(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 387 878**

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

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **10184351 .4**
(96) Fecha de presentación: **23.06.2006**
(97) Número de publicación de la solicitud: **2292788**
(97) Fecha de publicación de la solicitud: **09.03.2011**

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

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

(45) Fecha de publicación de la mención BOPI:
03.10.2012

(45) Fecha de la publicación del folleto de la patente:
03.10.2012

(73) Titular/es:
KEYGENE N.V.
P.O. Box 216
6700 AE Wageningen, NL

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

(74) Agente/Representante:
Sugrañes Moliné, Pedro

DESCRIPCIÓN

Estrategias para la identificación de alto rendimiento y la detección de polimorfismos

Campo técnico

La presente invención se refiere a los campos de la biología molecular y la genética. La invención se refiere a la rápida identificación de múltiples polimorfismos en una muestra de ácido nucleico. Los polimorfismos identificados pueden usarse para el desarrollo de sistemas de examen de alto rendimiento para detectar polimorfismos en muestras de prueba.

Antecedentes de la invención

La exploración del ADN genómico ha sido largamente deseada por la comunidad científica, en particular la médica. El ADN genómico contiene la clave para la identificación, el diagnóstico y el tratamiento de enfermedades tales como cáncer y enfermedad de Alzheimer. Además de la identificación y el tratamiento de la enfermedad, la exploración del ADN genómico puede proporcionar ventajas significativas en esfuerzos de cría de animales y plantas, lo que puede proporcionar respuestas a problemas de nutrición y alimentos en el mundo.

Se sabe que muchas enfermedades están asociadas con componentes genéticos específicos, en particular con polimorfismos en genes específicos. La identificación de polimorfismos en muestras grandes tales como genomas es en la actualidad una tarea laboriosa y que requiere mucho tiempo. Sin embargo, tal identificación es de gran valor en áreas tales como investigación biomédica, desarrollo de productos de farmacia, tipificación de tejidos, genotipado y estudios poblacionales.

Sumario de la invención

La presente invención proporciona un método de identificación eficaz y detección fiable de polimorfismos en una muestra de ácido nucleico compleja, por ejemplo muy grande (por ejemplo ADN o ARN) de una manera rápida y económica usando una combinación de métodos de alto rendimiento.

Esta integración de métodos de alto rendimiento entre sí proporciona una plataforma que es particularmente adecuada para la identificación y detección fiable de polimorfismos en muestras de ácido nucleico sumamente complejas en las que la identificación convencional y el mapeo de polimorfismos serían laboriosos y requerirían mucho tiempo.

Una de las cosas que los presentes inventores han encontrado es una solución para la identificación de polimorfismos, preferiblemente polimorfismos de nucleótido único, pero asimismo para (micro)satélites y/o indels, en particular en genomas grandes. El método es único en su aplicabilidad a genomas grandes y pequeños parecidos, pero proporciona ventajas particulares para genomas grandes, en particular especies poliploidales.

Para identificar SNP (y posteriormente detectar los SNP identificados), hay varias posibilidades disponibles en la técnica. En una primera opción, puede secuenciarse el genoma completo, y esto puede hacerse así para varios individuos. Esto es principalmente un ejercicio teórico, ya que esto es engorroso y caro y, a pesar del rápido desarrollo de la tecnología, simplemente no es viable hacerlo para cada organismo, especialmente los que tienen genomas más grandes. Una segunda opción es usar la información de secuencia disponible (fragmentada), tal como bibliotecas de EST. Esto permite la generación de cebadores de PCR, la resecuenciación y la comparación entre individuos. De nuevo, esto requiere información de secuencia inicial que no está disponible o sólo en una cantidad limitada. Además, tienen que desarrollarse ensayos de PCR separados para cada región, lo que añade un enorme coste y tiempo de desarrollo.

La tercera opción es limitarse uno mismo a parte del genoma para cada individuo. La dificultad reside en que la parte proporcionada del genoma debe ser la misma para diferentes individuos con el fin de proporcionar un resultado comparable para lograr una identificación de SNP satisfactoria. Los presentes inventores han solucionado ahora este dilema mediante la integración de métodos sumamente reproducibles para seleccionar parte del genoma con secuenciación de alto rendimiento para la identificación de polimorfismos integrados con preparación de muestras y plataformas de identificación de alto rendimiento. La presente invención acelera el proceso de descubrimiento de polimorfismos y usa los mismos elementos en el proceso posterior para el aprovechamiento de los polimorfismos descubiertos para permitir un genotipado de alto rendimiento fiable y eficaz.

Las aplicaciones previstas adicionales del método de la presente invención incluye el examen de bibliotecas de microsatélites enriquecidas, realizar la obtención del perfil de transcrito de ADNc-AFLP (hibridación de tipo Northern digital), secuenciación de genomas complejos, secuenciación de bibliotecas de EST (en ADNc completo o ADNc-AFLP), descubrimiento de microARN (secuenciación de bibliotecas de insertos pequeñas), secuenciación de cromosomas artificiales bacterianos (BAC) (cóntigos), enfoque de análisis de segregantes agrupados AFLP/ADNc-AFLP, detección de rutina de fragmentos de AFLP, por ejemplo para retrocruzamientos asistidos por marcador

(MABC), etcétera.

Definiciones

En la siguiente descripción y ejemplos, se usan varios términos. Con el fin de proporcionar una comprensión clara y sistemática de la memoria descriptiva y las reivindicaciones, incluyendo el alcance que se da a tales términos, se proporcionan las siguientes definiciones. A menos que se defina lo contrario en el presente documento, todos los términos técnicos y científicos usados tienen el mismo significado que entiende comúnmente un experto habitual en la técnica a la que pertenece esta invención. Las descripciones de todas las publicaciones, solicitudes de patente, patentes y otras referencias se incorporan en el presente documento en su totalidad como referencia.

Polimorfismo: Polimorfismo se refiere 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 nucleótido único (SNP), que es una variación, que se produce cuando se altera un único nucleótido: adenina (A), timina (T), citosina (C) o guanina (G). Una variación debe producirse generalmente en al menos el 1% de la población para que se considere un SNP. Los SNP constituyen por ejemplo el 90% de todas las variaciones genéticas humanas, y se producen cada de 100 a 300 bases a lo largo del genoma humano. Dos de cada tres SNP sustituyen citosina (C) por timina (T). Las variaciones en las secuencias de ADN de por ejemplo seres humanos o plantas pueden afectar a cómo soportan enfermedades, bacterias, virus, productos químicos, fármacos, etc.

Ácido nucleico: Un ácido nucleico según la presente invención puede incluir cualquier polímero u oligómero de bases de pirimidina y purina, preferiblemente citosina, timina y uracilo, y adenina y guanina, respectivamente (véase Albert L. Lehninger, Principles of Biochemistry, en 793-800 (Worth Pub. 1982) que se incorpora en el presente documento como referencia en su totalidad para todos los fines). La presente invención contempla cualquier desoxirribonucleótido, ribonucleótido o componente de ácido nucleico peptídico, y cualquier variante química de los mismos, tales como formas metiladas, hidroximetiladas o glicosiladas de estas bases, y similares. Los polímeros u oligómeros pueden ser heterogéneos u homogéneos en composición, y pueden aislarse de fuentes que se producen de manera natural o pueden producirse de manera sintética o artificial. Además, los ácidos nucleicos pueden ser ADN o ARN, o una mezcla de los mismos, y pueden existir permanente o transitoriamente en forma mono o bicatenaria, incluyendo homodúplex, heterodúplex y estados híbridos.

Reducción de la complejidad: El término reducción de la complejidad se usa para indicar un método en el que se reduce la complejidad de una muestra de ácido nucleico, tal como ADN genómico, mediante la generación de un subconjunto de la muestra. Este subconjunto puede ser representativo de la muestra completa (es decir, compleja) y es preferiblemente un subconjunto reproducible. Reproducible significa en este contexto que cuando se reduce la complejidad de la misma muestra usando el mismo método, se obtiene el mismo subconjunto o al menos uno comparable. El método usado para la reducción de la complejidad puede ser cualquier método para la reducción de la complejidad conocido en la técnica. Los ejemplos de métodos para la reducción de la complejidad incluyen por ejemplo AFLP® (Keygene N.V., Países Bajos; véase por ejemplo el documento EP 0 534 858), los métodos descritos por Dong (véase por ejemplo los documentos WO 03/012118, WO 00/24939), unión indexada (Unrau *et al*., véase a continuación), etc. Los métodos de reducción de la complejidad usados en la presente invención tienen en común que son reproducibles. Reproducible en el sentido de que cuando se reduce la complejidad de la misma muestra de la misma manera, se obtiene el mismo subconjunto de la muestra, en oposición a una reducción de la complejidad más aleatoria tal como microdisección o el uso de ARNm (ADNc) que representa una parte del genoma transcrito en un tejido seleccionado y por ello su reproducibilidad depende de la selección de tejido, el tiempo de aislamiento, etc.

Etiquetado: El término etiquetado se refiere a la adición de una etiqueta a una muestra de ácido nucleico con el fin de poder distinguirla de una segunda muestra de ácido nucleico o adicionales. El etiquetado puede realizarse por ejemplo mediante la adición de un identificador de secuencia durante la reducción de la complejidad o mediante cualquier otro medio conocido en la técnica. Tal identificador de secuencia puede ser por ejemplo una secuencia de bases única de longitud variable pero definida usada únicamente para identificar una muestra de ácido nucleico específica. Los ejemplos típicos de las mismas son por ejemplo secuencias ZIP. Usando tal etiqueta, puede determinarse el origen de una muestra tras el procesamiento adicional. En el caso de combinar productos procesados que se originan a partir de diferentes muestras de ácido nucleico, las diferentes muestras de ácido nucleico deben identificarse usando etiquetas diferentes.

Biblioteca etiquetada: El término biblioteca etiquetada se refiere a una biblioteca de ácido nucleico etiquetado.

Secuenciación: El término secuenciación se refiere a la determinación del orden de nucleótidos (secuencias de bases) en una muestra de ácido nucleico, por ejemplo ADN o ARN.

Alinear y alineación: Con el término “alinear” y “alineación” quiere decirse la comparación de dos o más secuencias de nucleótidos basada en la presencia de tramos cortos o largos de nucleótidos similares o idénticos. Se conocen en la técnica varios métodos para la alineación de secuencias de nucleótidos, tal como se explicará adicionalmente a

continuación.

Sondas de detección: El término “sondas de detección” se usa para indicar sondas diseñadas para detectar una secuencia de nucleótidos específica, en particular secuencias que contienen uno o más polimorfismos.

Examen de alto rendimiento: El examen de alto rendimiento, abreviado a menudo como HTS, es un método de experimentación científica especialmente relevante en los campos de biología y química. A través de una combinación de robótica moderna y otro equipo de laboratorio especializado, permite a un investigador examinar de manera eficaz grandes cantidades de muestras simultáneamente.

Ácido nucleico de muestra de prueba: El término “ácido nucleico de muestra de prueba” se usa para indicar una muestra de ácido nucleico que se investiga para detectar polimorfismos usando el método de la presente invención.

Endonucleasa de restricción: Una endonucleasa de restricción o enzima de restricción es una enzima que reconoce una secuencia de nucleótidos específica (sitio diana) en una molécula de ADN bicatenaria, y escindirá ambas hebras de la molécula de ADN en cada sitio diana.

Fragmentos de restricción: Las moléculas de ADN producidas mediante digestión con una endonucleasa de restricción se denominan fragmentos de restricción. Cualquier genoma dado (o ácido nucleico, independientemente de su origen) se digerirá por una endonucleasa de restricción particular para dar un conjunto diferenciado de fragmentos de restricción. Los fragmentos de ADN que resultan de la escisión con endonucleasas de restricción pueden usarse adicionalmente en una variedad de técnicas y pueden detectarse por ejemplo mediante electroforesis en gel.

Electroforesis en gel: Con el fin de detectar fragmentos de restricción, puede requerirse un método analítico para fraccionar moléculas de ADN bicatenarias basándose en el tamaño. La técnica más comúnmente usada para lograr tal fraccionamiento es la electroforesis en gel (capilar). La velocidad a la que los fragmentos de ADN se mueven en tales geles depende de su peso molecular; por tanto, las distancias recorridas disminuyen a medida que la longitud del fragmento aumenta. Los fragmentos de ADN fraccionados mediante electroforesis en gel pueden visualizarse directamente mediante un procedimiento de tinción, por ejemplo tinción con plata o tinción usando bromuro de etidio, si el número de fragmentos incluidos en el patrón es suficientemente pequeño. Alternativamente, el tratamiento adicional de los fragmentos de ADN puede incorporar marcadores detectables en los fragmentos, tales como fluoróforos o marcadores radiactivos.

Ligamiento: La reacción enzimática catalizada por una enzima ligasa en la que dos moléculas de ADN bicatenarias se unen covalentemente entre sí se denomina ligamiento. En general, ambas hebras de ADN se unen covalentemente entre sí, pero también es posible impedir el ligamiento de una de las dos hebras a través de modificación química o enzimática de uno de los extremos de las hebras. En ese caso, la unión covalente se producirá en sólo una de las dos hebras del ADN.

Oligonucleótido sintético: Moléculas de ADN monocatenarias que tienen preferiblemente desde aproximadamente 10 hasta aproximadamente 50 bases, que pueden sintetizarse químicamente se denominan oligonucleótidos sintéticos. En general, estas moléculas de ADN sintéticas están diseñadas para tener una secuencia de nucleótidos única o deseada, aunque es posible sintetizar familias de moléculas que tienen secuencias relacionadas y que tienen diferentes composiciones de nucleótidos en posiciones específicas dentro de la secuencia de nucleótidos. El término oligonucleótido sintético se usará para referirse a moléculas de ADN que tienen una secuencia de nucleótidos diseñada o deseada.

Adaptadores: Moléculas de ADN bicatenarias cortas con un número limitado de pares de bases, por ejemplo de aproximadamente 10 a aproximadamente 30 pares de bases de longitud, que están diseñados de manera que puedan ligarse a los extremos de los fragmentos de restricción. Los adaptadores están compuestos generalmente por dos oligonucleótidos sintéticos que tienen secuencias de nucleótidos que son parcialmente complementarias entre sí. Cuando se mezclan los dos oligonucleótidos sintéticos en disolución en condiciones apropiadas, se aparearán entre sí formando una estructura bicatenaria. Tras aparearse, un extremo de la molécula adaptadora está diseñado de manera que sea compatible con el extremo de un fragmento de restricción y pueda ligarse al mismo; el otro extremo del adaptador puede estar diseñado de modo que no pueda ligarse, pero no es necesario que éste sea el caso (adaptadores ligados dobles).

Fragmentos de restricción ligados a adaptadores: Fragmentos de restricción cuyos extremos se han ocupado mediante adaptadores.

Cebadores: En general, el término cebadores se refiere a una hebra de ADN que puede cebar la síntesis de ADN. La ADN polimerasa no puede sintetizar ADN *de novo* sin cebadores: sólo puede extender una hebra de ADN existente en una reacción en la que se usa la hebra complementaria como molde para dirigir el orden de nucleótidos que han de ensamblarse. Se hará referencia a las moléculas de oligonucleótidos sintéticos que se usan en una reacción en cadena de la polimerasa (PCR) como cebadores.

Amplificación del ADN: El término amplificación del ADN se usará normalmente para indicar la síntesis *in vitro* de moléculas de ADN bicatenarias usando PCR. Se indica que existen otros métodos de amplificación y pueden usarse en la presente invención sin apartarse de lo esencial.

Descripción detallada de la invención

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

a) proporcionar una primera muestra de ácido nucleico de interés;

b) realizar una reducción de la complejidad en la primera muestra de ácido nucleico de interés para proporcionar una primera biblioteca de la primera muestra de ácido nucleico;

c) realizar consecutiva o simultáneamente las etapas a) y b) con una segunda muestra de ácido nucleico de interés o adicionales para obtener una segunda biblioteca o adicionales de la segunda muestra de ácido nucleico de interés o adicionales;

d) secuenciar al menos una parte de la primera biblioteca y la segunda biblioteca o adicionales;

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

f) determinar uno o más polimorfismos entre la primera muestra de ácido nucleico y la segunda muestra de ácido nucleico o adicionales en la alineación de la etapa e);

g) usar el uno o más polimorfismos determinados en la etapa f) para diseñar una o más sondas de detección;

h) proporcionar un ácido nucleico de muestra de prueba de interés;

i) realizar la reducción de la complejidad de la etapa b) en el ácido nucleico de muestra de prueba de interés para proporcionar una biblioteca de prueba del ácido nucleico de muestra de prueba;

j) someter la biblioteca de prueba a examen de alto rendimiento para identificar la presencia, ausencia o la cantidad de los polimorfismos determinados en la etapa f) usando la una o más sondas de detección diseñadas en la etapa g).

