total para el enriquecimiento en polimorfismos situados en secuencias de bajo número de copia.

[0156] 3b) Una alternativa a la laboriosa preparación de ADN de C_0t elevado podría ser la incubación de ADNdc desnaturalizado y re-hibridado con una nueva nucleasa, de cangrejo de Kamchatka, que corta dúplex de ADN cortos perfectamente apareados a una tasa más alta que los dúplex de ADN no apareados perfectamente, tal como describen Zhulidov y colaboradores (Simple cDNA normalization using Kamchatka crab duplex-specific nucleasa, Nucleic Acids Research 32:e37, 2004) y Shagin y colaboradores (A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas, Genome Research 12:1935-1942, 2006). Específicamente, se propone incubar las mezclas de restricción/ligación de AFLP con dicha endonucleasa para empobrecer la mezcla en secuencias altamente duplicadas, seguido de la amplificación mediante AFLP selectiva de las secuencias genómicas remanentes de bajo número de copia o únicas.

[0157] 3c) La filtración de metilos es un método para enriquecer en fragmentos de ADN genómico hipometilado utilizando la endonucleasa de restricción McrBC, que corta el ADN metilado en la secuencia [A/G]C, en la que C se encuentra metilado (ver Pablo D. Rabinowicz, Robert Citek, Muhammad A. Budiman, Andrew Nunberg, Joseph A. Bedell, Nathan Lakey, Andrew L. O'Shaughnessy, Lidia U. Nascimento, W. Richard McCombie y Robert A. Martienssen, Differential methylation of genes and repeats in land plants, Genome Research 15:1431-1440, 2005). Puede utilizarse la McrBC para enriquecer en la fracción de secuencias de bajo número de copia de un genoma, que se utilizará como material de partida para la exploración para polimorfismos.

4) Utilización de ADNc y no ADN genómico para el reconocimiento de secuencias génicas

[0158] Finalmente, se propone utilizar ADNc cebado con oligo-dT y no ADN genómico como material de partida para la exploración de polimorfismos, opcionalmente en combinación con la utilización de nucleasa específica de dúplex de cangrejo indicada en 3b, anteriormente, para la normalización. Observar que la utilización de ADNc cebado con oligo-dT también excluye las secuencias de cloroplastos. Alternativamente, se utilizan moldes de ADNc-AFLP en lugar de ADNc cebado con oligo-dT para facilitar la amplificación de las secuencias remanentes de bajo número de copia análogamente a AFLP (ver también 3b, anteriormente).

Ejemplo 8. Estrategia para el enriquecimiento en repeticiones de secuencias simples.

[0159] El presente ejemplo describe la estrategia propuesta de descubrimiento de secuencias repetidas de secuencias simples análogamente a la identificación de SNPs descrita en el Ejemplo 4.

[0160] Específicamente, se lleva a cabo la restricción-ligación de ADN genómico de dos o más muestras, por ejemplo utilizando las endonucleasas de restricción PstI/MseI. La amplificación mediante AFLP selectiva se lleva a cabo tal como se describe en el Ejemplo 4. A continuación, se enriquece en fragmentos que contienen los motivos SSR seleccionados mediante uno de los dos métodos siguientes:

1) hibridación de transferencia southern sobre filtros que contienen oligonucleótidos correspondientes a los motivos SSR deseados (por ejemplo $(CA)_{15}$ en el caso del enriquecimiento para repeticiones CA/GT), seguido de la amplificación de los fragmentos unidos de un modo similar al descrito por Armour y colaboradores (Armour J., Sismani C., Patsalis P. y Cross G., Measurement of locus copy number by hybridization with amplifiables probes, Nucleic Acids Research 28(2):605-609, 2000), o mediante:

2) enriquecimiento utilizando sondas de hibridación de oligonucleótidos de captura biotinilados para capturar fragmentos (AFLP) en solución tal como describen Kijas y colaboradores (Kijas J.M., Fowler J.C., Garbett C.A. y Thomas M.R., Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles, Biotechniques 16:656-662, 1994).

[0161] A continuación, los fragmentos de AFLP enriquecidos en motivo SSR se amplifican utilizando los

mismos cebadores de AFLP utilizados en la etapa de preamplificación, con el fin de generar una biblioteca de secuencias. Una alícuota de los fragmentos amplificados se clonan T/A y 96 clones se secuencian para estimar la fracción de clones positivos (clones que contienen el motivo SSR deseado, por ejemplo motivos CA/GT de más de 5 unidades repetidas. Se detecta otra alícuota de la mezcla enriquecida en fragmentos de AFLP mediante electroforesis en gel de poliacrilamida (PAGE), opcionalmente tras la amplificación selectiva adicional para obtener una huella genética legible, con el fin de inspeccionar visualmente si se ha realizado el enriquecimiento en fragmentos que contienen SSR. Tras completar con éxito dichas etapas de control, las bibliotecas de secuencias se someten a secuenciación de alto rendimiento de 454.

[0162] La estrategia anteriormente indicada para la identificación *de novo* de SSRs se ilustra esquemáticamente en la figura 8A, y puede adaptarse para otros motivos de secuencia mediante la sustitución correspondiente de las secuencias oligonucleótidas de captura.

Ejemplo 9. Estrategia para evitar las etiquetas mixtas.

[0163] Las etiquetas mixtas se refieren a la observación de que, aparte de la combinación de cebadores de AFLP etiquetados esperada en cada muestra, se observa una fracción reducida de secuencias que contiene una etiqueta de muestra 1 en un extremo y una etiqueta de muestra 2 en el otro extremo (ver también la Tabla 1 en el Ejemplo 4). Esquemáticamente se ilustra la configuración de las secuencias que contienen etiquetas mixtas a continuación.

[0164] Representación esquemática de las combinaciones esperadas de etiquetas de muestra:

```
          EcoRI tag                                          MseI tag
PSP 11:   5'-CGTC ------------------------------------------ ACCA-3'
          3'-GCAG------------------------------------------ TGGT-5'

PI-201234 5'-CAAG ------------------------------------- GGCT-3'
          3'-GTTC ------------------------------------- CCGA-5'
```

[0165] Representación esquemática de las etiquetas mixtas.

```
etiqueta EcoRI                                    etiqueta MseI
5'-CGTC ------------------------------------------ GGCT-3'
3'-GCAG------------------------------------------ CCGA-5'

5'-CAAG ------------------------------------------ ACCA-3'
3'-GTTC ------------------------------------------  TGGT-5'
```

[0166] La observación de etiquetas mixtas impide la asignación correcta de las secuencias a PSP11 ó PI-201234.

[0167] En la figura 5A se muestra un ejemplo de una secuencia de etiqueta mixta en la operación de secuenciación del pimiento descrita en el Ejemplo 4. En el panel 2 de la figura 5A se muestra una vista general de la configuración de los fragmentos observados que contienen etiquetas esperadas y etiquetas mixtas.