En la etapa a), se proporciona una primera muestra de ácido nucleico de interés. Dicha primera muestra de ácido nucleico de interés es preferiblemente una muestra de ácido nucleico compleja tal como ADN genómico total o una biblioteca de ADNc. Se prefiere que la muestra de ácido nucleico compleja sea ADN genómico total.

En la etapa b), se realiza una reducción de la complejidad en la primera muestra de ácido nucleico de interés para proporcionar una primera biblioteca de la primera muestra de ácido nucleico.

En una realización de la invención, la etapa de reducción de la complejidad de la muestra de ácido nucleico comprende cortar enzimáticamente la muestra de ácido nucleico en fragmentos de restricción, separar los fragmentos de restricción y seleccionar un grupo particular de fragmentos de restricción. Opcionalmente, los fragmentos seleccionados se ligan entonces a secuencias adaptadoras que contienen secuencias de unión/moldes de cebadores de PCR.

En una realización de la reducción de la complejidad, se usa un tipo de endonucleasa IIs para digerir la muestra de ácido nucleico y los fragmentos de restricción se ligan selectivamente a secuencias adaptadoras. Las secuencias adaptadoras pueden contener diversos nucleótidos en la proyección que va a ligarse y sólo el adaptador con el conjunto de nucleótidos coincidentes en la proyección se liga al fragmento y posteriormente se amplifica. Esta tecnología se representa en la técnica como ligadores de indexación. Pueden encontrarse ejemplos de este principio entre otros en Unrau P. y Deugau K.V. (1994) Gene 145:163-169.

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

En otra realización de la invención, la etapa de reducción de la complejidad comprende realizar una PCR cebada arbitrariamente sobre la muestra.

Aún en otra realización de la invención, la etapa de reducción de la complejidad comprende eliminar secuencias repetidas desnaturalizando y volviendo a aparear el ADN y entonces eliminando dúplex bicatenarios.

En otra realización de la invención, la etapa de reducción de la complejidad comprende hibridar la muestra de ácido nucleico con una perla magnética que se une a una sonda oligonucleotídica que contiene una secuencia deseada. Esta realización puede comprender además exponer la muestra hibridada a una ADN nucleasa monocatenaria para eliminar el ADN monocatenario, ligando una secuencia adaptadora que contiene una enzima de restricción de la clase IIs para liberar la perla magnética. Esta realización puede comprender o no la amplificación de la secuencia de ADN aislada. Además, la secuencia adaptadora puede usarse o no como molde para el cebador oligonucleotídico de PCR. En esta realización, la secuencia adaptadora puede contener o no una etiqueta o identificador de secuencia.

En otra realización, el método de reducción de la complejidad comprende exponer la muestra de ADN a una proteína de unión a apareamientos erróneos y digerir la muestra con una exonucleasa de 3' a 5' y luego una nucleasa monocatenaria. Esta realización puede incluir o no el uso de una perla magnética unida a la proteína de unión a apareamientos erróneos.

En otra realización de la presente invención, la reducción de la complejidad comprende el método CHIP tal como se describe en el presente documento en otra parte o el diseño de cebadores de PCR dirigidos contra motivos conservados tales como SSR, regiones NBS (regiones de unión a nucleótidos), secuencias promotoras/potenciadoras, secuencias consenso de telómeros, genes de caja MADS, familias de genes de ATP-asa y otras familias de genes.

En la etapa c), las etapas a) y b) se realizan consecutiva o simultáneamente con una segunda muestra de ácido nucleico o adicionales de interés para obtener una segunda biblioteca o adicionales de la segunda muestra de ácido nucleico o adicionales de interés. Dicha segunda muestra de ácido nucleico o adicionales de interés es también preferiblemente una muestra de ácido nucleico compleja tal como ADN genómico total. Se prefiere que la muestra de ácido nucleico compleja sea ADN genómico total. También se prefiere que dicha segunda muestra de ácido nucleico o adicionales se refiera a la primera muestra de ácido nucleico. La primera muestra de ácido nucleico y el segundo ácido nucleico o adicionales pueden ser por ejemplo líneas diferentes de una planta, tales como líneas de pimiento diferentes, o variedades diferentes. Las etapas a) y b) pueden realizarse para simplemente una segunda muestra de ácido nucleico de interés, pero también pueden realizarse adicionalmente para una tercera, cuarta, quinta, etc. muestra de ácido nucleico de interés.

Debe indicarse que el método según la presente invención será lo más útil cuando la reducción de la complejidad se realiza usando el mismo método y en sustancialmente las mismas condiciones, preferiblemente idénticas, para la primera muestra de ácido nucleico y la segunda muestra de ácido nucleico o adicionales. En tales condiciones, se obtendrán fracciones similares (comparables) de las muestras de ácido nucleico (complejas).

En la etapa d), se secuencia al menos una parte de la primera biblioteca y de la segunda biblioteca o adicionales. Se prefiere que la cantidad de solapamiento de fragmentos secuenciados de la primera biblioteca y la segunda biblioteca o adicionales sea al menos del 50%, más preferiblemente al menos del 60%, aún más preferiblemente al menos del 70%, incluso más preferiblemente al menos del 80%, aún más preferiblemente al menos del 90% y lo más preferiblemente al menos del 95%.

La secuenciación puede realizarse en principio mediante cualquier medio conocido en la técnica, tal como el método de terminación de la cadena didesoxi. Sin embargo, se prefiere que la secuenciación se realice usando métodos de secuenciación de alto rendimiento, tales como los métodos dados a conocer en los documentos WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007 y WO 2005/003375 (todos a nombre de 454 Corporation), por Seo *et al*. (2004) Proc. Natl. Acad. Sci. USA 101:5488-93, y tecnologías de Helios, Solexa, US Genomics. Lo más preferido es que la secuenciación se realiza usando el aparato y/o método dado a conocer en los documentos WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007 y WO 2005/003375 (todos a nombre de 454 Corporation). La tecnología descrita permite la secuenciación de 40 millones de bases en una única ronda y es 100 veces más rápida y barata que la tecnología de la competencia. La tecnología de secuenciación consiste en líneas generales en 4 etapas: 1) fragmentación del ADN y ligamiento de un adaptador específico a una biblioteca de ADN monocatenario (ADNmc); 2) apareamiento de ADNmc a perlas y emulsificación de las perlas en microrreactores de agua en aceite; 3) deposición de perlas que portan ADN en una placa PicoTiter®; y 4) secuenciación simultánea en 100.000 pocillos mediante generación de una señal luminosa de pirofosfato. El método se explicará en más detalle a continuación.

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

La herramienta de búsqueda de alineación local básica del NCBI ("*Basic Local Alignment Search Tool*") (BLAST)

(Altschul *et al*., 1990) está disponible de varias fuentes, incluyendo el Centro Nacional de Información Biológica ("*National Center for Biological Information*") (NCBI, Bethesda, Md.) e Internet, para su uso conjuntamente con los programas de análisis de secuencias blastp, blastn, blastx, tblastn y tblastx. Puede accederse al mismo en <http://www.ncbi. nlm.nih.gov/BLAST/>. Está disponible una descripción de cómo determinar la identidad de secuencia usando este programa en <http: //www.ncbi.nlm.nih.gov/BLAST/blasthelp.html>. Una aplicación adicional puede ser en prospección de microsatélites (véase Varshney *et al*. (2005) Trends in Biotechn. 23(1):48-55.

Normalmente, la alineación se realiza sobre los datos de secuencias que se han recortado por los adaptadores/cebador y/o identificadores, es decir, usando sólo los datos de secuencias de los fragmentos que se originan a partir de la muestra de ácido nucleico. Normalmente, los datos de secuencias se usan para identificar el origen del fragmento (es decir, de qué muestra procede), las secuencias derivadas del adaptador y/o identificador se eliminan de los datos y se realiza la alineación sobre este conjunto recortado.

En la etapa f), se determinan uno o más polimorfismos entre la primera muestra de ácido nucleico y la segunda muestra de ácido nucleico o adicionales en la alineación. La alineación puede hacerse de manera que las secuencias derivadas de la primera muestra de ácido nucleico y la segunda muestra de ácido nucleico o adicionales puedan compararse. Pueden identificarse entonces diferencias que reflejan polimorfismos.

En la etapa g), el uno o más polimorfismos determinados en la etapa g) se usan para diseñar sondas de detección, por ejemplo para la detección mediante hibridación sobre chips de ADN o una plataforma de detección a base de perlas. Las sondas de detección están diseñadas de manera que se refleja en las mismas un polimorfismo. En el caso de polimorfismos de nucleótido único (SNP), las sondas de detección contienen normalmente los alelos de SNP variantes en la posición central tal como para maximizar la diferenciación de alelos. Tales sondas pueden usarse ventajosamente para examinar muestras de prueba que tienen un determinado polimorfismo. Las sondas pueden sintetizarse usando cualquier método conocido en la técnica. Las sondas están diseñadas normalmente de manera que son adecuadas para métodos examen de alto rendimiento.

En la etapa h), se proporciona un ácido nucleico de muestra de prueba de interés. El ácido nucleico de muestra de prueba puede ser cualquier muestra, pero es preferiblemente otra línea o variedad que va a mapearse para detectar polimorfismos. Comúnmente, se usa una colección de muestras de prueba que representan el germoplasma de los organismos estudiados para 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 mapeo genético en la etapa de validación con el fin de determinar la posición en el mapa genético del polimorfismo también.

En la etapa i), la reducción de la complejidad de la etapa b) se realiza sobre el ácido nucleico de muestra de prueba de interés para proporcionar una biblioteca de prueba del ácido nucleico de muestra de prueba. Se prefiere sumamente que durante todo el método según la presente invención se use el mismo método para la reducción de la complejidad usando sustancialmente las mismas condiciones, preferiblemente idénticas, cubriendo así una fracción similar de la muestra. Sin embargo, no se requiere obtener una biblioteca de prueba etiquetada, aunque puede estar presente una etiqueta en los fragmentos de la biblioteca de prueba.

En la etapa j), se somete la biblioteca de prueba a examen de alto rendimiento para identificar la presencia, ausencia o cantidad de los polimorfismos determinados en la etapa f) usando las sondas de detección diseñadas en la etapa g). Un experto en la técnica conoce varios métodos para el examen de alto rendimiento usando sondas. Se prefiere que se inmovilicen una o más sondas diseñadas usando la información obtenida en la etapa g) sobre una matriz, tal como un chip de ADN, y que tal matriz se ponga en contacto posteriormente con la biblioteca de prueba en condiciones de hibridación. Los fragmentos de ADN dentro de la biblioteca de prueba complementarios a una o más sondas en la matriz se hibridarán en tales condiciones con tales sondas, y por tanto podrán detectarse. También se prevén otros métodos de examen de alto rendimiento dentro del alcance de la presente invención, tales como inmovilización de la biblioteca de prueba obtenida en la etapa j) y puesta en contacto de dicha biblioteca de prueba inmovilizada con las sondas diseñadas en la etapa h) en condiciones de hibridación.

Se proporciona otra técnica de examen por secuenciación de alto rendimiento entre otros por Affymetrix usando detección basada en chips de SNP y tecnología de perlas proporcionada por Illumina.

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 la segunda biblioteca etiquetada o adicionales para obtener una biblioteca combinada.

Se prefiere que el etiquetado se realice durante la etapa de reducción de la complejidad para reducir la cantidad de etapas requeridas para obtener la primera biblioteca etiquetada de la primera muestra de ácido nucleico. Tal etiquetado simultáneo puede lograrse por ejemplo mediante AFLP, usando adaptadores que comprenden un identificador único (nucleótido) para cada muestra.

Se pretende que el etiquetado distinga entre muestras de diferente origen, por ejemplo obtenidas de diferentes líneas de plantas, cuando las bibliotecas de dos o más muestras de ácido nucleico se combinan para obtener una biblioteca de combinación. Por tanto, se usan preferiblemente etiquetas diferentes para preparar las bibliotecas etiquetadas de la primera muestra de ácido nucleico y la segunda muestra de ácido nucleico o adicionales. Cuando se usan por ejemplo cinco muestras de ácido nucleico, se pretende obtener cinco bibliotecas etiquetadas de manera diferente, indicando las cinco etiquetas diferentes las muestras originales respectivas.

La etiqueta puede ser cualquier etiqueta conocida en la técnica para distinguir muestras de ácido nucleico, pero es preferiblemente una secuencia identificadora corta. Tal secuencia identificadora puede ser por ejemplo una secuencia de bases única de longitud variable usada para indicar el origen de la biblioteca obtenida mediante reducción de la complejidad.

En una realización preferida, el etiquetado de la primera biblioteca y la segunda biblioteca o adicionales se realiza usando etiquetas diferentes. Tal como se comentó anteriormente, se prefiere que cada biblioteca de una muestra de ácido nucleico se identifique mediante su propia etiqueta. El ácido nucleico de muestra de prueba no necesita estar etiquetado.

En una realización preferida de la invención, la reducción de la complejidad se realiza por medio de AFLP® (Keygene N.V., Países Bajos; véase por ejemplo el documento EP 0 534 858 y Vos *et al*. (1995). AFLP: a new technique for DNA fingerprinting, Nucleic Acids Research, vol. 23, n.º 21, 07-4414).

AFLP es un método para la amplificación selectiva de fragmentos de restricción. AFLP no necesita ninguna información de secuencia previa y puede realizarse sobre cualquier ADN de partida. En general, AFLP comprende las etapas de:

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

(b) ligar los fragmentos de restricción así obtenidos con un adaptador oligonucleotídico sintético bicatenario, un extremo del cual es compatible con uno o ambos de los extremos de los fragmentos de restricción, para producir de ese modo fragmentos de restricción ligados a adaptador, preferiblemente etiquetados, del ADN de partida;

(c) poner en contacto los fragmentos de restricción ligados a adaptador, preferiblemente etiquetados, en condiciones de hibridación con al menos un cebador oligonucleotídico que contiene al menos un nucleótido selectivo en su extremo 3';

(d) amplificar el fragmento de restricción ligado a adaptador, preferiblemente etiquetado, hibridado con los cebadores mediante PCR o una técnica similar de modo para provocar el alargamiento adicional de los cebadores hibridados a lo largo de los fragmentos de restricción del ADN de partida con el que los cebadores se hibridaron; y (e) detectar, identificar o recuperar el fragmento de ADN amplificado o alargado así obtenido.

AFLP proporciona por tanto un subconjunto reproducible de fragmentos ligados a adaptador. Otros métodos adecuados para la reducción de la complejidad son inmunoprecipitación de cromatina ("*Chromatine Immuno Precipitation*") (ChiP). Esto significa que se aísla ADN nuclear, mientras que proteínas tales como factores de transcripción se unen al ADN. Con ChiP, se usa en primer lugar un anticuerpo contra la proteína, dando como resultado un complejo Ac-proteína-ADN. Purificando este complejo y precipitándolo, se selecciona ADN al que se une esta proteína. Posteriormente, el ADN puede usarse para la construcción y secuenciación de la biblioteca. Es decir, este es un método para realizar una reducción de la complejidad de un modo no aleatorio dirigido a zonas funcionales específicas; en el presente ejemplo factores de transcripción específicos.

Un variante útil de la tecnología AFLP usa nucleótidos no selectivos (es decir, +0/+0 cebadores) y algunas veces se denomina PCR de ligador. Esto proporciona también una reducción de la complejidad muy adecuada.

Para una descripción adicional de AFLP, sus ventajas, sus realizaciones, así como las técnicas, enzimas, adaptadores, cebadores y compuestos y herramientas adicionales usados en el mismo, se hace referencia a los documentos US 6.045.994, EP-B-0 534 858, EP 976835 y EP 974672, WO01/88189 y Vos *et al*. Nucleic Acids Research, 1995, 23, 4407-4414.

Por tanto, en una realización preferida del método de la presente invención, la reducción de la complejidad se realiza

- digiriendo la muestra de ácido nucleico con al menos una endonucleasa de restricción para fragmentarlo en fragmentos de restricción;

- ligando los fragmentos de restricción obtenidos con al menos un adaptador oligonucleotídico sintético bicatenario que tiene un extremo compatible con uno o ambos extremos de los fragmentos de restricción para producir fragmentos de restricción ligados a adaptador;

- poner en contacto dichos fragmentos de restricción ligados a adaptador con uno o más cebadores oligonucleotídicos en condiciones de hibridación; y

- amplificar dichos fragmentos de restricción ligados a adaptador mediante alargamiento del uno o más cebadores oligonucleotídicos,

en la que al menos uno de los uno o más cebadores oligonucleotídicos incluye una secuencia de nucleótidos que tiene la misma secuencia de nucleótidos que las partes terminales de las hebras en los extremos de dichos fragmentos de restricción ligados a adaptador, incluyendo los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción e incluyendo al menos parte de los nucleótidos presentes en los adaptadores, en los que, opcionalmente, al menos uno de dichos cebadores incluye en su extremo 3' una secuencia seleccionada que comprende al menos un nucleótido ubicado inmediatamente adyacente a los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción.

AFLP es un método sumamente reproducible para la reducción de la complejidad y es por tanto particularmente adecuado para el método según la presente invención.

En una realización preferida 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, cuando es importante distinguir entre secuencias derivadas de bibliotecas separadas. La incorporación de una etiqueta oligonucleotídica en un adaptador o cebador es muy conveniente ya que no se requieren etapas adicionales para etiquetar una biblioteca.

En otra realización, la etiqueta es una secuencia identificadora. Tal como se comentó anteriormente, tal secuencia identificadora puede ser de longitud variable dependiendo de la cantidad de muestras de ácido nucleico que van a compararse. Una longitud de aproximadamente 4 bases (4^4 = 256 secuencias de etiqueta diferentes posibles) es suficiente para distinguir entre el origen de un número limitado de muestras (hasta 256), aunque se prefiere que las secuencias de etiqueta difieran en más de una base entre las muestras que van a distinguirse. Según se necesite, la longitud de las secuencias de etiqueta puede ajustarse por consiguiente.

En una realización, la secuenciación se realiza sobre un soporte sólido tal como una perla (véase por ejemplo los documentos WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007 y WO 2005/003375 (todos a nombre de 454 Corporation). Tal método de secuenciación es particularmente adecuado para una secuenciación barata y eficaz de muchas muestras simultáneamente.

En una realización preferida, la secuenciación comprende las etapas de:

- aparear fragmentos ligados a adaptador con perlas, apareándose cada perla con un único fragmento ligado a 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.

En la primera etapa, se ligan adaptadores de secuenciación a fragmentos dentro de la biblioteca de combinación. Dicho adaptador de secuenciación incluye al menos una región "clave" para aparearse a una perla, una región de cebador de secuenciación y una región de cebador de PCR. Por tanto, se obtienen fragmentos ligados a adaptador.

En una etapa adicional, se aparean fragmentos ligados a adaptador con perlas, apareándose cada perla con un único fragmento ligado a adaptador. Al grupo de fragmentos ligados a adaptador, se le añaden perlas en exceso para garantizar el apareamiento de un único fragmento ligado a adaptador por perla para la mayoría de las perlas (distribución de Poisson).

En una etapa posterior, se emulsionan las perlas en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una única perla. Están presentes reactivos de PCR en los microrreactores de agua en aceite que permiten que tenga lugar una reacción de PCR dentro de los microrreactores. Posteriormente, se rompen los microrreactores, y se enriquecen las perlas que comprenden ADN (perlas positivas para ADN).

En la siguiente etapa, se cargan las perlas en pocillos, comprendiendo cada pocillo una única perla. Los pocillos son preferiblemente parte de una placa PicoTiter™ que permite la secuenciación simultánea de una gran cantidad de fragmentos.

Tras la adición de perlas que portan enzimas, se determina la secuencia de los fragmentos usando pirosecuenciación. En etapas sucesivas, se someten la placa PicoTiter y las perlas así como las perlas de enzimas en la misma a diferentes desoxirribonucleótidos en presencia de reactivos de secuenciación convencionales, y tras la incorporación de un desoxirribonucleótido se genera una señal luminosa que se registra. La incorporación del nucleótido correcto generará una señal de pirosecuenciación que puede detectarse.

Se conoce por sí misma la pirosecuenciación en la técnica y se describe entre otros en www.biotagebio.com; www.pyrosequencing.com/tab technology. La tecnología se aplica además en por ejemplo los documentos WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007 y WO 2005/003375 (todos a nombre de 454 Corporation).

El examen de alto rendimiento de la etapa k) se realiza preferiblemente mediante inmovilización de las sondas diseñadas en la etapa h) sobre una matriz, seguido por la puesta en contacto de la matriz que comprende las sondas con una biblioteca de prueba en condiciones de hibridación. Preferiblemente, la etapa de puesta en contacto se realiza en condiciones de hibridación rigurosas (véase Kennedy *et al*. (2003) Nat. Biotech.; publicado en Internet el 7 de septiembre de 2003: 1-5). Un experto en la técnica es consciente de métodos adecuados para la inmovilización de sondas sobre una matriz y de métodos de puesta en contacto en condiciones de hibridación. Se revisa la tecnología típica que es adecuada para este fin en Kennedy *et al*. (2003) Nat. Biotech.; publicado en Internet el 7 de septiembre de 2003: 1-5.

Una aplicación ventajosa particular se encuentra en la cría de cultivos poliploidales. Secuenciando cultivos poliploidales con una alta cobertura, identificando SNP y los diversos alelos y desarrollando sondas para amplificación específica de alelos, pueden realizarse avances significativos en la cría de cultivos poliploidales.

Como parte de la invención, se ha encontrado que la combinación de generación de subconjuntos seleccionados aleatoriamente usando amplificación selectiva para una pluralidad de muestras y tecnología de secuenciación de alto rendimiento presenta determinados problemas complejos que han de solucionarse para la mejora adicional del método descrito en el presente documento para la identificación eficaz y de alto rendimiento de polimorfismos. Más en detalle, se ha encontrado que cuando se combinan muestras múltiples (es decir, la primera y la segunda o adicionales) en un grupo tras realizar una reducción de la complejidad, se produce el problema de que muchos fragmentos parecen derivarse de dos muestras o, en otras palabras, se identificaron muchos fragmentos que no podían asignarse de manera única a una muestra y por tanto no podían usarse en el procedimiento de identificación de polimorfismos. Esto condujo a una reducción de la fiabilidad del método y a que pudiesen identificarse adecuadamente menos polimorfismos (SNP, indels, SSR).

Tras un análisis cuidadoso y detallado de toda la secuencia de nucleótidos de los fragmentos que no podían asignarse, se encontró que esos fragmentos contenían dos adaptadores que comprendían etiquetas diferentes y se formaron probablemente entre la generación de las muestras de complejidad reducida y el ligamiento de los adaptadores de secuenciación. El fenómeno se representa como un “etiquetado mixto”. El fenómeno representado como “etiquetado mixto”, tal como se usa en el presente documento, se refiere por tanto a fragmentos que contienen una etiqueta que se refiere al fragmento en un lado de la muestra, mientras que el lado opuesto del fragmento contiene una etiqueta que se refiere al fragmento en otra muestra. Por tanto, un fragmento parece derivarse de dos muestras (lo que no es así). Esto conduce a la identificación errónea de polimorfismos y por tanto no es deseable.

Se ha especulado que la formación de fragmentos de heterodúplex entre dos muestras se encuentra detrás de esta anomalía.

La solución a este problema se ha encontrado en un rediseño de la estrategia para la conversión de muestras de las que se reduce la complejidad a fragmentos apareados a perlas que pueden amplificarse antes de la secuenciación de alto rendimiento. En esta realización, se somete cada muestra a reducción de la complejidad y purificación opcional. Después de eso, se hacen romos los extremos de cada muestra (pulido de extremos) seguido por ligamiento del adaptador de secuenciación que puede aparearse con la perla. Se combinan entonces los fragmentos ligados a adaptador de secuenciación y se ligan a las perlas para la polimerización en emulsión y la posterior secuenciación de alto rendimiento.

Como parte adicional de esta invención, se encontró que la formación de concatámeros dificultaba la identificación apropiada de polimorfismos. Los concatámeros se han identificado como fragmentos que se forman tras la reducción de la complejidad de productos “con extremos romos” o “pulidos”, por ejemplo mediante ADN polimerasa de T4, y en lugar de ligar a los adaptadores que permiten el apareamiento a las perlas, se ligan entre sí, creando de ese modo concatámeros, es decir, un concatámero es el resultado de la dimerización de fragmentos con extremos romos.

La solución a este problema se encontró en el uso de determinados adaptadores modificados específicamente. Los fragmentos amplificados obtenidos a partir de la reducción de la complejidad contienen normalmente una proyección 3’-A debido a las características de determinadas polimerasas preferidas que no tienen actividad de corrección de lectura exonucleasa 3’-5’. La presencia de una proyección 3’-A de este tipo es también el motivo por el cual se hacen romos los extremos de los fragmentos antes del ligamiento al adaptador. Proporcionando un adaptador que

puede aparearse a una perla, conteniendo el adaptador una proyección 3'-T, se encontró que tanto el problema de las "etiquetas mixtas" como el de los concatámeros pueden solucionarse en una etapa. Una ventaja adicional del uso de estos adaptadores modificados es que la etapa de "hacer romos los extremos" convencional y la etapa de fosforilación posterior pueden omitirse.

Por tanto, en una realización preferida adicional, tras la etapa de reducción de la complejidad de cada muestra, se realiza una etapa sobre los fragmentos de restricción ligados a adaptador amplificados obtenidos a partir de la etapa de reducción de la complejidad, mediante la cual se ligan a estos fragmentos adaptadores de secuenciación, adaptadores de secuenciación que contienen una proyección 3'-T y que pueden aparearse con las perlas.

Se ha encontrado además que, cuando los cebadores usados en la etapa de reducción de la complejidad están fosforilados, la etapa de pulido de extremos (obtención de extremos romos) y la fosforilación intermedia antes del ligamiento pueden evitarse.

Por tanto, en una realización sumamente preferida de la invención, la invención se refiere a un método para identificar uno más polimorfismos, comprendiendo dicho método las etapas de:

a) proporcionar una pluralidad de muestras de ácido nucleico de interés;

b) realizar una reducción de la complejidad sobre cada una de las muestras para proporcionar una pluralidad de bibliotecas de las muestras de ácido nucleico, en el que la reducción de la complejidad se realiza

- digiriendo cada muestra de ácido nucleico con al menos una endonucleasa de restricción para fragmentarlo en fragmentos de restricción;

- ligando los fragmentos de restricción obtenidos con al menos un adaptador oligonucleotídico sintético, bicatenario que tiene un extremo compatible con uno o ambos extremos de los fragmentos de restricción para producir fragmentos de restricción ligados a adaptador;

- poniendo en contacto dichos fragmentos de restricción ligados a adaptador con uno o más cebadores oligonucleotídicos fosforilados en condiciones de hibridación; y

- amplificando dichos fragmentos de restricción ligados a adaptador mediante el alargamiento del uno o más cebadores oligonucleotídicos, en el que al menos uno del uno o más cebadores oligonucleotídicos incluye una secuencia de nucleótidos que tiene la misma secuencia de nucleótidos que las partes terminales de las hebras en los extremos de dichos fragmentos de restricción ligados a adaptador, incluyendo los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción e incluyendo al menos parte de los nucleótidos presentes en los adaptadores, en los que, opcionalmente, al menos uno de dichos cebadores incluye en su extremo 3' una secuencia seleccionada que comprende al menos un nucleótido ubicado inmediatamente adyacente a los nucleótidos implicados en la formación de la secuencia diana para dicha endonucleasa de restricción y en los que el adaptador y/o el cebador contienen una etiqueta;

c) combinar dichas bibliotecas con una biblioteca combinada;

d) ligar adaptadores de secuenciación que pueden aparearse a perlas a los fragmentos con extremos ocupados por adaptador amplificados en la biblioteca combinada, usando un adaptador de secuenciación que porta una proyección 3'-T y sometiendo los fragmentos apareados a polimerización en emulsión;

e) secuenciar al menos una parte de la biblioteca combinada;

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

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

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

i) proporcionar un ácido nucleico de muestra de prueba de interés;

j) realizar la reducción de la complejidad de la etapa b) sobre el ácido nucleico de muestra de prueba de interés para proporcionar una biblioteca de prueba del ácido nucleico de muestra de prueba;

k) someter la biblioteca de prueba a examen de alto rendimiento para identificar la presencia, ausencia o cantidad de los polimorfismos determinados en la etapa g) usando las sondas de detección diseñadas en la etapa h).

Breve descripción de los dibujos

La figura 1A muestra un fragmento según la presente invención apareado sobre una perla ("perla 454") y la secuencia de cebador usado para la amplificación previa de las dos líneas de pimiento. "Fragmento de ADN" indica el fragmento obtenido tras la digestión con una endonucleasa de restricción, "adaptador génico clave" indica un adaptador que proporciona un sitio de apareamiento para los cebadores oligonucleotídicos (fosforilados) usados para generar una biblioteca, "KRS" indica una secuencia identificadora (etiqueta), "adaptador de SEQ. de 454" indica un adaptador de secuenciación y "adaptador de PCR de 454" indica un adaptador para permitir la amplificación en emulsión del fragmento de ADN. El adaptador de PCR permite el apareamiento de la perla y la amplificación y puede contener una proyección 3'-T.

La figura 1B muestra un cebador esquemático usado en la etapa de reducción de la complejidad. Un cebador de este tipo comprende generalmente una región de sitio de reconocimiento indicada 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 extremo 3' del mismo).

La figura 2 muestra una estimación de la concentración de ADN usando electroforesis en gel de agarosa al 2%. S1 indica PSP11; S2 indica PI201234. 50, 100, 250 y 500 ng indican respectivamente 50 ng, 100 ng, 250 ng y 500 ng para estimar las cantidades de ADN de S1 y S2. Las figuras 2C y 2D muestran la determinación de la concentración de ADN usando espectrofotometría Nanodrop.

La figura 3 muestra los resultados de evaluaciones de la calidad intermedias del ejemplo 3.

La figura 4 muestra diagramas de flujo del sistema de procesamiento de datos de secuencia, es decir, las etapas tomadas desde la generación de los datos de secuenciación hasta la identificación de los supuestos SNP, SSR e indels, mediante las etapas de la eliminación de información de secuencias conocidas en el recorte y etiquetado que dan como resultado datos de secuencias recortadas que se agrupan y ensamblan para producir cóntigos y singletons (fragmentos que no pueden ensamblarse en un cóntigo) tras lo cual pueden identificarse y evaluarse supuestos polimorfismos. La figura 4B explica en mayor detalle el proceso de prospección de polimorfismos

La figura 5 aborda el problema de las etiquetas mixtas y proporciona en el panel 1 un ejemplo de una etiqueta mixta, que porta etiquetas asociadas con la muestra 1 (MS1) y la muestra 2 (MS2). El panel 2 proporciona una explicación esquemática del fenómeno. Se ligan fragmentos de restricción AFLP derivados de la muestra 1 (S1) y de la muestra 2 (S2) con adaptadores ("adaptador de Keygene") en ambos lados que portan etiquetas específicas de muestras S1 y S2. Tras la amplificación y secuenciación, los fragmentos esperados e son aquéllos con las etiquetas S1-S1 y las etiquetas S2-S2. Lo que se observa adicional e inesperadamente son también fragmentos que portan etiquetas S1-S2 o S2-S1. El panel 3 explica la causa planteada como hipótesis de la generación de etiquetas mixtas por lo cual se forman productos de heterodúplex a partir de fragmentos de las muestras 1 y 2. Posteriormente, los heterodúplex se liberan, debido a la actividad exonucleasa 3'-5' de la ADN polimerasa de T4 o Klenow, de los extremos sobresalientes en 3'. Durante la polimerización, se rellenan los huecos con nucleótidos y se introduce la etiqueta incorrecta. Esto funciona para heterodúplex de aproximadamente la misma longitud (panel superior) pero también para heterodúplex de longitud más variable. El panel 4 proporciona a la derecha el protocolo convencional que conduce a la formación de etiquetas mixtas y a la derecha el protocolo modificado.

La figura 6 aborda el problema de la formación de concatámeros, por lo cual en el panel 1 se proporciona un ejemplo típico de un concatámero, por lo cual las diversas secciones de etiqueta y adaptador están subrayadas y con su origen (es decir, MS1, MS2, ES1 y ES2 correspondientes respectivamente 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 demuestra los fragmentos esperados que portan etiquetas S1-S1 y etiquetas S2-S2 y los observados pero inesperados S1-S1-S2-S2, que son un concatámero de fragmentos de la muestra 1 y de la muestra 2. Solución del panel 3 para evitar la generación de concatámeros así como etiquetas mixtas introduciendo una proyección en los adaptadores AFLP, adaptadores de secuenciación modificados y la omisión de la etapa de pulido de extremos cuando se ligan adaptadores de secuenciación. No se encuentra formación de concatámeros porque los fragmentos AFLP no pueden ligarse entre sí y no se producen fragmentos mixtos ya que se omite la etapa de pulido de extremos. El panel 4 proporciona el protocolo modificado usando adaptadores modificados para evitar la formación de concatámeros así como etiquetas mixtas.