[0168] La explicación molecular propuesta para las etiquetas mixtas es que durante la etapa de preparación de la biblioteca de secuencias, la ADN polimerasa de T4 o el enzima Klenow generan extremos romos en los fragmentos de ADN al eliminar los extremos protuberantes 3 prima antes de la ligación de adaptadores (Margulies *et al.*, 2005). Aunque lo anterior puede funcionar bien en el caso de que se procese una única muestra de ADN, al procesar una mezcla de dos o más muestras de ADN etiquetadas diferentemente, el rellenado por parte de la polimerasa resulta en la incorporación de una secuencia de etiqueta incorrecta en el caso de que se haya formado un heterodúplex entre las cadenas complementarias derivadas de muestras diferentes (figura 5B, panel 3, etiquetas mixtas). La solución encontrada es agrupar las muestras tras la etapa de purificación posterior a la ligación de adaptadores durante la etapa de construcción de la biblioteca de fragmentos de 454, tal como se muestra en la figura 5C, panel 4.

Ejemplo 10. Estrategia para evitar etiquetas mixtas y concatámeros utilizando un diseño mejorado de preparación de biblioteca de secuencias de 454

[0169] Aparte de la observación de frecuencias bajas de lecturas de secuencia que contienen etiquetas

mixtas tal como se describe en el Ejemplo 9, se ha observado una frecuencia baja de lecturas de secuencia de fragmentos de AFLP concatenados.

[0170] Un ejemplo de una lectura de secuencia derivada de un concatámero se muestra en la figura 6A, panel 1. Esquemáticamente, se muestra en la figura 6A, panel 2, la configuración de secuencias que contienen etiquetas esperadas y concatámeros.

[0171] La explicación molecular propuesta para la presencia de fragmentos de AFLP concatenados es que, durante la etapa de preparación de la biblioteca de secuencias de 454, se generan extremos romos en los fragmentos de ADN al eliminar la ADN polimerasa de T4 o el enzima Klenow los extremos protuberantes 3 prima antes de la ligación de adaptadores (Margulies *et al.*, 2005). En consecuencia, los fragmentos de ADN de extremos romos de la muestra compiten con los adaptadores durante la etapa de ligación y podrían ligarse entre sí antes de ligarse a los adaptadores. Este fenómeno de hecho es independiente de si se incluye una única muestra de ADN o una mezcla de múltiples muestras (etiquetadas) en la etapa de preparación de la biblioteca, y por lo tanto también podría producirse durante la secuenciación convencional tal como describen Margulies y colaboradores. En el caso de que se utilicen muestras con múltiples etiquetas, tal como se describe en el Ejemplo 4, los concatámeros complican la asignación correcta de lecturas de secuencia a las muestras basada en la información de etiquetas y por lo tanto deben evitarse.

[0172] La solución propuesta a la formación de concatámeros (y de etiquetas mixtas) es sustituir la ligación de adaptadores de extremos romos con la ligación de adaptadores que contienen un extremo protuberante 3 prima T, análogamente a la clonación T/A de productos de PCR, tal como se muestra en la figura 6B, panel 3. Convenientemente, se propone que estos adaptadores modificados que contienen una T en el extremo protuberante 3' contengan una C en el extremo protuberante 3' opuesto (que no se ligará al fragmento de ADN de muestra, para evitar la formación de concatámeros entre extremos romos de secuencias adaptadoras (ver la figura 6B, panel 3). El flujo de operaciones adaptado que resulta para el procedimiento de construcción de una biblioteca de secuencias al utilizar el enfoque de adaptadores modificados se muestra esquemáticamente en la figura 6C, panel 4.

LISTADO DE SECUENCIAS

[0173]

<110> Keygene NV

<120> Estrategias para la identificación y detección de alto rendimiento de los polimorfismos