Figura 7. Alineación múltiple "10037 C1a989contig2" de secuencias de fragmentos AFLP de pimiento, que contienen un supuesto polimorfismo de nucleótido único (SNP). Obsérvese que el SNP (indicado mediante la flecha negra) está definido por un alelo A presente en ambas lecturas de la muestra 1 (PSP11), indicado 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 inferiores. Los nombres de las lecturas se muestras a la izquierda. La secuencia consenso de esta alineación múltiple es (5'- 3'):

TAACACGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACA[A/G]TGTTGGTTTT GGTGCTAACTTCAACCCCACTACTGTTTTGCTCTATTTTTG.

Figura 8A. Representación esquemática de la estrategia de enriquecimiento para seleccionar como diana repeticiones de secuencias simples (SSR) en combinación con secuenciación de alto rendimiento para el descubrimiento de SSR *de novo* SSR.

Figura 8B: Validación de un SNP G/A en pimiento usando detección SNPWave. P1 = PSP11; P2 = PI201234. Se indican ocho descendientes RIL mediante los números 1-8.

Ejemplos

Ejemplo 1

Se generó una mezcla de ligamiento por restricción con EcoRI/MseI (1) a partir de ADN genómico de las líneas de pimiento PSP-11 y PI20234. Se diluyó 10 veces la mezcla de ligamiento por restricción y se preamplificaron 5 microlitros de cada muestra (2) con los 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%. Se diluyeron 20 veces los productos de preamplificación, seguido por una preamplificación de KRSEcoRI +1(A) y KRSMseI +2(CA) AFLP. Las secciones KRS (identificador) están subrayadas y los nucleótidos selectivos están en negrita en el extremo 3' en la secuencia de cebador SEQ ID 1-4 a continuación. 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 huella de EcoRI +3(A) y MseI +3(C) (3) AFLP (4). Se purificaron por separado los productos de preamplificación de las dos líneas de pimiento en una columna QiagenPCR (5). Se midió la concentración de las muestras en el instrumento nanodrop. Se mezclaron y secuenciaron un total de 5006,4 ng de PSP-11 y 5006,4 ng de PI20234.

Conjunto I de cebadores usado para la preamplificación de PSP-11

E01LKRS1 5'-CGTCAGACTGCGTACCAATTCA-3' [SEQ ID 1]

M15FCRRS1 5'-TGGTGATGAGTCCTGAGTAACA-3' [SEQ ID 2]

Conjunto II de cebadores usado para la preamplificación de PI20234

E01LKRS2 5'-CAAGAGACTGCGTACCAATTCA-3' [SEQ ID 3]

M15KKRS2 5'-AGCCGATGAGTCCTGAGTAACA-3' [SEQ ID 4]

(1) Mezcla de ligamiento por restricción con EcoRI/MseI

Mezcla de restricción (40 ul/muestra)

ADN	6 µl (±6300 ng)
ECoRI (5U)	0,1 µl
MseI (2U)	0,05 µl
5xRL	8 µl
MQ	25,85 µl
Total	40 µl

Incubación durante 1 h a 37ºC

Adición de:

Mezcla de ligamiento (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

Incubación durante 3 h a 37ºC

Adaptador EcoRI

91M35/91M36:	*-CTCGTAGACTGCGTACC	:91M35 [SEQ ID 5]
± bio	CATCTGACGCATGGTTAA	:91M36 [SEQ ID 6]

Adaptador MseI

92A18/92A19:	5-GACGATGAGTCCTGAG-3	:92A18 [SEQ ID 7]
	3-TACTCAGGACTCAT-5	:92A19 [SEQ ID 8]

(2) Preamplificación

Preamplificación (A/C):

Mezcla RL (10x)	5 µl
EcoRI-pr E01L(50 ng/ul)	0,6 µl
MseI-pr M02K(50 ng/ul)	0,6 µl
dNTP (25 mM)	0,16 µl
Taq.pol.(5U)	0,08 µl
10XPCR	2,0 µl
MQ	11,56 µl
Total	20 ml/reacción

Perfil térmico de la preamplificación

Se realizó una preamplificación selectiva en un volumen de reacción de 50 µl. Se realizó la PCR en un instrumento PE GeneAmp PCR System 9700 y se inició un perfil de 20 ciclos con una etapa de desnaturalización de 94ºC durante 30 segundos, seguido por una etapa de apareamiento de 56ºC durante 60 segundos y una etapa de extensión de 72ºC durante 60 segundos.

EcoRI +1(A)[1]	
E01 L	92R11: 5-AGACTGCGTACCAATTCA-3 [SEQ ID 9]
MseI +1(C)[1]	
M02k	93E42: 5-GATGAGTCCTGAGTAAC-3 [SEQ ID 10]

Preamplificación A/CA:

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

Se realizó la preamplificación selectiva en un volumen de reacción de 50 µl. Se realizó la PCR en un instrumento PE GeneAmp PCR System 9700 y se inició un perfil de 30 ciclos con una etapa de desnaturalización de 94ºC durante 30 segundos, seguido por una etapa de apareamiento de 56ºC durante 60 segundos y una etapa de extensión de 72ºC durante 60 segundos.

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

05F212	E01LKRS1	CGTCAGACTGCGTACCAATTCA	-3' [SEQ ID 11]
05F213	E01LKRS2	CAAGAGACTGCGTACCAATTCA	-3' [SEQ ID 12]
05F214	M15KKRS1	TGGTGATGAGTCCTGAGTAA**CA**	-3' [SEQ ID 13]
05F215	M15KKRS2	AGCCGATGAGTCCTGAGTAA**CA**	-3' [SEQ ID 14]

nucleótidos selectivos en negrita y etiquetas (KRS) subrayadas
Muestra PSP11: E07.LKRS1/M15KKR**S1**
Muestra PI120234: E01LKRS2/M15KKR**S2**

(4) Protocolo de AFLP

Se realizó la amplificación selectiva en un volumen de reacción de 20 µl. Se realizó la PCR en un instrumento PE GeneAmp PCR System 9700. Se inició un perfil de 13 ciclos con una etapa de desnaturalización de 94ºC durante 30 segundos, seguido por una etapa de apareamiento de 65ºC durante 30 segundos, con una fase de descenso en la que la temperatura de apareamiento se disminuyó 0,7ºC en cada ciclo, y una etapa de extensión de 72ºC durante 60 segundos. A este perfil le siguió un perfil de 23 ciclos con una etapa de desnaturalización de 94ºC durante 30 segundos, seguido por una etapa de apareamiento de 56ºC durante 30 segundos y una etapa de extensión de 72ºC durante 60 segundos.

EcoRI	+3(**AAC**) y MseI +(**CAG**)	
E32	92SO2: 5-GACTGCGTACCAATTC**ACC**-3	[SEQ ID 15]
M49	92G23: 5-GATGAGTCCTGAGTAA**CAG**-3	[SEQ ID 16]

(5) Columna Qiagen

Se realizó la purificación Qiagen según las instrucciones del fabricante: manual de QIAquick® Spin (http://wwwl.qiagen.com/literature/handbooks/PDF/DNACleanupAndConcentration/QQ Spin/1021422 HBQQSpin 072002WW.pdf)

Ejemplo 2: PIMIENTO

Se usó ADN de las líneas de pimiento PSP-11 y PI20234 para generar producto de AFLP mediante el uso de cebadores específicos de *sitios de reconocimiento Keygene* de AFLP. (Estos cebadores de AFLP son esencialmente iguales a cebadores de AFLP convencionales, por ejemplo descritos en el documento EP 0 534 858, y contendrán generalmente una *región de sitio de reconocimiento,* una región constante y uno o más nucleótidos selectivos en una región selectiva. A partir de las líneas de pimiento PSP-11 o 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 por inactivación durante 10 minutos a 80ºC. Se ligaron los fragmentos de restricción obtenidos con adaptador oligonucleotídico sintético bicatenario, un extremo del cual es compatible con uno o ambos de los extremos de los fragmentos de restricción de EcoRI y/o MseI. Se realizaron reacciones de preamplificación de AFLP (20 µl/reacción) con cebadores de AFLP +1/+1 sobre una mezcla de ligamiento por restricción diluida 10 veces. Perfil de PCR: 20*(30 s a 94ºC + 60 s a 56ºC + 120 s a 72ºC). Se realizaron reacciones de AFLP adicionales (50 µl/reacción) con diferentes cebadores específicos de sitios de reconocimiento Keygene de AFLP +1 *EcoRI y* +2 *MseI* (tabla a continuación, las etiquetas están en negrita, los nucleótidos selectivos están subrayados) sobre un producto de preamplificación de AFLP +1/+1 EcoRI/MseI diluido 20 veces. Perfil de PCR: 30*(30 s a 94ºC + 60 s a 56ºC + 120 s a 72ºC). Se purificó el producto de AFLP usando el kit de purificación de PCR QIAquick (QIAGEN) siguiendo el manual de QIAquick® Spin 07/2002 página 18 y se midió la concentración con un espectrofotómetro NanoDrop® ND-1000. Se reunió un total de 5 µg de producto de AFLP +1/+2 PSP-11 y 5 µg de producto de AFLP +1/+2 PI20234 AFLP y se solubilizó en 23,3 µl de TE. Finalmente, se obtuvo una mezcla con una concentración de producto de AFLP +1/+2 de 430 ng/µl.

Tabla

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

Ejemplo 3: Maíz

Se usó ADN de las líneas de maíz B73 y M017 para generar producto de AFLP mediante el uso de cebadores específicos de *sitios de reconocimiento Keygene* de AFLP. (Estos cebadores de AFLP son esencialmente iguales a cebadores de AFLP convencionales, por ejemplo descritos en el documento EP 0 534 858, y contendrán generalmente una *región de sitio de reconocimiento,* una región constante y uno o más nucleótidos selectivos en el extremo 3' de los mismos).

Se digirió ADN las líneas de 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 por inactivación durante 10 minutos a 80ºC. Se ligaron los fragmentos de restricción obtenidos con adaptador oligonucleotídico sintético bicatenario, un extremo del cual es compatible con uno o ambos de los extremos de los fragmentos de restricción de TaqI y/o MseI.

Se realizaron reacciones de preamplificación de AFLP (20 µl/reacción) con cebadores de AFLP +1/+1 sobre una mezcla de ligamiento por restricción diluida 10 veces. Perfil de PCR: 20*(30 s a 94ºC + 60 s a 56ºC + 120 s a 72ºC). Se realizaron reacciones de AFLP adicionales (50 µl/reacción) con diferentes cebadores de sitios de reconocimiento Keygene de AFLP +2 *TaqI* y *MseI* (tabla a continuación, las etiquetas están en negrita, los nucleótidos selectivos están subrayados) sobre un producto de preamplificación de AFLP +1/+1 TaqI/MseI diluido 20 veces. Perfil de PCR: 30*(30 s a 94ºC + 60 s a 56ºC + 120 s a 72ºC). Se purificó el producto de AFLP usando el kit de purificación de PCR QIAquick (QIAGEN) siguiendo el manual de QIAquick® Spin 07/2002 página 18 y se midió la concentración con un espectrofotómetro NanoDrop® ND-1000. Se reunió un total de 1,25 µg de cada producto de AFLP B73 +2/+2 diferente y 1,25 µg de cada producto de AFLP M017 +2/+2 diferente y se solubilizó en 30 µl de TE. Finalmente, se obtuvo una mezcla con una concentración de producto de AFLP +2/+2 de 333 ng/µl.

Tabla

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

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

SECUENCIACIÓN POR 454

Se procesaron muestras de fragmentos de AFLP de pimiento y maíz preparadas tal como se describió anteriormente por 454 Life Sciences tal como se describe (Margulies *et al*., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437 (7057):376-80. Epub 31 de julio de 2005).

PROCESAMIENTO DE DATOS

Sistema de procesamiento:

Datos de entrada:

Se recibieron datos de secuencias sin procesar para cada ronda:

- 200.000 – 400.000 lecturas

- puntuaciones de calidad de la lectura automática de bases

Recorte y etiquetado

Se analizaron estos datos de secuencias para detectar la presencia de sitios de reconocimiento Keygene (KRS) al comienzo y al final de la lectura. Estas secuencias de KRS consisten tanto en una secuencia de adaptador de AFLP como en una secuencia de marcador de muestra y son específicas para una determinada combinación de cebadores de AFLP en una determinada muestra. Las secuencias de KRS se identifican mediante BLAST y se recortan y se restauran los sitios de restricción. Las lecturas se marcan con una etiqueta para la identificación del origen del KRS. Se seleccionan secuencias recortadas según su longitud (mínimo de 33 nt) para que participen en el procesamiento adicional.

Agrupación y ensamblaje

Se realiza un análisis *MegaBlast* sobre todas las lecturas recortadas, seleccionadas por tamaño para obtener agrupaciones de secuencias homólogas. Consecutivamente, se ensamblan todas las agrupaciones con *CAP3* para dar como resultado cóntigos ensamblados. A partir de ambas etapas, se identifican lecturas de secuencias únicas que no coinciden con ninguna otra lectura. Estas lecturas se marcan como singletons.

El sistema de procesamiento que lleva a cabo las etapas descritas en el presente documento anteriormente se muestra en la figura 4A

Prospección de polimorfismos y evaluación de la calidad

Los cóntigos resultantes del análisis de ensamblaje forman la base de la detección de polimorfismos. Cada “apareamiento erróneo” en la alineación de cada agrupación es un posible polimorfismo. Se definen criterios de selección para obtener una puntuación de calidad:

- número de lecturas por cóntigo

- frecuencia de “alelos” por muestra

- aparición de secuencia de homopolímero

- aparición de polimorfismos vecinos

Se identifican SNP e indels con una puntuación de calidad por encima del umbral como supuestos polimorfismos. Para la prospección de SSR, se usa la herramienta MISA (identificación de MIcroSatélites) (http://pgrc.ipk-gatersleben.de/misa). Esta herramienta identifica di, tri, tetranucleótidos y motivos SSR de compuestos con criterios predefinidos y resume las apariciones de estas SSR.

La prospección de polimorfismos y el procedimiento de asignación de calidad se muestran en la figura 4B.

Resultados

La tabla a continuación resume los resultados del análisis combinado de secuencias obtenidas a partir de 2 rondas de secuenciación de 454 para las muestras de pimiento combinadas y 2 rondas para las muestras de maíz combinadas.

	Pimiento	Maíz
Número total de lecturas	457178	492145
Número de lecturas recortadas	399623	411008
Número de singletons	105253	313280
Número de cóntigos	31863	14588
Número de lecturas en cóntigos	294370	97728
Número total de secuencias que contienen SSR	611	202
Número de secuencias que contienen SSR diferentes	104	65
Número de motivos SSR diferentes (di, tri, tetra y compuesto)	49	40

Número de SNP con puntuación Q ≥ 0,3 *	1636	782
Número de indels *	4090	943
* ambos con selección frente a SNP vecinos, secuencia flanqueante de al menos 12 pb y sin producirse en secuencias de homopolímeros mayores de 3 nucleótidos.		

Ejemplo 4. Descubrimiento de polimorfismos de nucleótido único (SNP) en pimiento.

Aislamiento de ADN

Se aisló ADN genómico a partir de las dos líneas originales de una población consanguínea recombinante (RIL) de pimiento y 10 progenies RIL. Las líneas originales son PSP11 y PI201234. Se aisló ADN genómico de material de hoja de plántulas individuales usando un procedimiento CTAB modificado descrito por Stuart y Via (Stuart, C.N., Jr y Via, L.E. (1993) A rapid CTAB ADN isolation technique useful for RAPD fingeprinting and other PCR applications. Biotechniques, 14, 748-750). 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 moldes de AFLP usando cebadores de AFLP etiquetados

Se prepararon moldes de AFLP de las líneas originales de pimiento PSP11 y PI201234 usando la combinación de endonucleasas de restricción *EcoRI*/*MseI* tal como se describe por Zabeau & Vos, 1993: Selective restriction fragment amplification; a general method for ADN fingerprinting. Documento EP 0534858-Al, B1; patente estadounidense 6045994) y Vos *et al* (Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. *et al.* (1995) AFLP: a new technique for ADN fingerprinting. Nucl. Acids Res., 21, 4407-4414).

Específicamente, se llevó a cabo la restricción de ADN genómico con EcoRI y MseI tal como sigue:

Restricción del ADN

ADN	100-500 ng
EcoRI	5 unidades
MseI	2 unidades
tampón 5xRL	8 µl

Agua MilliQ hasta 40 µl

La incubación fue durante 1 hora a 37ºC. Tras la restricción con enzimas, se inactivaron las enzimas mediante incubación durante 10 minutos a 80ºC.

Ligamiento de adaptadores

ATP 10 mM	1 µl
ADN ligasa de T4	1 µl
adaptador EcoRI (5 pmol/µl)	1 µl
adaptador MseI (50 pmol/µl)	1 µl
tampón 5xRL.	2 µl

Agua MilliQ hasta 40 µl.

La incubación fue durante 3 horas a 37ºC.

Amplificación AFLP selectiva

Tras la restricción-ligamiento, se diluyó la reacción de restricción/ligamiento 10 veces con $T_{10}E_{0,1}$ y se usaron 5 µl de mezcla diluida como molde en una etapa de amplificación selectiva. Obsérvese que puesto que se pretendía una amplificación selectiva +1/+2, se realizó en primer lugar una etapa de preamplificación selectiva +1/+1 (con cebadores de AFLP convencionales). Las condiciones de reacción de la amplificación +1/+1 (+A/+C) fueron las siguientes.

Mezcla de restricción-ligamiento (diluida 10 veces)	5 µl
Cebador EcoRI +1 (50 ng/µl):	0,6 µl

cebador MseI +1 (50 ng/µl)	0,6 µl
dNTP (20mM)	0,2 µl
Taq polimerasa (5 U/µl Amplitaq, PE)	0,08 µl
Tampón de PCR 10X	2,0 µl
Agua MilliQ hasta	20 µl

Las secuencias de los cebadores eran:

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

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

Se realizaron las amplificaciones por PCR usando un instrumento PE9700 con un bloque de oro o plata usando las siguientes condiciones: 20 veces (30 s a 94ºC, 60 s a 56ºC y 120 s a 72ºC).

Se comprobó la calidad de los productos de preamplificación +1/+1 generados en un gel de agarosa al 1% usando un marcador de tamaño molecular de 100 pares de bases y un marcador de tamaño molecular de 1 Kb para comprobar la distribución de longitud de fragmentos. Tras la amplificación selectiva +1/+1, se diluyó la reacción 20 veces con $T_{10}E_{0,1}$ y se usaron 5 µl de mezcla diluida como molde en la etapa de amplificación selectiva +1/+2 usando cebadores de AFLP etiquetados.

Finalmente, se realizaron amplificaciones AFLP selectivas +1/+2 (A/+CA):

Producto de amplificación selectiva +1/+1 (diluido 20 veces)	5,0 pl
Cebador KRS EcoRI +A (50 ng/µl)	1,5 µl
Cebador KRS MseI + CA (50 ng/µl)	1,5 µl
dNTP (20 mM)	0,5 µl
Taq polimerasa (5U/µl Amplitaq, Perkin Elmer)	0,2 µl
Tampón de PCR 10X	5,0 µl
MQ hasta	50 µl

Las secuencias de los cebadores de AFLP etiquetados eran:

PSP11:

O5F212: EcoRI+1: 5'-CGTCAGACTGCGTACCAATTCA-3' [SEQ ID 1] y

O5F214: MseI+2: 5'-TGGTGATGAGTCCTGAGTAACA-3' [SEQ ID 2]

PI201234:

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

05F215: MseI+1: 5'-AGCCGATGAGTCCTGAGTAACA-3' [SEQ ID 4]

Obsérvese que estos cebadores contienen etiquetas de 4 pb (subrayadas anteriormente) en sus extremos 5 prima para distinguir los productos de amplificación que se originan a partir de las respectivas líneas de pimiento al final del procedimiento de secuenciación.

Representación esquemática de productos de amplificación AFLP +1/+2 de pimiento tras la amplificación con cebadores de AFLP que contienen secuencias de etiqueta en 5 prima de 4 pb.

	Etiqueta EcoRI	Etiqueta MseI
PSP 11:	5'-CGTC	ACCA-3'
	3'_GCAG	TGGT-5'
PI201234	5'-CAAG	GGCT-3'
	3'-GTTC	CCGA-5'

Se realizaron las amplificaciones por PCR (24 por muestra) usando un instrumento PE9700 con un bloque de oro o plata usando las siguientes condiciones: 30 veces (30 s a 94 ºC + 60 s a 56 ºC + 120 s a 72ºC).

Se comprobó la calidad de los productos de amplificación generados en un gel de agarosa al 1% usando un marcador de tamaño molecular de 100 pares de bases y un marcador de tamaño molecular de 1 Kb para comprobar

la distribución de longitud de fragmentos.

Purificación y cuantificación de la reacción de AFLP

Tras agrupar dos reacciones de AFLP selectivas +1/+2 de 50 microlitros por muestra de pimiento, se purificaron los 12 productos de reacción de AFLP de 100 µl resultantes usando el kit de purificación de PCR QIAquick (QIAGEN), siguiendo el manual QIAquick® Spin (página 18). En cada columna, se cargó un máximo de 100 µl de producto. Se eluyeron los productos amplificados 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 instrumento Nanodrop (figura 2).

Se usaron las mediciones de la concentración con 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 secuencias de 454.

Preparación de la biblioteca de secuencias y secuenciación de alto rendimiento

Se sometieron productos de amplificación mezclados de ambas líneas de pimiento a secuenciación de alto rendimiento usando la tecnología de secuenciación de 454 Life Sciences tal como se describe por Margulies *et al.*, (Margulies *et al.*, Nature 437, págs. 376-380 y suplementos en Internet). Específicamente, en primer lugar se pulieron los extremos de los productos de PCR de AFLP y posteriormente se ligaron a adaptadores para facilitar la amplificación por PCR en emulsión y la secuenciación de fragmentos posterior tal como se describe por Margulies y colaboradores. Las secuencias de adaptadores de 454, los cebadores de PCR en emulsión, la secuencia-cebadores y las condiciones de rondas de secuenciación se describen todos por Margulies y colaboradores. El orden lineal de elementos funcionales en un fragmento de PCR en emulsión amplificado en perlas de Sepharose en el procedimiento de secuenciación de 454 fue tal como sigue tal como se muestra a modo de ejemplo en la figura 1A:

Adaptador de PCR de 454 – adaptador de secuencia de 454 – etiqueta 1 de cebador de AFLP de 4 pb – secuencia 1 de cebador de AFLP incluyendo nucleótido(s) selectivo(s) – secuencia interna de fragmento de AFLP – secuencia 2 de cebador de AFLP incluyendo nucleótido(s) selectivo(s), etiqueta 2 de cebadores de AFLP de 4 pb – adaptador de secuencia de 454 – adaptador de PCR de 454 – perla de Sepharose

Se realizaron dos rondas de secuenciación de 454 de alto rendimiento por 454 Life Sciences (Branford, CT; Estados Unidos de América).

Procesamiento de datos de secuenciación de 454.

Se procesaron datos de secuencias que resultan de 2 rondas de secuenciación de 454 usando un sistema bioinformático (Keygene N.V.). Específicamente, se convirtieron lecturas de secuencias con lectura automática de bases 454 sin procesar en formato FASTA y se inspeccionaron para determinar la presencia de secuencias de adaptadores de AFLP etiquetados usando un algoritmo BLAST. Con coincidencias de alta confianza con las secuencias de cebadores de AFLP etiquetados conocidos, se recortaron las secuencias, se restauraron los sitios de endonucleasas de restricción y se asignaron las etiquetas apropiadas (muestra 1 EcoRI (ES1), muestra 1 MseI (MS1), muestra 2 EcoRI (ES2) o muestra 2 MseI (MQ2), respectivamente). A continuación, se agruparon todas las secuencias recortadas mayores de 33 bases usando un procedimiento megaBLAST basándose en homologías de secuencia global. A continuación, se ensamblaron agrupaciones en uno o más cóntigos y/o singletons por agrupación, usando un algoritmo de alineación múltiple CAP3. Se inspeccionaron cóntigos que contenían más de una secuencia para detectar los apareamientos erróneos de secuencias, que representan supuestos polimorfismos. Se asignaron a los apareamientos erróneos de secuencias puntuaciones de calidad basándose en los siguientes criterios:

* los números de lecturas en un cóntigo
* la distribución de alelos observada

Los dos criterios anteriores forman la base para la denominada puntuación Q asignada a cada supuesto SNP/indel. Las puntuaciones Q oscilan entre 0 y 1; una puntuación Q de 0,3 sólo puede alcanzarse en el caso de que ambos alelos se observen al menos dos veces.

* ubicación en homopolímeros de una determinada longitud (ajustable; parámetro por defecto para evitar polimorfismos ubicados en homopolímeros de 3 bases o mayores).

* número de cóntigos en la agrupación.

* distancia hasta los apareamientos erróneos de secuencias vecinas más cercanos (ajustable; importante para determinados tipos de ensayos de genotipado que estudian con sonda secuencias flanqueantes).

* el nivel de asociación de alelos observados con la muestra 1 o la muestra 2; en el caso de una asociación perfecta, constante entre los alelos de un supuesto polimorfismo y las muestras 1 y 2, el polimorfismo (SNP) se indica como un supuesto polimorfismo (SNP) de “élite”. Se cree que un polimorfismo de élite tiene una alta probabilidad de estar ubicado en una secuencia del genoma de copia única o con bajo número de copias en el caso de que se hayan usado dos líneas homocigotas en el procedimiento de descubrimiento. A la inversa, una asociación débil de un polimorfismo con un origen de muestra conlleva un alto riesgo de haber descubierto polimorfismos falsos que surgen de la alineación de secuencias no alélicas en un cóntigo.

Se identificaron secuencias que contienen motivos SSR usando la herramienta de búsqueda MISA (herramienta de identificación de MIcroSatélites; disponible de http://pgrc.ipk-gatersleben.de/misa/

Se muestra la estadística global de la ronda en la tabla a continuación.

Tabla. Estadística global de una ronda de secuenciación de 454 para el descubrimiento de SNP en pimiento

Combinación de enzimas	Ronda
Recorte	
Todas las lecturas	254308
Erróneas	5293 (2 %)
Correctas	249015 (98%)
Concatámeros	2156 (8,5 %)
Etiquetas mixtas	1120 (0,4 %)
Lecturas correctas	
Recortado un extremo	240817 (97%)
Recortados ambos extremos	8198 (3 %)
Número de lecturas de la muestra 1	136990 (55%)
Número de lecturas de la muestra 2	112025 (45 %)
Agrupamiento	
Número de cóntigos	21918
Lecturas en cóntigos	190861
Número promedio de lecturas por cóntigo	8,7
Prospección de SNP	
SNP con puntuación Q ≥ 0,3 *	1483
Indel con puntuación Q ≥ 0,3 *	3300
Prospección de SSR	
Número total motivos SSR identificados	359
Número de lecturas que contienen uno o más motivos SSR	353
Número de motivos SSR con tamaño unitario 1 (homopolímero)	0
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 prospección de SNP / indels fueron los siguientes: sin polimorfismos vecinos con puntuación Q mayor de 0,1 dentro de 12 bases en cualquier lado, no presentes en homopolímero de 3 o más bases. Los criterios de prospección no tuvieron en cuenta la asociación constante con la muestra 1 y 2, es decir, los SNP e indels no son necesariamente supuestos SNP/indels de élite.	

Se muestra un ejemplo de una alineación múltiple que contiene un supuesto polimorfismo de nucleótido único de

élite en la figura 7.

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

Con el fin de validar el supuesto SNP A/G identificado en el ejemplo 1, se diseñó un ensayo de sitio etiquetado de secuencia (STS) para este SNP usando cebadores de PCR flanqueantes. Las secuencias de los cebadores de PCR eran las siguientes:

cebador 1.2f: 5'-AAACCCAAACTCCCCCAATC-3', [SEQ ID 37] y

cebador 1.2r: 5'-AGCGGATAACAATTTCACACAGGACATCAGTAGTCACACTGGTACAAAAATAGAGCAAAACAGTAGTG-3' [SEQ ID 38]

Obsérvese que el cebador 1.2r contenía un sitio de unión a cebador de secuencia M13 y un fragmento de relleno de longitud en su extremo 5 prima. Se llevó a cabo la amplificación por PCR usando productos de amplificación AFLP +A/+CA de PSP11 y PI210234 preparados tal como se describió en el ejemplo 4 como molde. Las condiciones de PCR fueron las siguientes:

Durante 1 reacción de PCR se mezclaron los siguientes componentes:

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

5 µl de cebador 1.2f 1 pmol/µl (diluido directamente a partir de una disolución madre 500 µM)

5 µl de cebador 1.2r 1 pmol/µl (diluido directamente a partir de una disolución madre 500 µM)

5 µl de mezcla de PCR
- 2 ml de tampón de PCR 10 x
- 1 ml de dNTP 5 mM
- 1,5 ml de $MgCl_2$ 5 mM
- 0,5 ml de H_2O

5 µl de mezcla de enzimas
- 0,5 ml de tampón de PCR 10 x (Applied Biosystems)
- 0,1 ml de ADN polimerasa AmpliTaq 5 U/ml (Applied Biosystems)
- 4,4 ml de H_2O

Se usó el siguiente perfil de PCR:

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

Se clonaron los productos de PCR en el vector pCR2.1 (kit de clonación TA; Invitrogen) usando el método de clonación TA y se transformaron en células de *E. coli* competentes INVαF'. Se sometieron los transformantes a selección azul/blanca. Se seleccionaron tres transformantes blancos independientes cada uno para PSP11 y PI-201234 y se hicieron crecer durante la noche en medio líquido selectivo para el aislamiento de plásmidos.

Se aislaron plásmidos usando el kit QIAprep Spin Miniprep (QIAGEN). Posteriormente, se secuenciaron los insertos de estos plásmidos según el protocolo a continuación y se resolvieron en el instrumento MegaBACE 1000 (Amersham). Se inspeccionaron las secuencias obtenidas para detectar la presencia del alelo de SNP. Dos plásmidos independientes que contenían el inserto PI-201234 y 1 plásmido que contenía el inserto PSP11 contenían la secuencia consenso esperada que flanquea al SNP. La secuencia derivada del fragmento PSP11 contenía el alelo A esperado (subrayado) y la secuencia derivada del fragmento PI-201234 contenía el alelo G esperado (doblemente subrayado):

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

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA CCCCACTACTGTTTTGCTCTATTTTTGT [SEQ ID 39]

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

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAGTGTTGGTTTTGGTGCTAACTTCAA CCCCACTACTGTTTTGCTCTATTTTTG [SEQ ID 40]

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

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAGTGTTGGTTTTGGTGCTAACTTCAA CCCCACTACTGTTTTGCTCTATTTTTG [SEQ ID 41]

Este resultado indica que el supuesto SNP A/G de pimiento representa un polimorfismo genético verdadero detectable usando el ensayo de STS diseñado.

Ejemplo 6: Validación de SNP mediante detección SNPWave.

Con el fin de validar el supuesto SNP A/G identificado en el ejemplo 1, se definieron conjuntos de sondas de ligamiento SNPWave para ambos alelos de este SNP usando la secuencia consenso. Las secuencias de las sondas de ligamiento eran las siguientes:

Secuencias de sondas SNPWave (5'-3'):

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

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

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

Obsérvese que las sondas específicas de alelos 06A162 y 06A163 para los alelos A y G, respectivamente, difieren en 2 bases en tamaño, de manera que tras el ligamiento a la sonda específica de locus común 06A164, resultan tamaños de productos de ligamiento de 94 (42+54) y 96 (44+52) bases.

Se llevaron a cabo reacciones de PCR y ligamiento SNPWave tal como se describe por 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. (2004). SNPWave™ a flexible multiplexed SNP genotyping technology. Nucleic Acids Research 32: e47), usando 100 ng de ADN genómico de las líneas de pimiento PSP11 y PI201234 y 8 descendientes RIL como material de partida. Las secuencias de los cebadores de PCR eran:

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

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

Tras la amplificación por PCR, la purificación de producto de PCR y la detección en el instrumento MegaBACE1000 fueron tal como se describe por van Eijk y colaboradores (véase anteriormente). Se muestra en la figura 8B una imagen de pseudo-gel de los productos de amplificación obtenidos a partir de PSP11, PI201234 y 8 descendientes RIL.