<130> P27819PC00

<160> 46

<170> PatentIn versión 3.3

<210> 1

<211> 22

<212> ADN

<213> Artificial

<220>

<223> Cebador

<400> 1

gtcagactg cgtaccaatt ca 22

<210> 2

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 2

ggtgatgag tcctgagtaa ca 22

<210> 3

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 3

aagagactg cgtaccaatt ca 22

<210> 4

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 4

gccgatgag tcctgagtaa ca 22

<210> 5

<211> 17

<212> ADN

<213> Artificial

<220>

<223> adaptador

<400> 5

tcgtagact gcgtacc 17

<210> 6

<211> 18

<212> ADN

<213> Artificial

<220>

<223> adaptador

<400> 6

attggtacg cagtctac 18

<210> 7

<211> 16

<212> ADN

<213> Artificial

<220>

<223> adaptador

<400> 7

acgatgagt cctgag 16

<210> 8

<211> 14

<212> ADN

<213> Artificial

<220>

<223> adaptador

<400> 8

actcaggac tcat 14

<210> 9

<211> 18

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 9

gactgcgta ccaattca 18

<210> 10

<211> 17

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 10

atgagtcct gagtaac 17

<210> 11

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 11

gtcagactg cgtaccaatt ca 22

<210> 12

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 12

aagagactg cgtaccaatt ca 22

<210> 13

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 13

ggtgatgag tcctgagtaa ca 22

<210> 14

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 14

gccgatgag tcctgagtaa ca 22

<210> 15

<211> 19

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 15

actgcgtac caattcaac 19

<210> 16

<211> 19

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 16

atgagtcct gagtaacag 19

<210> 17

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 17

gtcagactg cgtaccaatt ca 22

<210> 18

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 18

ggtgatgag tcctgagtaa ca 22

<210> 19

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 19

aagagactg cgtaccaatt ca 22

<210> 20

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 20

aagagactg cgtaccaatt ca 22

<210> 21

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 21

cgtgtagac tgcgtaccga aa 22

<210> 22

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 22

cgtgatgag tcctgagtaa ca 22

<210> 23

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 23

gtagtagac tgcgtaccga ac 22

<210> 24

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 24

gtagatgag tcctgagtaa ca 22

<210> 25

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 25

tacgtagac tgcgtaccga ag 22

<210> 26

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 26

tacgatgag tcctgagtaa ca 22

<210> 27

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 27

acggtagac tgcgtaccga at 22

<210> 28

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 28

acggatgag tcctgagtaa ca 22

<210> 29

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 29

gtcgtagac tgcgtaccga aa 22

<210> 30

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 30

gtcgatgag tcctgagtaa ca 22

<210> 31

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 31

atggtagac tgcgtaccga ac 22

<210> 32

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 32

atggatgag tcctgagtaa ca 22

<210> 33

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 33

agcgtagac tgcgtaccga ag 22

<210> 34

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 34

agcgatgag tcctgagtaa ca 22

<210> 35

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 35

gatgtagac tgcgtaccga at 22

<210> 36

<211> 22

<212> ADN

<213> Artificial

<220>

<223> cebador

<400> 36

gatgatgag tcctgagtaa ca 22

<210> 37

<211> 20

<212> ADN

<213> artificial

<220>

<223> cebador

<400> 37

aacccaaac tcccccaatc 20

<210> 38

<211> 68

<212> ADN

<213> artificial

<220>

<223> cebador

<400> 38

gcggataac aatttcacac aggacatcag tagtcacact ggtacaaaaa tagagcaaaa 60

agtagtg 68

<210> 39

<211> 91

<212> ADN

<213> artificial

<220>

<223> sonda

<400> 39

aacccaaac tcccccaatc gatttcaaac ctagaacaat gttggttttg gtgctaactt 60

aaccccact actgttttgc tctatttttg t 91

<210> 40

<211> 90

<212> ADN

<213> artificial

<220>

<223> secuencia que contiene SNP PI-201234

<400> 40

aacccaaac tcccccaatc gatttcaaac ctagaacagt gttggttttg gtgctaactt 60

aaccccact actgttttgc tctatttttg 90

<210> 41

<211> 90

<212> ADN

<213> artificial

<220>

<400> 41

aacccaaac tcccccaatc gatttcaaac ctagaacagt gttggttttg gtgctaactt 60

aaccccact actgttttgc tctatttttg 90

<210> 42

<211> 42

<212> ADN

<213> artificial

<220>

<223> sonda SNPWave

<400> 42

atgagtcct gagtaaccca atcgatttca aacctagaac aa 42

<210> 43

<211> 44

<212> ADN

<213> artificial

<220>

<223> sonda SNPWave

<400> 43

atgagtcct gagtaaccac caatcgattt caaacctaga acag 44

<210> 44

<211> 50

<212> ADN

<213> artificial

<220>

<223> sonda snpwave

<400> 44

gttggtttt ggtgctaact tcaaccaaca tctggaattg gtacgcagtc 50

<210> 45

<211> 16

<212> ADN

<213> artificial

<220>

<223> cebador

<400> 45

actgcgtac caattc 16

<210> 46

<211> 16

<212> ADN

<213> artificial

<220>

<223> cebador

<400> 46

atgagtcct gagtaa 16

REIVINDICACIONES

1. Método para identificar uno o más polimorfismos, comprendiendo dicho método las etapas de:

a) proporcionar una primera muestra de ácidos nucleicos de interés;

b) llevar a cabo una reducción de complejidad de la primera muestra de ácidos nucleicos de interés, proporcionando una primera biblioteca de la primera muestra de ácidos nucleicos;

c) llevar a cabo consecutiva o simultáneamente las etapas a) y b) con una segunda o posterior muestra de ácidos nucleicos de interés, obteniendo una segunda o posterior biblioteca de la segunda o posterior muestra de ácidos nucleicos de interés;

d) secuenciar por lo menos una parte de la primer biblioteca y de la segunda o posterior biblioteca, en la que la secuenciación se lleva a cabo en un soporte sólido, tal como una perla;

e) alinear las secuencias obtenidas en la etapa d);

f) determinar uno o más polimorfismos entre la primera muestra de ácidos nucleicos y la segunda o posterior muestra de ácidos nucleicos en la alineación de la etapa e);

g) utilizar el polimorfismo o polimorfismos determinados en la etapa f) para diseñar sondas de detección;

h) proporcionar una muestra de ensayo de ácidos nucleicos de interés;

i) llevar a cabo la reducción de complejidad de la etapa b) en la muestra de ensayo de ácidos nucleicos de interés, proporcionando una biblioteca de ensayo de la muestra de ensayo de ácidos nucleicos;

j) someter la biblioteca de ensayo a cribado de alto rendimiento para identificar la presencia, ausencia o cantidad de polimorfismos determinados en la etapa f) utilizando las sondas de detección diseñadas en la etapa g), y en la que la reducción de complejidad de la etapa b) se lleva a cabo mediante:

- digestión de la muestra de ácidos nucleicos con por lo menos una endonucleasa de restricción para fragmentar en fragmentos de restricción;

- ligación de los fragmentos de restricción obtenidos con por lo menos un adaptador oligonucleótido sintético de doble cadena que presenta un extremo compatible con uno o ambos extremos de los fragmentos de restricción, produciendo fragmentos de restricción ligados con adaptadores;

- puesta en contacto de dichos fragmentos de restricción ligados con adaptadores con uno o más cebadores oligonucleótidos bajo condiciones de hibridación; y

- amplificación de dichos fragmentos de restricción ligados con adaptadores mediante alargamiento de uno o más cebadores oligonucleótidos,

- en el que por lo menos uno de entre el cebador o cebadores oligonucleótidos incluye una secuencia de nucleótidos que presenta la misma secuencia de nucleótidos que las partes terminales de las cadenas en los extremos de dichos fragmentos de restricción ligados con adaptadores, incluyendo los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción e incluyendo por lo menos parte de los nucleótidos presentes en los adaptadores, en el que, opcionalmente, por lo menos uno de dichos cebadores incluye en su extremo 3' una secuencia seleccionada que comprende por lo menos un nucleótido inmediatamente contiguo a los nucleótidos implicados en la formación de la secuencia diana de dicha endonucleasa de restricción.

2. Método según la reivindicación 1, en el que el adaptador y/o el cebador comprende una etiqueta.

3. Método según la reivindicación 2, en el que la etiqueta es una secuencia identificadora.

4. Método según la reivindicación 1, en el que por lo menos uno de los cebadores se encuentra fosforilado.

5. Método según cualquiera de las reivindicaciones anteriores, en el que la secuenciación se basa en la secuenciación mediante terminación de cadena dideoxi.

6. Método según la reivindicación 1, en el que la secuenciación comprende las etapas de:

- unir los fragmentos ligados con adaptadores a perlas, uniendo cada perla con un único fragmento ligado con adaptador;

- emulsionar las perlas en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una única perla;

- cargar las perlas en pocillos, comprendiendo cada pocillo una única perla; y

- generar una señal pirofosfato.

7. Método según la reivindicación 6, en el que, antes de la etapa de unión, se ligan adaptadores de secuenciación con los fragmentos dentro de la primera biblioteca etiquetada y de la segunda biblioteca etiquetada o la biblioteca combinada.

8. Método según la reivindicación 7, en el que los adaptadores de secuenciación portan un extremo protuberante 3'-T.

9. Método según cualquiera de las reivindicaciones anteriores, en el que el cribado de alto rendimiento se lleva a cabo mediante inmovilización de las sondas diseñadas en la etapa h) sobre una matriz, seguido de la puesta en contacto de la matriz que comprende las sondas con una biblioteca de ensayo bajo condiciones de hibridación.