Los resultados de SNPWave demuestran claramente que se detecta el SNP A/G mediante el ensayo SNPWave, dando como resultado productos de 92 pb (=AA genotipo homocigoto) para P1 (PSP11) y descendientes RIL 1, 2, 3, 4, 6 y 7), y productos de 94 pb (=GG genotipo homocigoto) para P2 (PI201233) y descendientes RIL 5 y 8.

Ejemplo 7: Estrategias para enriquecer bibliotecas de fragmentos de AFLP en secuencias con bajo número de copias.

Este ejemplo describe varios métodos de enriquecimiento para seleccionar como diana un bajo número de copias de secuencias de genoma únicas con el fin de aumentar el rendimiento de polimorfismos de élite tal como se describe en el ejemplo 4. Los métodos pueden dividirse en cuatro categorías:

1) métodos dirigidos a preparar ADN genómico de alta calidad, excluyendo secuencias de cloroplastos.

En este caso se propone preparar ADN nuclear en lugar de ADN genómico completo tal como se describe en el ejemplo 4, para excluir el aislamiento conjunto de ADN de cloroplastos abundante, lo que puede dar como resultado una reducción del número de secuencias de ADN genómico vegetal, dependiendo de las endonucleasas de

restricción y los cebadores de AFLP selectivos usados en el procedimiento de preparación de la biblioteca de fragmentos. Se ha descrito un protocolo para el aislamiento de ADN nuclear de tomate sumamente puro por Peterson, DG., Boehm, K.S. & Stack S. M. (1997). Isolation of Milligram Quantities of Nuclear ADN From Tomato (*Lycopersicon esculentum*), A Plant Containing High Levels of Polyphenolic Compounds. Plant Molecular Biology Reporter 15 (2), páginas 148-153.

2) Métodos dirigidos a usar endonucleasas de restricción en el procedimiento de preparación de moldes de AFLP que se espera que produzcan niveles elevados de secuencias con bajo número de copias.

En este caso se propone usar determinadas endonucleasas de restricción en el procedimiento de preparación de moldes de AFLP, que se espera que seleccionen como diana secuencias de genoma únicas o con bajo número de copias, dando como resultado bibliotecas de fragmentos enriquecidas en polimorfismos con aumento de la capacidad para poder convertirse en ensayos de genotipado. Un ejemplo de una endonucleasa de restricción que selecciona como diana una secuencia con bajo número de copias en genomas vegetales es PstI. Otras endonucleasas de restricción sensibles a metilación pueden seleccionar como diana también secuencias de genoma únicas o con bajo número de copias preferentemente.

3) Métodos dirigidos a eliminar selectivamente secuencias sumamente duplicadas basándose en la cinética de reapareamiento de secuencias repetidas frente a secuencias con bajo número de copias.

En este caso se propone eliminar selectivamente secuencias sumamente duplicadas (repeticiones) o bien de la muestra de ADN genómico total o bien del material de molde de (ADNc-)AFLP antes de la amplificación selectiva.

3a) La preparación de ADN High-C_0t es una técnica comúnmente usada para enriquecer secuencias con bajo número de copias que se aparean lentamente a partir de una mezcla de ADN genómico vegetal compleja (Yuan *et al*. 2003; High-Cot sequence analysis of the maize genome. Plant J. 34: 249-255). Se sugiere tomar High-C_0t en lugar de ADN genómico total como material de partida para enriquecer polimorfismos ubicados en secuencias con bajo número de copias.

3b) Una alternativa a la laboriosa preparación de high-C_0t puede ser incubar ADNbc de reapareamiento y desnaturalizado con una nucleasa novedosa del cangrejo de Kamchatka, que escinde dúplex de ADN cortos, perfectamente coincidentes a una tasa superior que dúplex de ADN no perfectamente coincidentes, tal como se describe por Zhulidov y colaboradores (2004; Simple cDNA normalization using Kamchatka crab duplex-specific nuclease. Nucleic Acids Research 32, e37) y Shagin y colaboradores (2006; a novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. Genome Research 12: 1935-1942). Específicamente, se propone incubar mezclas de restricción/ligamiento de AFLP con esta endonucleasa para eliminar de la mezcla secuencias sumamente duplicadas, seguido por amplificación AFLP selectiva de las secuencias de genoma únicas o con bajo número de copias restantes.

3c) La metil-filtración es un método para enriquecer fragmentos de ADN genómico hipometilados usando la endonucleasa de restricción McrBC que corta ADN metilado en la secuencia [A/G]C, en la que la C está metilada (véase 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 usarse McrBC para enriquecer la fracción de secuencias con bajo número de copias de un genoma como material de partida para el descubrimiento de polimorfismos.

4) El uso de ADNc en contraposición a ADN genómico con el fin de seleccionar como diana secuencias génicas.

Finalmente, se propone en este caso usar ADNc cebado con oligodT en contraposición a ADN genómico como material de partida para el descubrimiento de polimorfismos, opcionalmente en combinación con el uso de la nucleasa específica de dúplex de cangrejo descrita en el punto 3b para la normalización. Obsérvese que el uso de ADNc cebado con oligodT también excluye secuencias de cloroplastos. Alternativamente, se usan moldes de ADNc-AFLP en lugar de ADNc cebado con oligodT para facilitar la amplificación de las secuencias con bajo número de copias restantes en analogía a AFLP (véase también el punto 3b anterior).

Ejemplo 8: Estrategia para el enriquecimiento de repeticiones de secuencias simples

Este ejemplo describe la estrategia propuesta para el descubrimiento de secuencias de repeticiones de secuencias simples, en analogía al descubrimiento de SNP descrito en el ejemplo 4.

Específicamente, se realiza restricción-ligamiento de ADN genómico de dos o más muestras, por ejemplo usando las endonucleasas de restricción PstI/MseI. Se realiza amplificación AFLP selectiva tal como se describe en el ejemplo 4. A continuación, se enriquecen fragmentos que contienen los motivos SSR seleccionados mediante uno de dos métodos:

1) hibridación por transferencia de tipo Southern sobre filtros que contienen oligonucleótidos que coinciden con los motivos SSR previstos (por ejemplo $(CA)_{15}$ en el caso de enriquecimiento de repeticiones CA/GT), seguido por amplificación de fragmentos unidos de un modo similar al descrito por Armour y colaboradores (Armour, J., Sismani, C., Patsalis, P., y Cross, G. (2000). Measurement of locus copy number by hybridization with amplifiable probes. Nucleic Acids Research vol. 28, n.º 2, págs. 605-609) o mediante

2) enriquecimiento usando sondas de hibridación oligonucleotídicas de captura biotiniladas para capturar fragmentos (AFLP) en disolución tal como se describe por Kijas y colaboradores (Kijas, J.M., Fowler, J.C., Garbett C.A., y Thomas, M.R., (1994). Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. Biotechniques, vol. 16, págs. 656-662.

A continuación, se amplifican los fragmentos de AFLP enriquecidos en motivos SSR usando los mismos cebadores de AFLP que se usan en la etapa de preamplificación, para generar una biblioteca de secuencias. Se clona T/A una alícuota de los fragmentos amplificados y se secuencian 96 clones para estimar la fracción de clones positivos (clones que contienen el motivo SSR previsto, por ejemplo motivos CA/GT mayores de 5 unidades de repetición. Se detecta otra alícuota de la mezcla de fragmentos de AFLP enriquecida mediante electroforesis en gel de poliacrilamida (PAGE), opcionalmente tras amplificación selectiva adicional para obtener una huella legible, con el fin de inspeccionar visualmente si están enriquecidos fragmentos que contienen SSR. Tras completar satisfactoriamente estas etapas de control, se someten las bibliotecas de secuencias a secuenciación de 454 de alto rendimiento.

La estrategia anterior para el descubrimiento de SSR *de novo* SSR se representa esquemáticamente en la figura 8A, y puede adaptarse para otros motivos de secuencia sustituyendo las secuencias oligonucleotídicas de captura por consiguiente.

Ejemplo 9. Estrategia para evitar etiquetas mixtas.

Etiquetas mixtas se refiere a la observación de que además de la combinación de cebadores de AFLP etiquetados esperada por muestra, se observa una baja fracción de secuencias que contienen una etiqueta de la muestra 1 en un extremo, y una etiqueta de la muestra 2 en el otro extremo (véase también la tabla 1 en el ejemplo 4). Esquemáticamente, la configuración de secuencias que contienen etiquetas mixtas se representa en el presente documento a continuación.

Representación esquemática de las combinaciones de etiquetas de muestras esperadas.

	Etiqueta EcoRI	Etiqueta MseI
PSP 11:	5’-CGTC	ACCA-3’
	3’-GCAG	TGGT-5’
PI-201234	5’-CAAG	GGCT-3’
	3’-GTTC	CCGA-5’

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’

La observación de etiquetas mixtas excluye la asignación correcta de la secuencia a o bien PSP11 o bien PI-201234.

Se muestra en la figura 5A un ejemplo de una secuencia de etiquetas mixtas observada en la ronda de secuenciación de pimiento descrita en el ejemplo 4. Se muestra en el panel 2 de la figura 5A una revisión de la configuración de fragmentos observados que contienen etiquetas esperadas y etiquetas mixtas.

La explicación molecular propuesta para las etiquetas mixtas es que durante la etapa de preparación de bibliotecas de secuencias, se hacen romos los extremos de fragmentos de ADN usando ADN polimerasa de T4 o enzima Klenow para eliminar extremos sobresalientes en 3 prima, antes del ligamiento al adaptador (Margulies *et al.*, 2005). Aunque esto puede funcionar bien cuando se procesa una única muestra de ADN, en el caso de usar una mezcla de

dos o más muestras de ADN etiquetadas de manera diferente, el relleno por la polimerasa da como resultado la incorporación de la secuencia de etiqueta errónea en el caso en el que se ha formado un heterodúplex entre las hebras complementarias derivadas de muestras diferentes (figura 5B panel 3, etiquetas mixtas). Se ha encontrado la solución agrupando las muestras tras la etapa de purificación que seguía al ligamiento al adaptador en la epata de construcción de bibliotecas de secuencias 454 tal como se muestra en la figura 5C panel 4.

Ejemplo 10. Estrategia para evitar etiquetas mixtas y concatámeros usando un diseño mejorado para la preparación de bibliotecas de secuencias 454.

Además de la observación bajas frecuencias de lecturas de secuencias que contienen etiquetas mixtas tal como se describe en el ejemplo 9, se ha observado una baja frecuencia de lecturas de secuencias observadas a partir de fragmentos de AFLP concatenados.

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

La explicación molecular propuesta para la aparición de fragmentos de AFLP concatenados es que durante la etapa de preparación de bibliotecas de secuencias 454, se hacen romos los extremos de fragmentos de ADN usando ADN polimerasa de T4 o enzima Klenow para eliminar extremos sobresalientes en 3 prima, antes del ligamiento al adaptador (Margulies *et al*., 2005). Como resultado, fragmentos de ADN de muestras con extremos romos están en competición con los adaptadores durante la etapa de ligamiento y pueden ligarse entre sí antes de ligarse a los adaptadores. Este fenómeno es de hecho independiente de si se incluye una única muestra de ADN o una mezcla de múltiples muestras (etiquetadas) en la etapa de preparación de bibliotecas, y por tanto también puede producirse durante la secuenciación convencional tal como se describe por Margulies y colaboradores. En el caso de usar muestras de etiquetado múltiple tal como se describe en el ejemplo 4, los concatámeros complican la asignación correcta de lecturas de secuencias a muestras basándose en la información de etiquetas y por tanto deben evitarse.

La solución propuesta a la formación de concatámeros (y etiquetas mixtas) es reemplazar el ligamiento de adaptadores con extremos romos por ligamiento de adaptadores que contienen una proyección T en 3 prima, en analogía 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 que contienen una proyección T en 3 prima contengan una proyección C en el extremo 3' opuesto (que no se ligará al fragmento de ADN de muestra, para prevenir la formación de concatámeros con extremos romos de secuencias adaptadoras (véase la figura 6B panel 3). El flujo de trabajo adaptado resultante en el procedimiento de construcción de bibliotecas de secuencias cuando se usa el enfoque de adaptador modificado se muestra esquemáticamente en la figura 6C panel 4.

LISTA DE SECUENCIAS

<110> Keygene NV

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

<130> P27819EP01

<160> 46

<170> Patente en versión 3.3

<210> 1
<211> 22
<212> ADN
<213> Artificial

<220>
<223> Cebador

<400> 1

cgtcagactg cgtaccaatt ca 22

<210> 2
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 2

tggtgatgag tcctgagtaa ca 22

<210> 3
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 3

caagagactg cgtaccaatt ca 22

<210> 4
<211> 22
<212> ADN
<213> Artificial

<220>

<223> cebador

<400> 4

agccgatgag tcctgagtaa ca
22

<210> 5
<211> 17
<212> ADN
<213> Artificial

<220>
<223> adaptador

<400> 5

ctcgtagact gcgtacc
17

<210> 6
<211> 18
<212> ADN
<213> Artificial

<220>
<223> adaptador

<400> 6

aattggtacg cagtctac
18

<210> 7
<211> 16
<212> ADN
<213> Artificial

<220>
<223> adaptador

<400> 7

gacgatgagt cctgag
16

<210> 8
<211> 14
<212> ADN
<213> Artificial

<220>
<223> adaptador

<400> 8

tactcaggac tcat
14

<210> 9
<211> 18
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 9

agactgcgta ccaattca
18

<210> 10
<211> 17
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 10

gatgagtcct gagtaac
17

<210> 11
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 11

cgtcagactg cgtaccaatt ca
22

<210> 12
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 12

caagagactg cgtaccaatt ca
22

<210> 13
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 13

tggtgatgag tcctgagtaa ca
22

<210> 14
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 14

agccgatgag tcctgagtaa ca
22

<210> 15
<211> 19
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 15

gactgcgtac caattcaac
19

<210> 16
<211> 19
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 16

gatgagtcct gagtaacag
19

<210> 17
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 17

cgtcagactg cgtaccaatt ca
22

<210> 18
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 18

tggtgatgag tcctgagtaa ca
22

<210> 19
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 19

caagagactg cgtaccaatt ca
22

<210> 20
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 20

caagagactg cgtaccaatt ca
22

<210> 21
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 21

acgtgtagac tgcgtaccga aa
22

<210> 22
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 22

acgtgatgag tcctgagtaa ca
22

<210> 23
<211> 22

<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 23

cgtagtagac tgcgtaccga ac
22

<210> 24
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 24

cgtagatgag tcctgagtaa ca
22

<210> 25
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 25

gtacgtagac tgcgtaccga ag
22

<210> 26
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 26

gtacgatgag tcctgagtaa ca
22

<210> 27
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 27

tacggtagac tgcgtaccga at
22

<210> 28
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 28

tacggatgag tcctgagtaa ca
22

<210> 29
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 29

agtcgtagac tgcgtaccga aa
22

<210> 30
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 30

agtcgatgag tcctgagtaa ca
22

<210> 31
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 31

catggtagac tgcgtaccga ac
22

<210> 32
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 32

catggatgag tcctgagtaa ca
22

<210> 33
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 33

gagcgtagac tgcgtaccga ag
22

<210> 34
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 34

gagcgatgag tcctgagtaa ca
22

<210> 35
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 35

tgatgtagac tgcgtaccga at
22

<210> 36
<211> 22
<212> ADN
<213> Artificial

<220>
<223> cebador

<400> 36

tgatgatgag tcctgagtaa ca
22

<210> 37
<211> 20
<212> ADN
<213> artificial

<220>
<223> cebador

<400> 37

aaacccaaac tcccccaatc
20

<210> 38
<211> 68
<212> ADN
<213> artificial

<220>
<223> cebador

<400> 38

agcggataac aatttcacac aggacatcag tagtcacact ggtacaaaaa tagagcaaaa
60

cagtagtg
68

<210> 39
<211> 91
<212> ADN
<213> artificial

<220>
<223> sonda

<400> 39

aaacccaaac tcccccaatc gatttcaaac ctagaacaat gttggttttg gtgctaactt
60

caacccccact actgttttgc tctatttttg t
91

<210> 40
<211> 90
<212> ADN
<213> artificial

<220>
<223> secuencia que contiene PI-201234 SNP

<400> 40

aaacccaaac tcccccaatc gatttcaaac ctagaacagt gttggttttg gtgctaactt
60

caacccccact actgttttgc tctatttttg
90

<210> 41
<211> 90
<212> ADN
<213> artificial

<220>
<223> PI-201234 SNP

<400> 41

aaacccaaac tcccccaatc gatttcaaac ctagaacagt gttggttttg gtgctaactt
60

caaccccact actgttttgc tctatttttg
90

<210> 42
<211> 42
<212> ADN
<213> artificial

<220>
<223> sonda SNPWave

<400> 42

gatgagtcct gagtaaccca atcgatttca aacctagaac aa
42

<210> 43
<211> 44
<212> ADN
<213> artificial

<220>
<223> sonda SNPWave

<400> 43

gatgagtcct gagtaaccac caatcgattt caaacctaga acag
44

<210> 44
<211> 50
<212> ADN
<213> artificial

<220>
<223> sonda snpwave

<400> 44

tgttggtttt ggtgctaact tcaaccaaca tctggaattg gtacgcagtc
50

<210> 45
<211> 16
<212> ADN
<213> artificial

<220>
<223> cebador

<400> 45

gactgcgtac caattc
16

<210> 46
<211> 16
<212> ADN
<213> artificial

gatgagtcct gagtaa
16

REIVINDICACIONES

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

a) proporcionar una pluralidad de muestras de ácido nucleico de interés;

b) realizar una reducción de la complejidad en cada una de las muestras para proporcionar una pluralidad de bibliotecas de muestras de ácido nucleico,

c) ligar adaptadores a las muestras de ácido nucleico de complejidad reducida en las bibliotecas, usando un adaptador que porta una proyección 3'-T;

d) secuenciar al menos una parte de las bibliotecas;

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

f) determinar uno o más polimorfismos entre la pluralidad de muestras de ácido nucleico en la alineación de la etapa e);

g) opcionalmente, examinar un ácido nucleico de muestra de prueba de interés para identificar la presencia, ausencia o cantidad del uno o más polimorfismos determinados en la etapa f) usando sondas de detección.

2. Método según la reivindicación 1, en el que el ácido nucleico de muestra de prueba es una muestra de ácido nucleico de complejidad reducida que se ha obtenido usando la reducción de la complejidad usada en la etapa b).

3. Método según la reivindicación 1, en el que la etapa b) comprende además la etapa de etiquetar la biblioteca para obtener una biblioteca etiquetada.

4. Método según la reivindicación 3, en el que la etiqueta se proporciona en el adaptador y/o el cebador.

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

6. Método según la reivindicación 4, en el que al menos uno de los cebadores está fosforilado.

7. Método según la reivindicación 6, en el que la secuenciación comprende secuenciar sobre un soporte sólido.

8. Método según cualquiera de las reivindicaciones anteriores, en el que el examen se realiza mediante inmovilización de las sondas diseñadas en la etapa g) según la reivindicación 1 sobre una matriz, seguido por la puesta en contacto de la matriz que comprende sondas con una biblioteca de prueba en condiciones de hibridación.

9. Uso del método según las reivindicaciones 1-8 para examinar bibliotecas de microsatélites enriquecidas, realizar la obtención del perfil de transcrito de ADNc-AFLP (hibridación de tipo Northern digital), secuenciación de genomas complejos, secuenciación de bibliotecas de etiquetas de secuencia expresada (en ADNc completo o ADNc-AFLP), descubrimiento de microARN (secuenciación de bibliotecas de insertos pequeñas), secuenciación de cromosomas artificiales bacterianos (cóntigos), enfoque de análisis de segregantes agrupados en combinación con AFLP/ADNc-AFLP, detección de rutina de fragmentos AFLP (retrocruzamientos asistidos por marcador).

Figura 1

Conjunto I de cebadores usado para la preamplificación de PSP-11

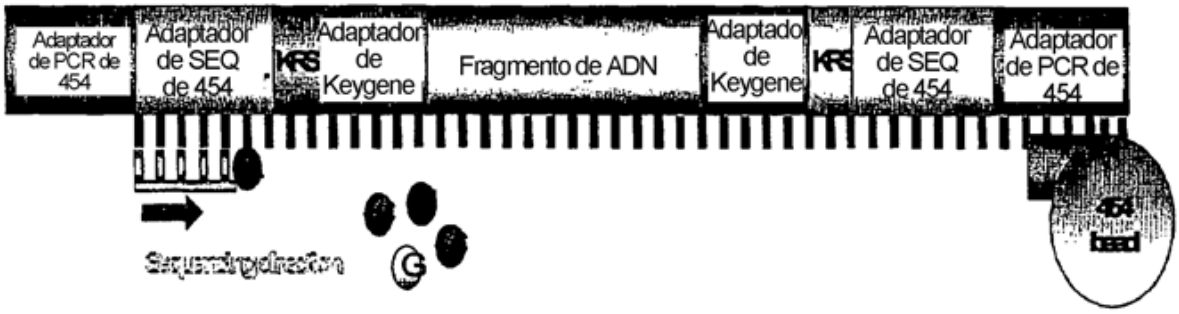
E01LKRS1 5'-CGTCAGACTGCGTACCAATTCA-3'

M15KKRS1 5'-TGGTGATGAGTCCTGAGTAACA-3'

Conjunto II de cebadores usado para la preamplificación de PI20234

E01LKRS2 5'-CAAGAGACTGCGTACCAATTCA-3'

M15KKRS2 5'-AGCCGATGAGTCCTGAGTAACA-3'



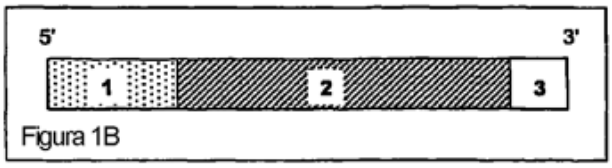


Figura 1B

kb

Figura 2
Control de calidad de ADN sobre un gel de agarosa al 1%

Figura 2A: Electroforesis en gel corto

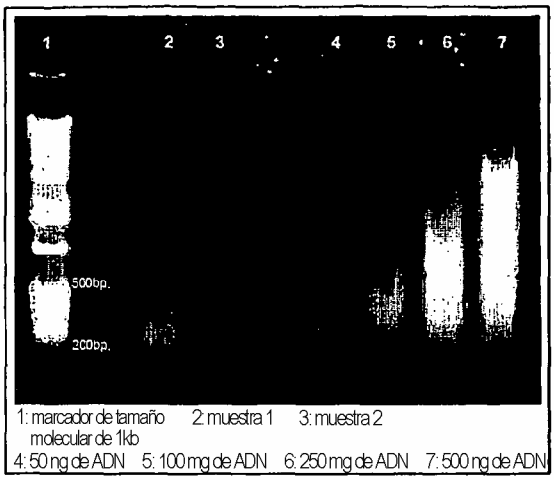


Figura 2B: Electroforesis en gel largo

Concentración de ADN medida con Nanodrop

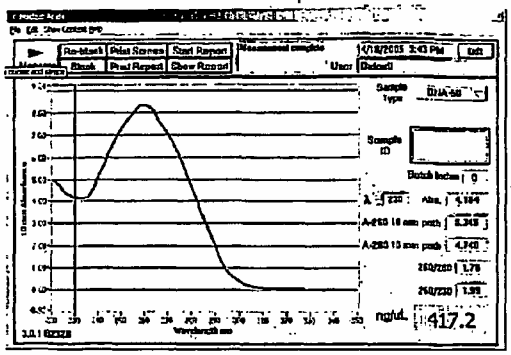


Figura 2C. Concentración de la muestra 1.

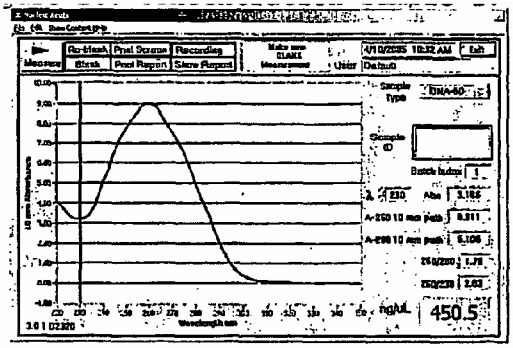


Figura 2D. Concentración de la muestra 2

Figura 3
Control de calidad de ADN sobre un gel de agarosa al 1%

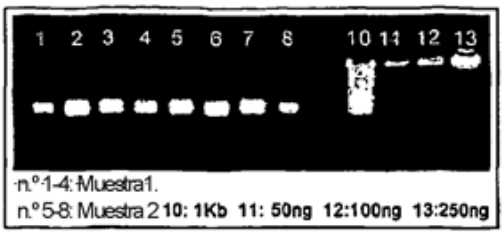


Figura 3A: Electroforesis en gel corto

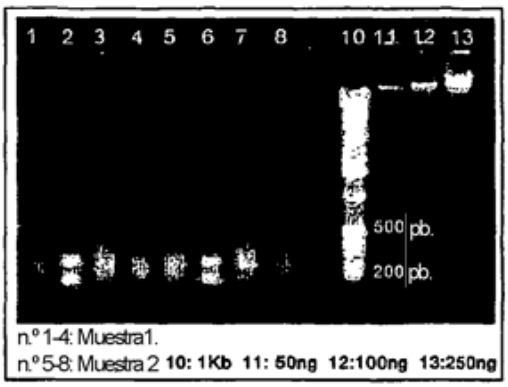


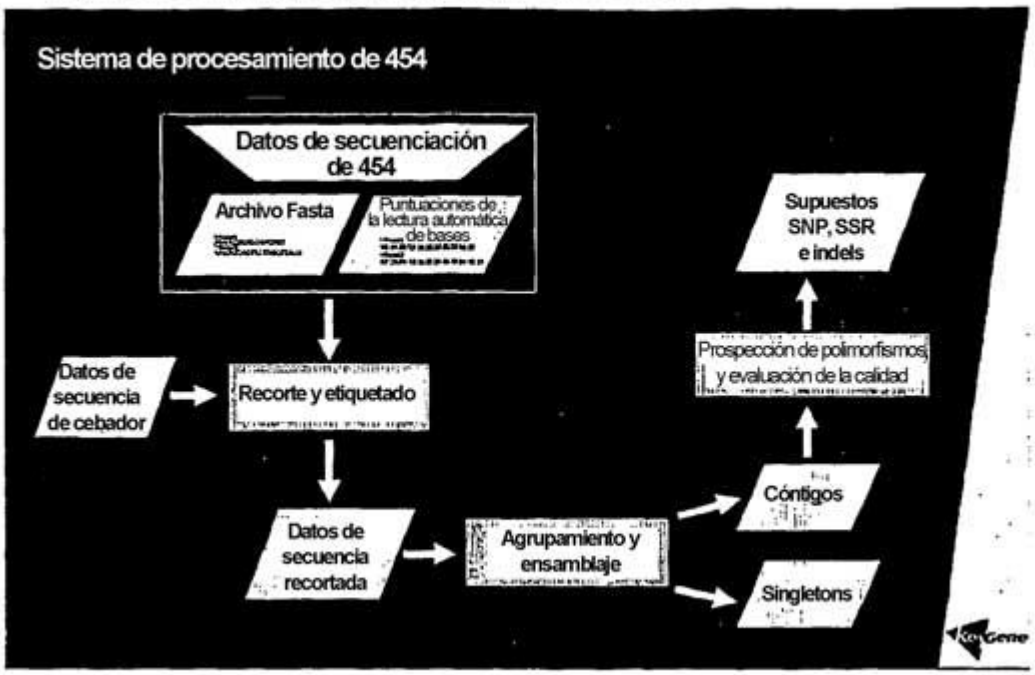
Figura 3A: Electroforesis en gel largo

Concentraciones de ADN medidas sobre Nanodrop

n.º	ID de la 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

n.º	ID de la 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. Sistema de procesamiento de datos de secuencia



Figiura 4B. Prospección de polimorfismos y el procedimiento de evaluación de la calidad

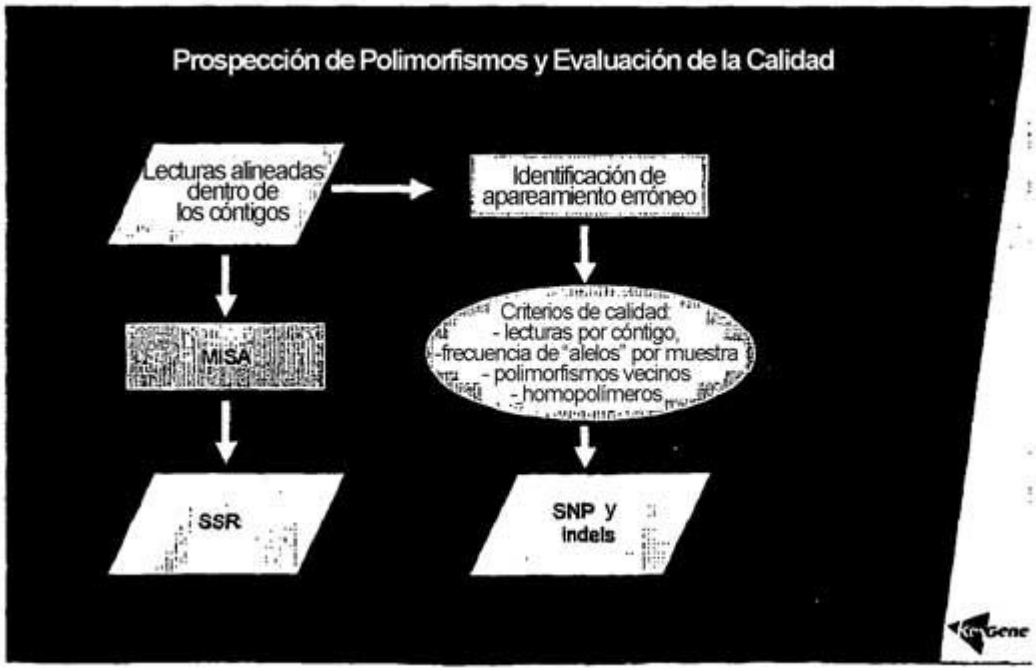


FIG 5A

Panel 1: Ejemplo de una etiqueta mixta

CAAGAGACTGCGTACCAATTCAACTTTGAGGTGAAAGATCGAAGGTTGCA
CAAGAGACTGCGTACCAATTCA (ES2)

AACACCAAGTGGCCGACCATCTCTTGCGTGTTACTCAGGACTCATCACCAC
(MS1) TGTTACTCAGGACTCATCACCA

Panel 2: Vista general de fragmentos obervados que contienen etiquetas esperadas y etiquetas mixtas