10. Método para identificar uno o más polimorfismos, comprendiendo dicho método las etapas de:

a) proporcionar una pluralidad de muestras de ácidos nucleicos de interés,

b) llevar a cabo una reducción de complejidad de cada una de las muestras para proporcionar una pluralidad de bibliotecas de las muestras de ácidos nucleicos, en la que la reducción de complejidad se lleva a cabo mediante:

- digestión de cada muestra de ácidos nucleicos con por lo menos una endonucleasa de restricción para fragmentarla en fragmentos de restricción;

- ligación de los fragmentos de restricción obtenidos con por lo menos un adaptador oligonucleótido sintético de doble cadena que presenta un extremo compatible con uno o con ambos extremos de los fragmentos de restricción, produciendo fragmentos de restricción ligados con adaptadores;

- puesta en contacto de dichos fragmentos de restricción ligados con adaptadores, con uno o más cebadores oligonucleótidos fosforilados bajo condiciones de hibridación; y

- amplificación de dichos fragmentos de restricción ligados con adaptadores mediante alargamiento de uno o más de los cebadores oligonucleótidos, en la que por lo menos uno de entre el cebador o cebadores oligonucleótidos incluye una secuencia de nucleótidos que presenta la misma secuencia de nucleótidos que las partes terminales de las cadenas en los extremos de dichos fragmentos de restricción ligados con adaptadores, incluyendo los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción, e incluyendo por lo menos parte de los nucleótidos presentes en los adaptadores, en el que, opcionalmente, por lo menos uno de dichos cebadores incluye en su extremo 3' una secuencia seleccionada que comprende por lo menos un nucleótido situado inmediatamente contiguo a los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción y en el que el adaptador y/o el cebador contiene una etiqueta;

c) combinación de dichas bibliotecas para formar una biblioteca combinada;

d) ligar los adaptadores de secuenciación capaces de unirse a perlas, con los fragmentos con caperuza de adaptador amplificados en la biblioteca combinada, utilizando un adaptador de secuenciación que porta un extremo protuberante 3'-T, y someter los fragmentos unidos a perla a polimerización en emulsión;

e) secuenciar por lo menos una parte de la biblioteca combinada;

f) alinear las secuencias de cada muestra obtenidas en la etapa e);

g) determinar uno o más polimorfismos entre la pluralidad de muestras de ácidos nucleicos en la alineación de la etapa f);

h) utilizar uno o más polimorfismos determinados en la etapa g) para diseñar sondas de detección;

i) proporcionar una muestra de ensayo de ácidos nucleicos de interés;

j) llevar a cabo la reducción de complejidad de la etapa b) en la muestra de ensayo de ácidos nucleicos de interés, proporcionando una biblioteca de ensayo de la muestra de ensayo de ácidos nucleicos;

k) someter la biblioteca de ensayo a cribado de alto rendimiento para identificar la presencia, ausencia o cantidad de polimorfismos determinada en la etapa g) utilizando las sondas de detección diseñadas en la etapa h).

11. Utilización de los métodos según las reivindicaciones 1 a 6 para el cribado de bibliotecas de microsatélites enriquecidas, la realización de AFLP-ADNc de perfilado de transcritos (northern digital), la secuenciación de genomas complejos, la secuenciación de bibliotecas de etiquetas de secuencia expresadas (en ADNc completo o en ADNc-AFLP), la exploración de microARN (secuenciación de bibliotecas de inserciones de pequeño tamaño), la secuenciación de cromosomas artificiales bacterianos (BACs), el enfoque de análisis de segregantes agrupados en combinación con AFLP/AFLP-ADNc y la detección rutinaria de fragmentos de la AFLP (retrocruzamientos asistidos por un marcador).

Figura 1

Conjunto I de cebadores para la preamplificación de PSP-11

E01LKRS1 5'-CGTCAGACTGCGTACCAATTCA-3'

M15KKRS1 5'-TGGTGATGAGTCCTGAGTAACA-3'

Conjunto II de cebadores para la preamplificación de PI20234

E01LKRS2 5'-CAAGAGACTGCGTACCAATTCA-3'

M15KKRS2 5'-AGCCGATGAGTCCTGAGTAACA-3'

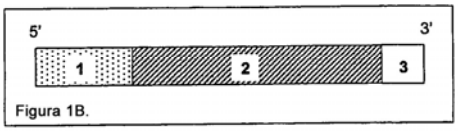


Figura 1B.

Figura 2

Control de calidad del ADN en un gel de agarosa al 1%

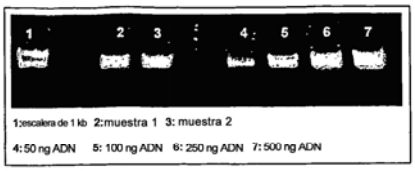


Figura 2A. Electroforesis en gel corta

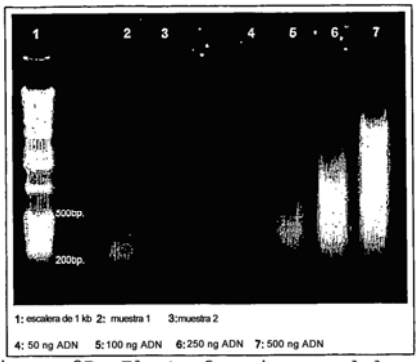


Figura 2B. Electroforesis en gel larga

Concentración de ADN medida con el NanoDrop

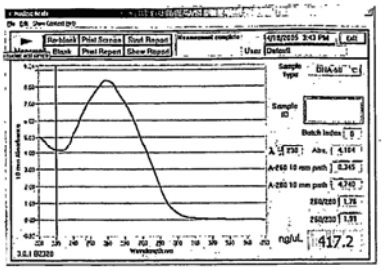


Figura 2C. Concentración muestra 1

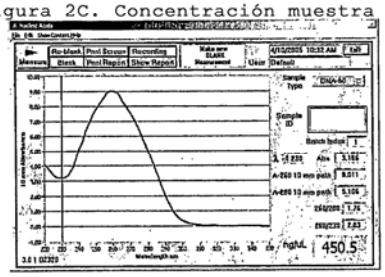


Figura 2D. Concentración muestra 2

Figura 3

Control de calidad del ADN en un gel de agarosa al 1%

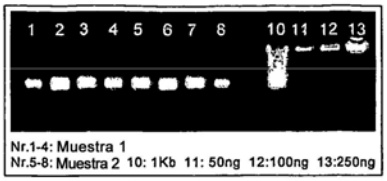


Figura 3A. Electroforesis en gel corta

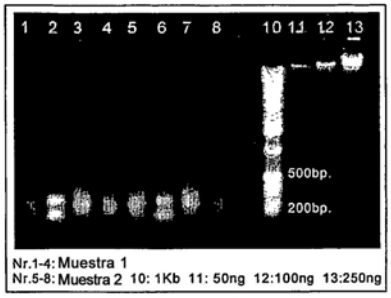


Figura 3B. Electroforesis en gel larga

Concentraciones de ADN medidas en el NanoDrop

Nr.	ID muestra	ng/uL	A260	260/280	260/230	Constante
1	P1.1	22.61	0.452	1.5	1.81	50
2	P1.2	19.08	0.382	1.67	2.49	50
3	P1.3	18.05	0.361	1.63	2.35	50
4	P1.4	15.19	0.304	1.71	2.1	50

Nr.	ID muestra	ng/uL	A260	260/280	260/230	Constante
5	P2.1	17.5	0.35	1.66	2.01	50
6	P2.2	16.67	0.333	1.96	2	50
7	P2.3	22.03	0.441	1.81	2.28	50
8	P2.4	9.8	0.196	1.78	1.98	50

Figura 4A. Esquema de procesamiento de datos de secuencia

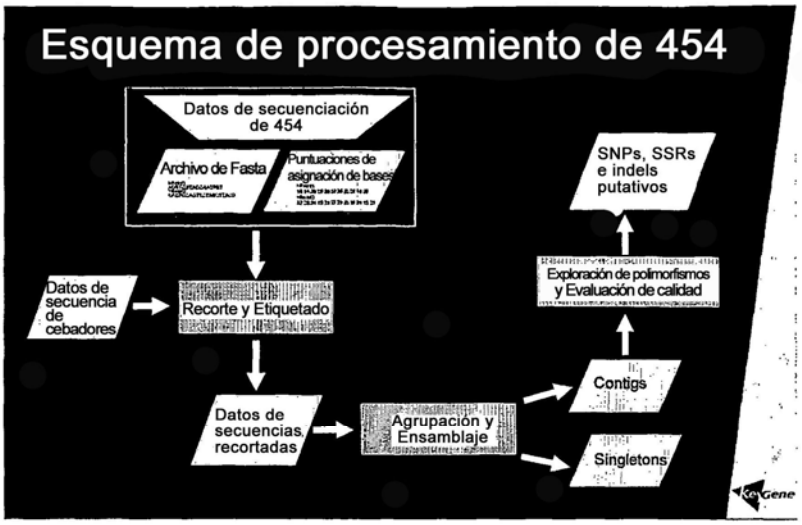


Figura 4B. Exploración de polimorfismos y procedimiento de asignación de calidad

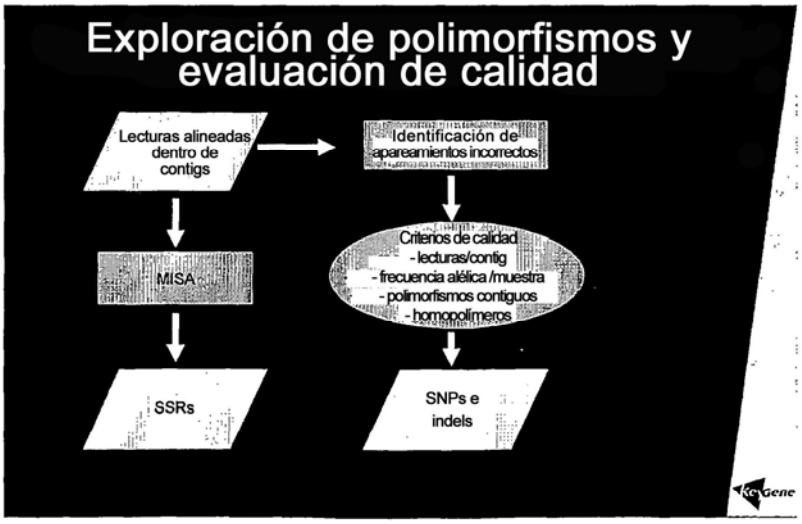


FIG 5A

Panel 1: Ejemplo de una etiqueta mixta

CAAGAGACTGCGTACCAATTCAACTTTGAGGTGAAAGATCGAAGGTTGCA
CAAGAGACTGCGTACCAATTCA (ES2)

AACACCAAGTGGCCGACCATCTCTTGCGTGTTACTCAGGACTCATCACCAC
(MS1) TGTTACTCAGGACTCATCACCA

Panel 2: vista general de fragmentos observados que contienen etiquetas esperadas y etiquetas mixtas

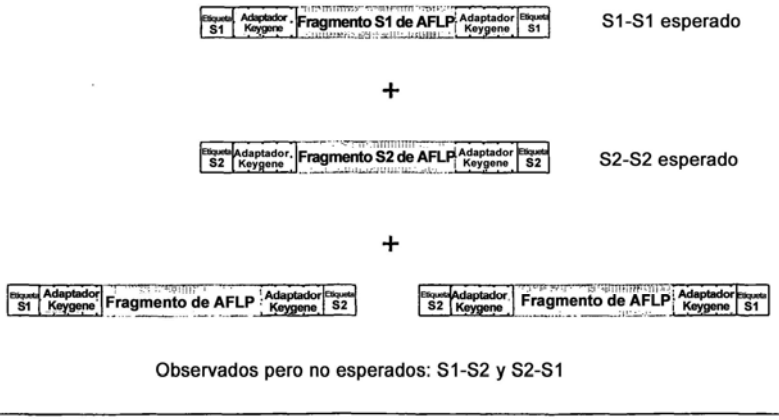


FIG 5B

Panel 3: causa hipotética de generación de etiquetas mixtas

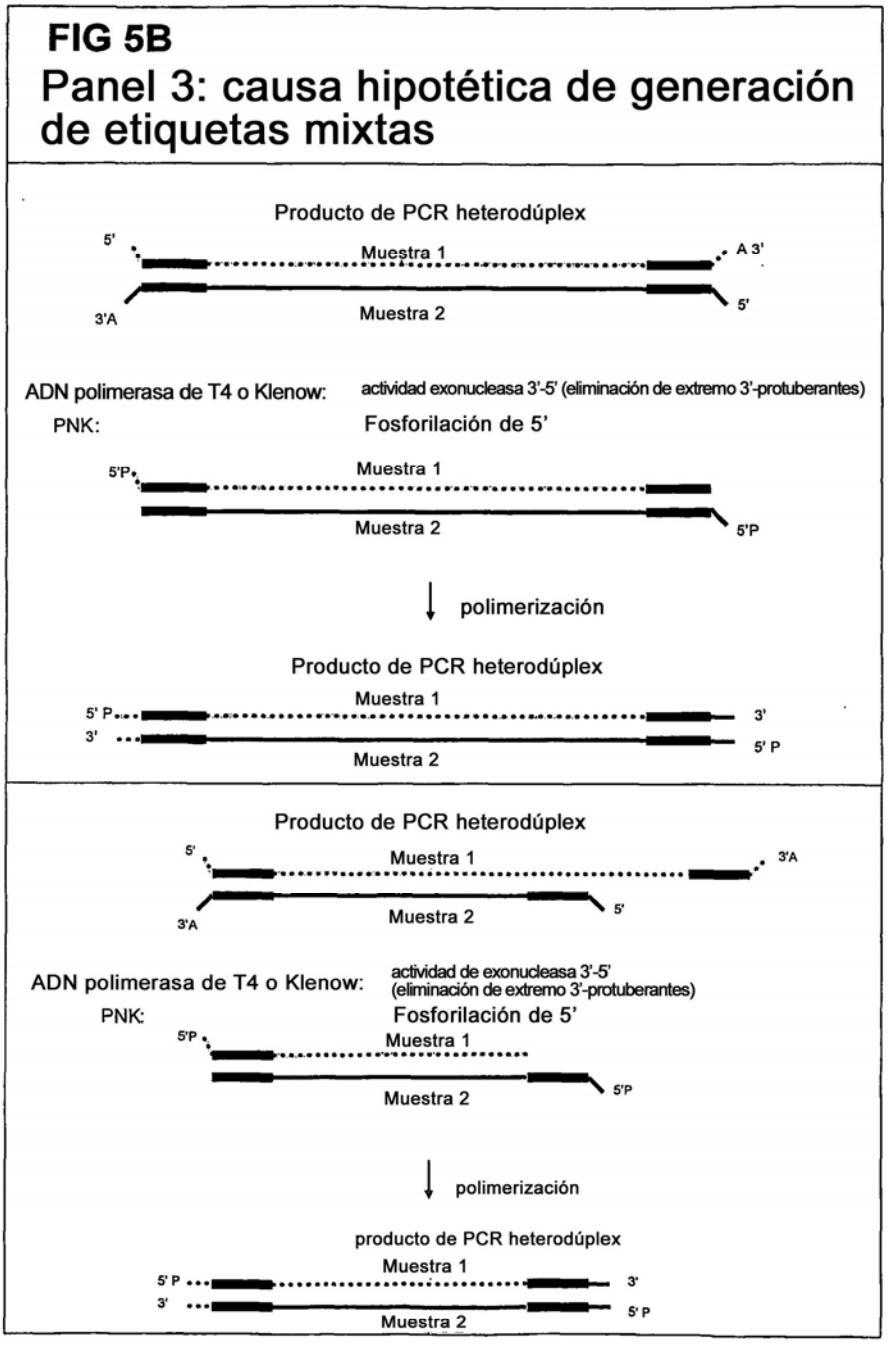


FIG 5C

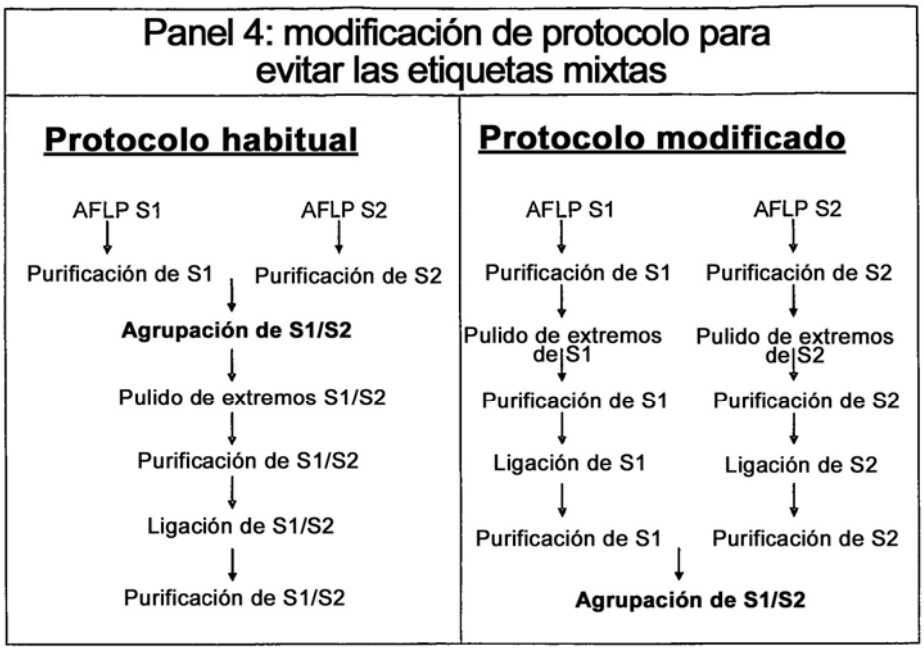

Panel 4: modificación de protocolo para evitar las etiquetas mixtas
Protocolo habitual
AFLP S1
AFLP S2
Purificación de S1
Purificación de S2
Agrupación de S1/S2
Pulido de extremos S1/S2
Purificación de S1/S2
Ligación de S1/S2
Purificación de S1/S2
Protocolo modificado
AFLP S1
AFLP S2
Purificación de S1
Purificación de S2
Pulido de extremos de S1
Pulido de extremos de S2
Purificación de S1
Purificación de S2
Ligación de S1
Ligación de S2
Purificación de S1
Purificación de S2
Agrupación de S1/S2

FIG 6A

Panel 1: Ejemplo de un concatámero

TGGTGATGAGTCCTGAGTAACGGGCCTTTCTTTGTACACT
TGGTGATGAGTCCTGAGTAACA (MS1)

TGAATTGGTACGCAGTCTGACG|AGCCGATGAGTCCTGAGTAACA
TGAATTGGTACGCAGTCTGACG|AGCCGATGAGTCCTGAGTAACA
(ES1) | (MS2)

TAAAGGTAAAGCGTGAATTGGTACG
TGAATTGGTACGCAGTCTCTTG (ES2)

Panel 2: Vista general de fragmentos observados que contienen etiquetas y concatámeros esperados

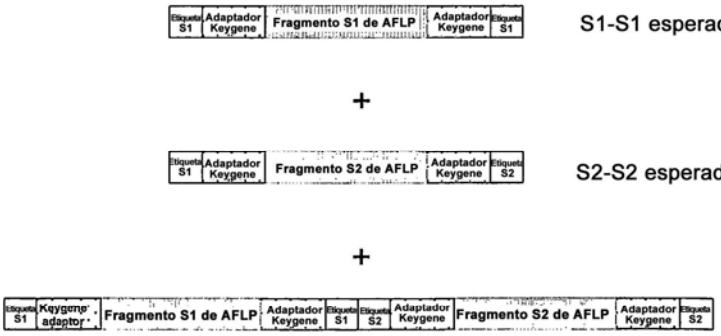


FIG 6B

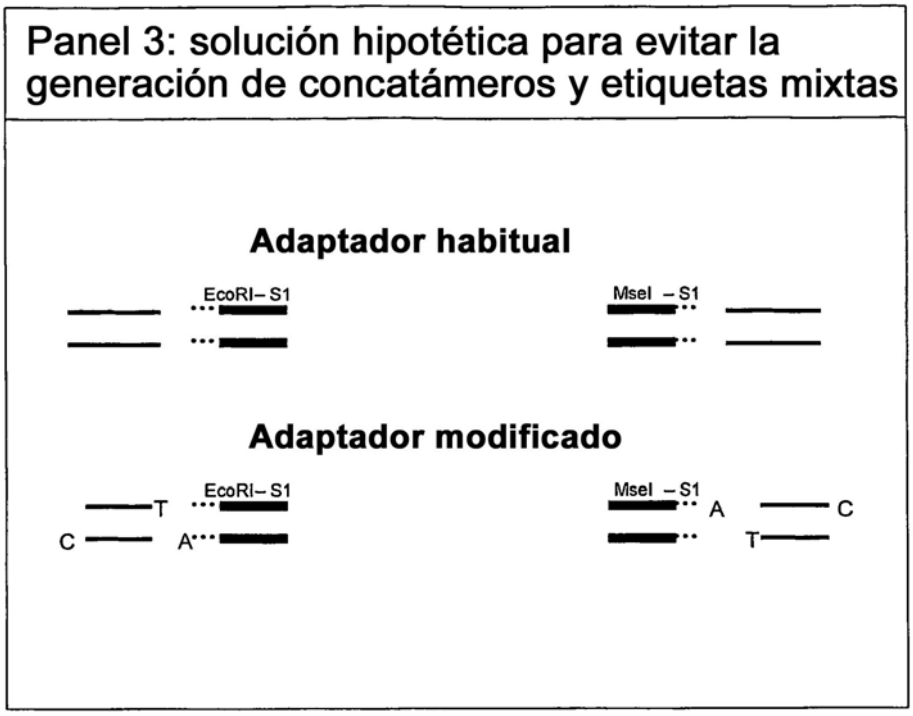

Panel 3: solución hipotética para evitar la generación de concatámeros y etiquetas mixtas
Adaptador habitual
EcoRI– S1
MseI – S1
Adaptador modificado
EcoRI– S1
MseI – S1
T
C
A
A
C
T

FIG 6C

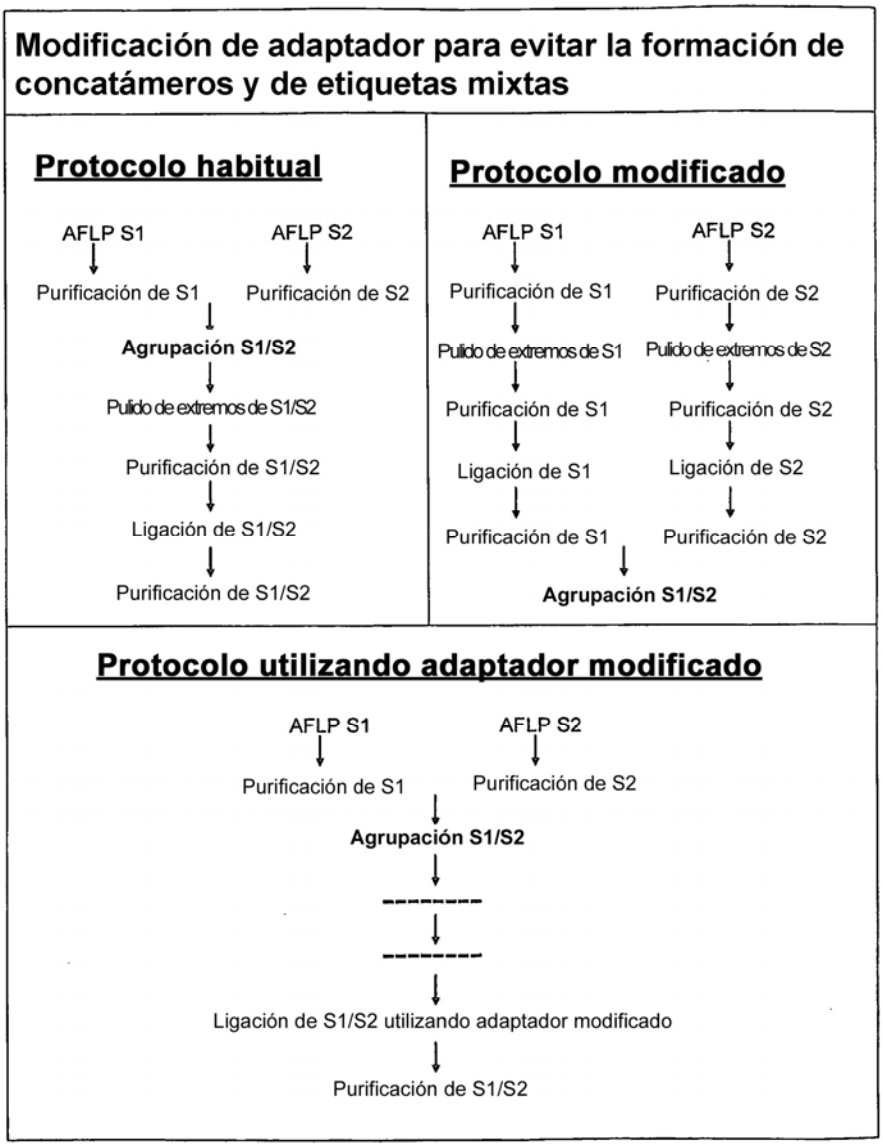

Modificación de adaptador para evitar la formación de concatámeros y de etiquetas mixtas
Protocolo habitual
AFLP S1
AFLP S2
Purificación de S1
Purificación de S2
Agrupación S1/S2
Pulido de extremos de S1/S2
Purificación de S1/S2
Ligación de S1/S2
Purificación de S1/S2
Protocolo modificado
AFLP S1
AFLP S2
Purificación de S1
Purificación de S2
Pulido de extremos de S1
Pulido de extremos de S2
Purificación de S1
Purificación de S2
Ligación de S1
Ligación de S2
Purificación de S1
Purificación de S2
Agrupación S1/S2
Protocolo utilizando adaptador modificado
AFLP S1
AFLP S2
Purificación de S1
Purificación de S2
Agrupación S1/S2
Ligación de S1/S2 utilizando adaptador modificado
Purificación de S1/S2

FIG 7

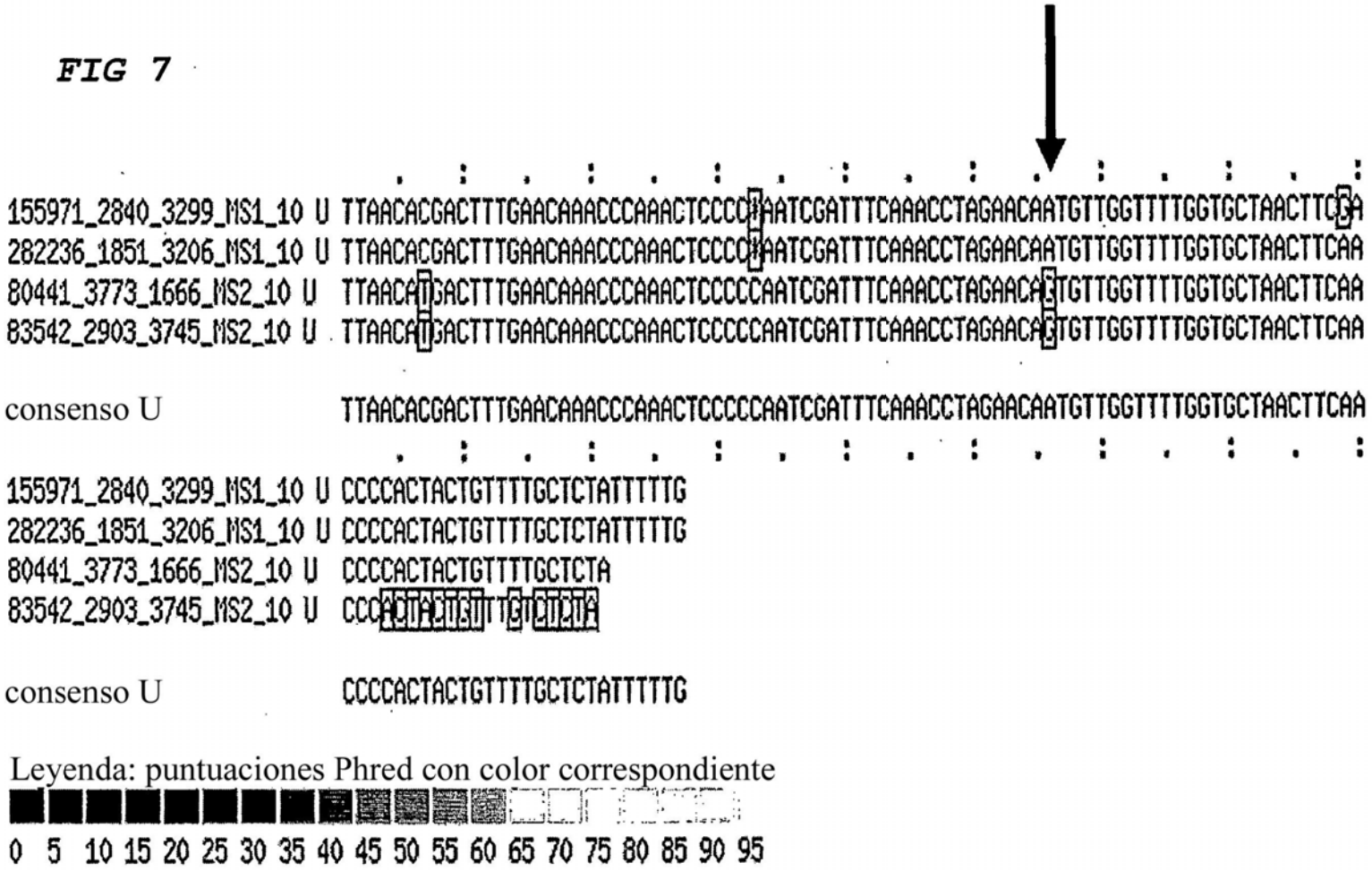

155971_2840_3299_MS1_10 U TTAACACGACTTTGAACAAACCCAAACTCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA
282236_1851_3206_MS1_10 U TTAACACGACTTTGAACAAACCCAAACTCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA
80441_3773_1666_MS2_10 U TTAACATGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACATGTTGGTTTTGGTGCTAACTTCAA
83542_2903_3745_MS2_10 U TTAACATGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACATGTTGGTTTTGGTGCTAACTTCAA
consenso U TTAACACGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA
155971_2840_3299_MS1_10 U CCCCACTACTGTTTTGCTCTATTTTTG
282236_1851_3206_MS1_10 U CCCCACTACTGTTTTGCTCTATTTTTG
80441_3773_1666_MS2_10 U CCCCACTACTGTTTTGCTCTA
83542_2903_3745_MS2_10 U CCCCACTACTGTTTTGCTCTATTTTTG
consenso U CCCCACTACTGTTTTGCTCTATTTTTG
Leyenda: puntuaciones Phred con color correspondiente
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95

FIG 8A.

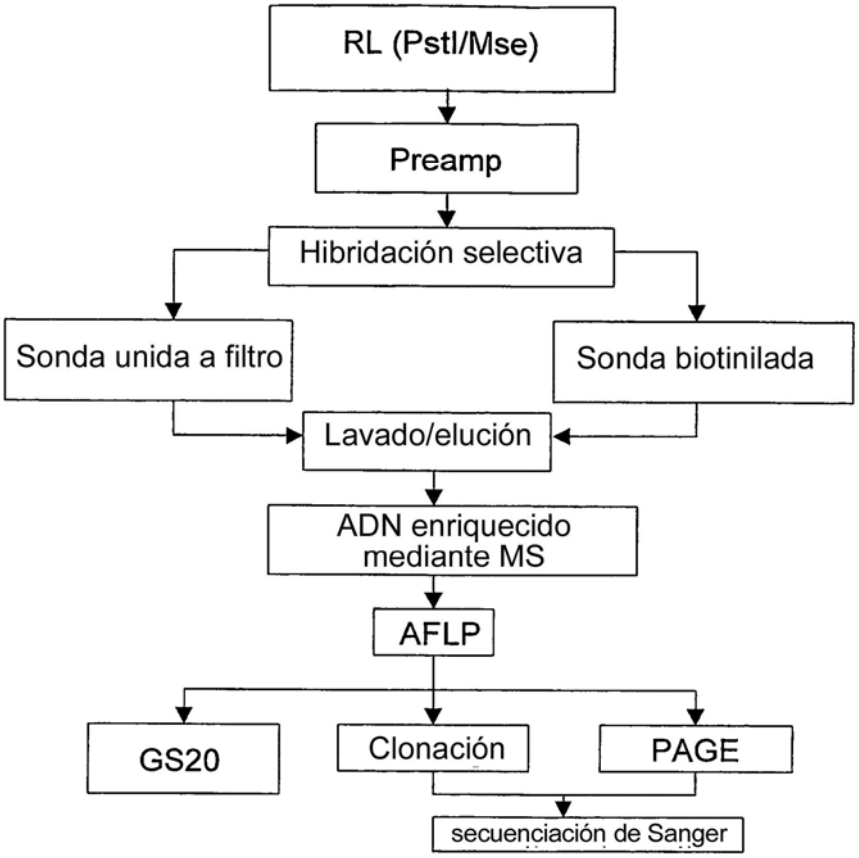
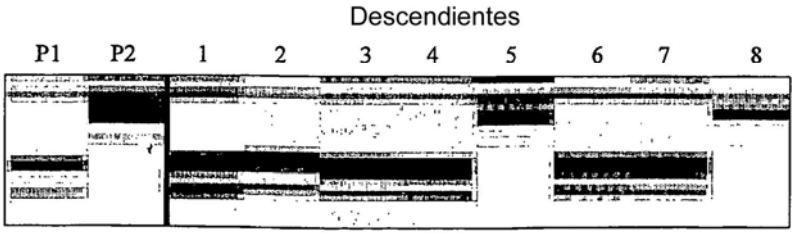

RL (PstI/Mse)
Preamp
Hibridación selectiva
Sonda unida a filtro
Sonda biotinilada
Lavado/elución
ADN enriquecido mediante MS
AFLP
GS20
Clonación
PAGE
secuenciación de Sanger

FIG 8B


Descendientes
P1
P2
1
2
3
4
5
6
7
8