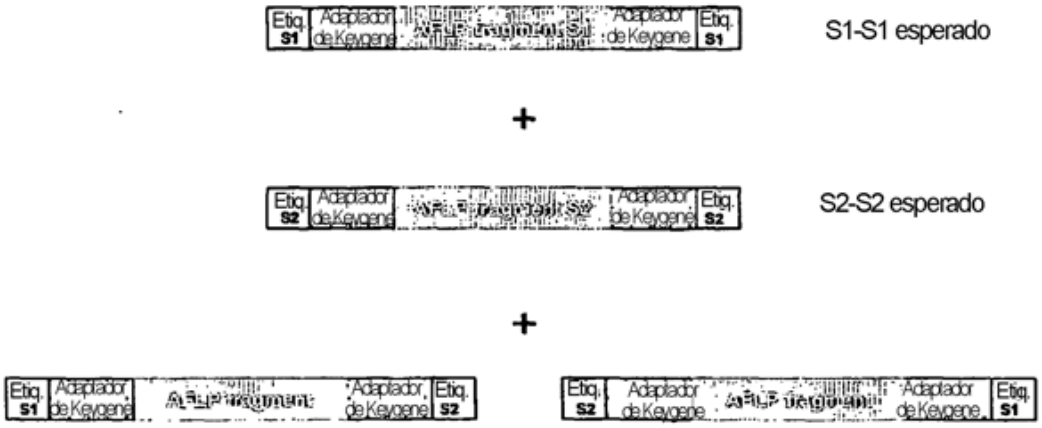


FIG 5B

Panel 3: Causa planteada como hipótesis de la generación de etiquetas mixtas

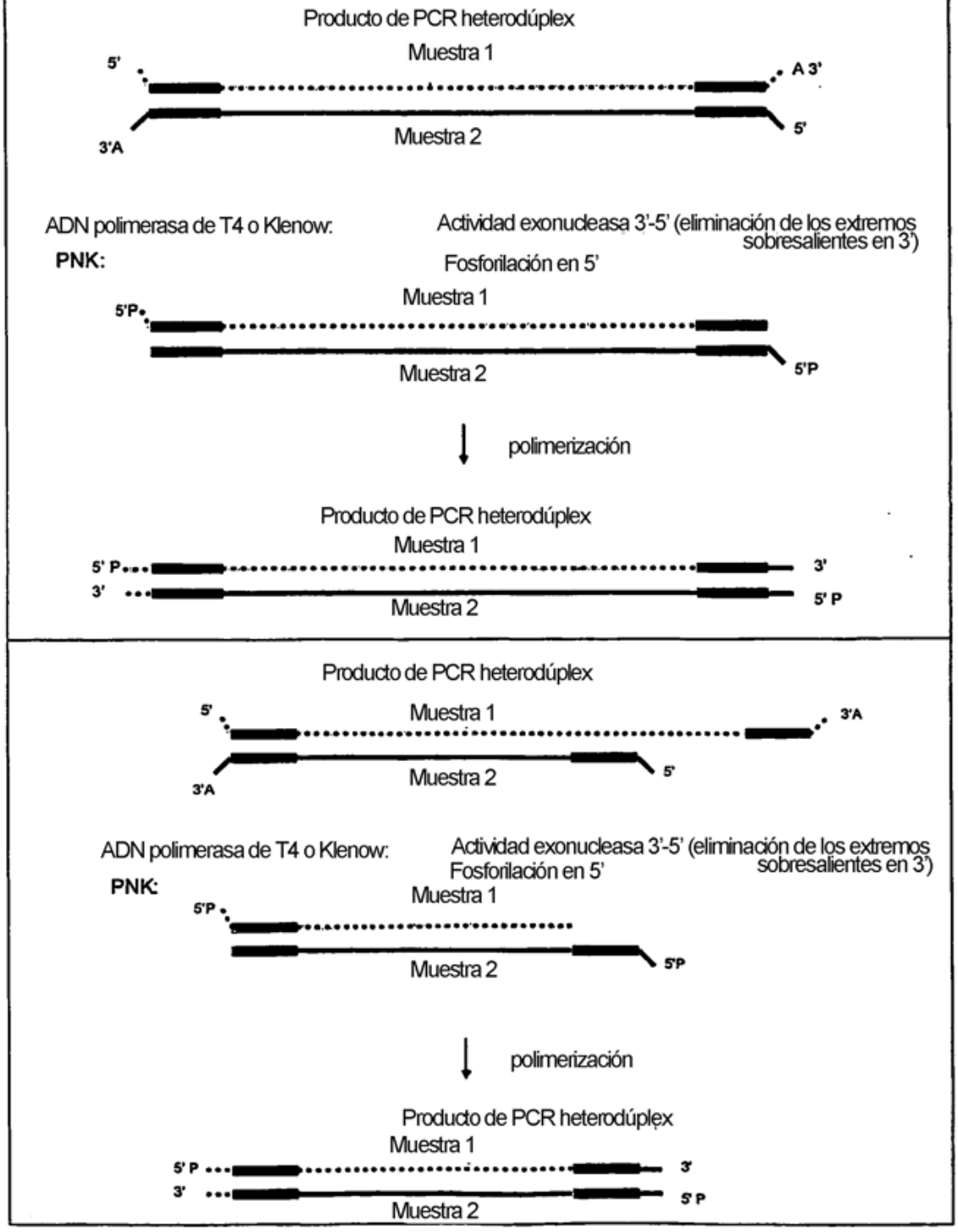


FIG 5C

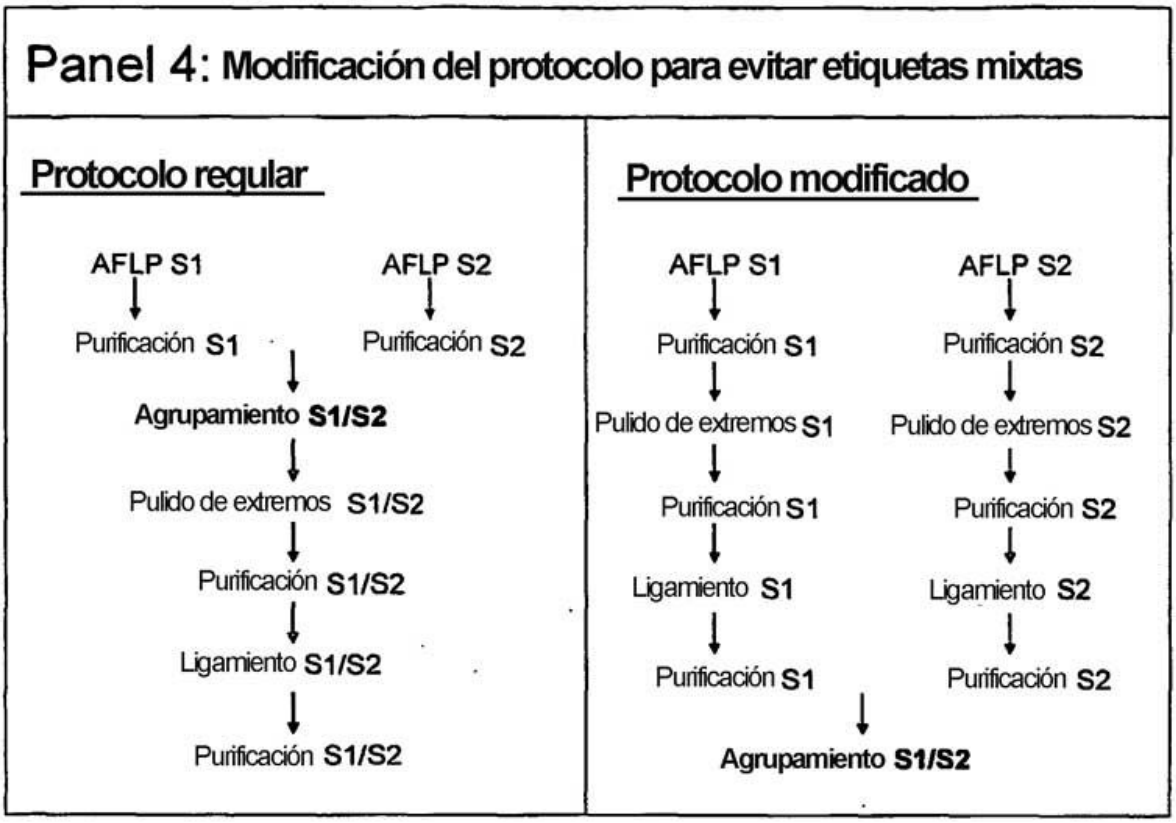

Panel 4: Modificación del protocolo para evitar etiquetas mixtas
Protocolo regular
AFLP S1
AFLP S2
Purificación S1
Purificación S2
Agrupamiento S1/S2
Pulido de extremos S1/S2
Purificación S1/S2
Ligamiento S1/S2
Purificación S1/S2
Protocolo modificado
AFLP S1
AFLP S2
Purificación S1
Purificación S2
Pulido de extremos S1
Pulido de extremos S2
Purificación S1
Purificación S2
Ligamiento S1
Ligamiento S2
Purificación S1
Purificación S2
Agrupamiento 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 global de fragmentos observados que contienen etiquetas esperadas y concatámeros

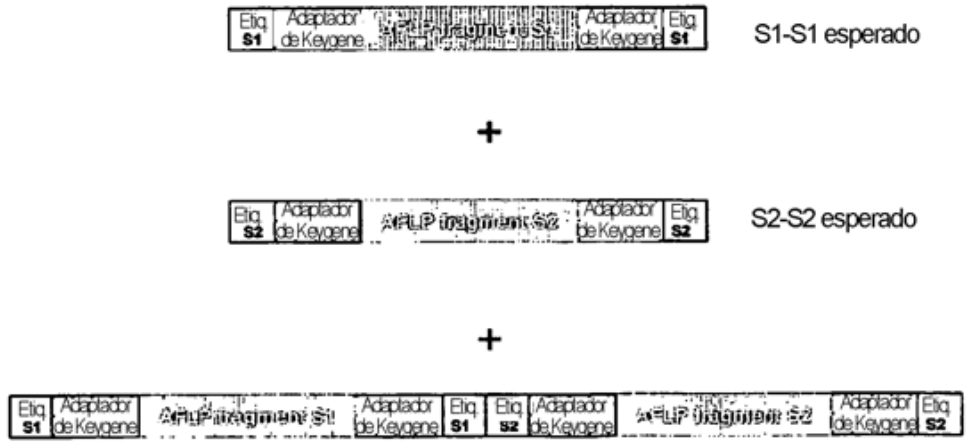


FIG 6B

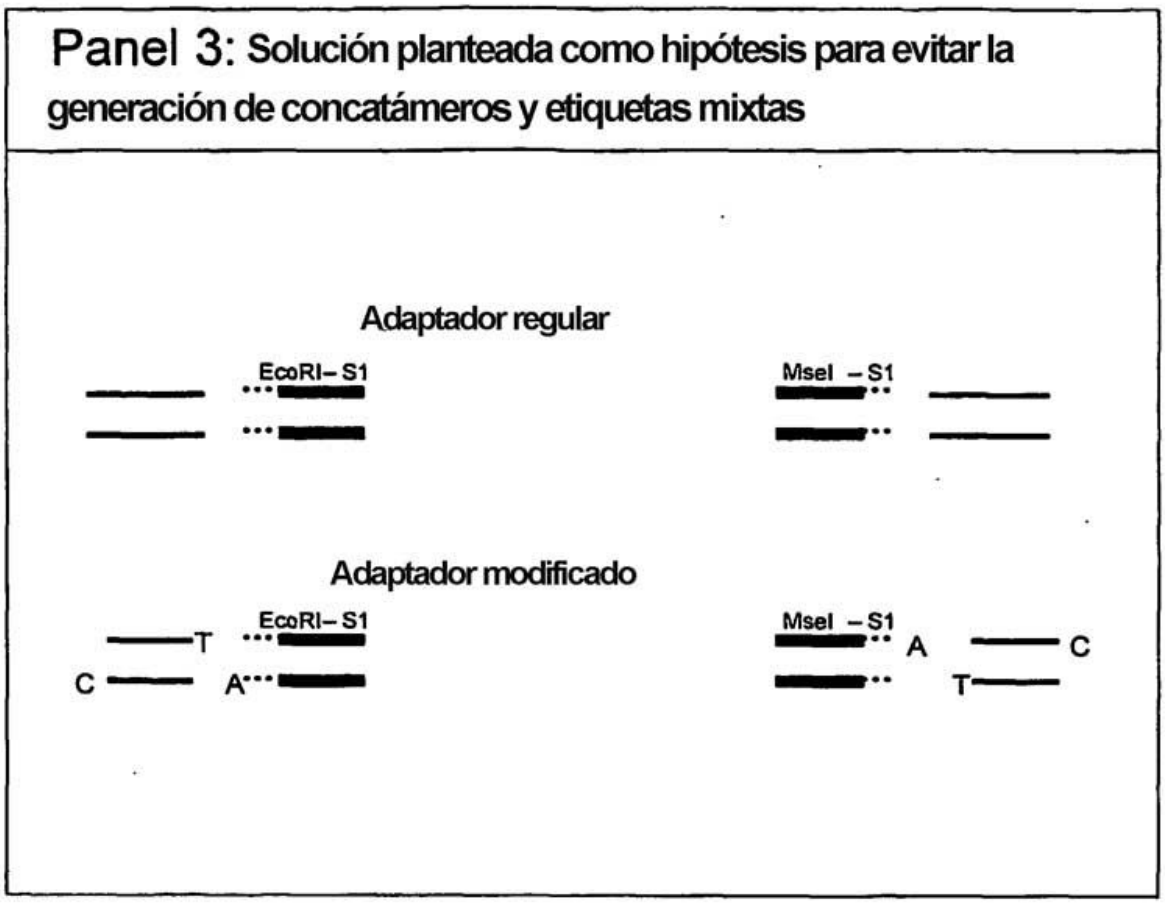

Panel 3: Solución planteada como hipótesis para evitar la generación de concatámeros y etiquetas mixtas
Adaptador regular
EcoRI– S1
MseI – S1
Adaptador modificado
EcoRI– S1
MseI – S1
T
C
A
A
C
T

FIG 6C

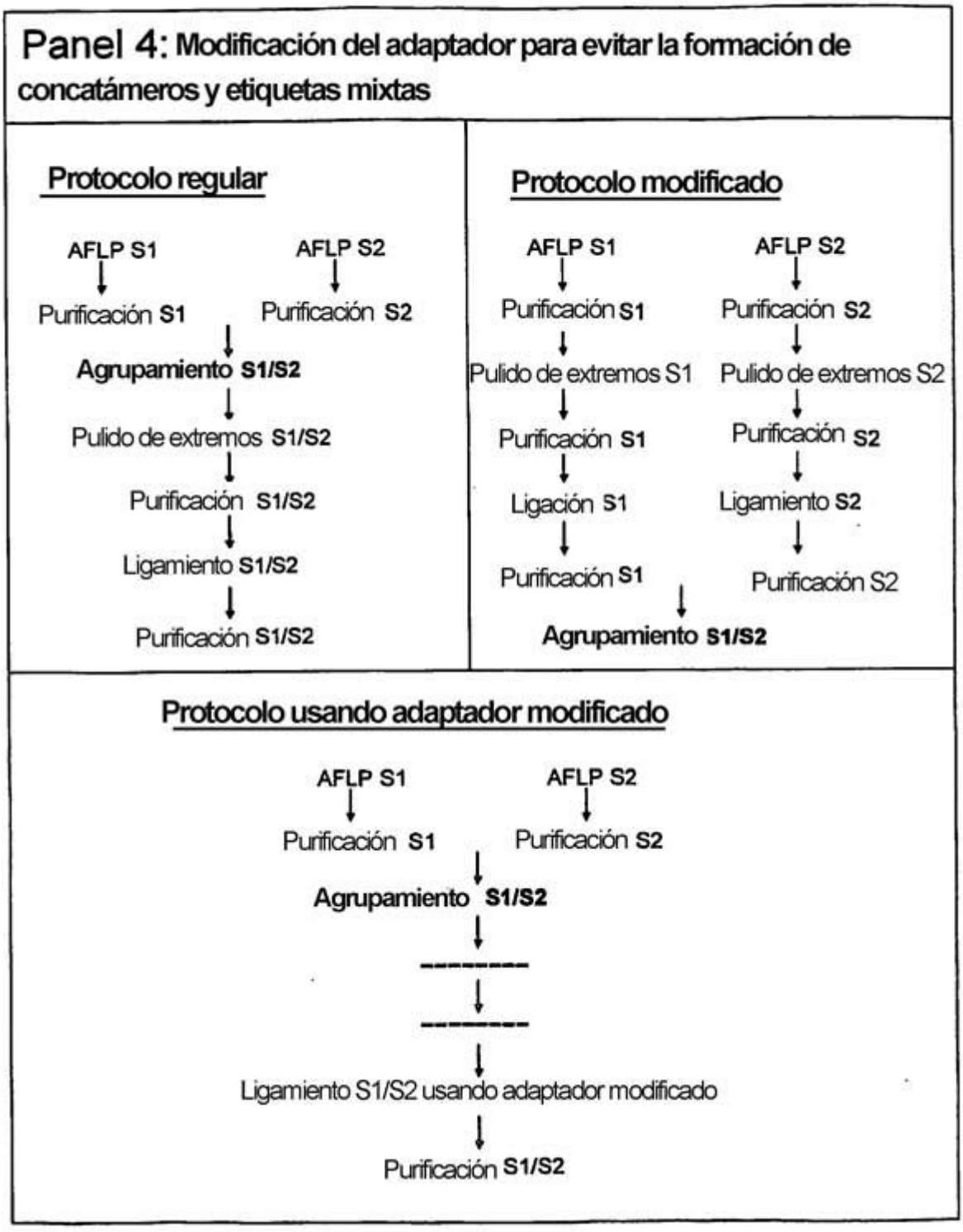

Panel 4: Modificación del adaptador para evitar la formación de concatámeros y etiquetas mixtas
Protocolo regular
AFLP S1
AFLP S2
Purificación S1
Purificación S2
Agrupamiento S1/S2
Pulido de extremos S1/S2
Purificación S1/S2
Ligamiento S1/S2
Purificación S1/S2
Protocolo modificado
AFLP S1
AFLP S2
Purificación S1
Purificación S2
Pulido de extremos S1
Pulido de extremos S2
Purificación S1
Purificación S2
Ligación S1
Ligamiento S2
Purificación S1
Purificación S2
Agrupamiento S1/S2
Protocolo usando adaptador modificado
AFLP S1
AFLP S2
Purificación S1
Purificación S2
Agrupamiento S1/S2
Ligamiento S1/S2 usando adaptador modificado
Purificación S1/S2

FIG 7

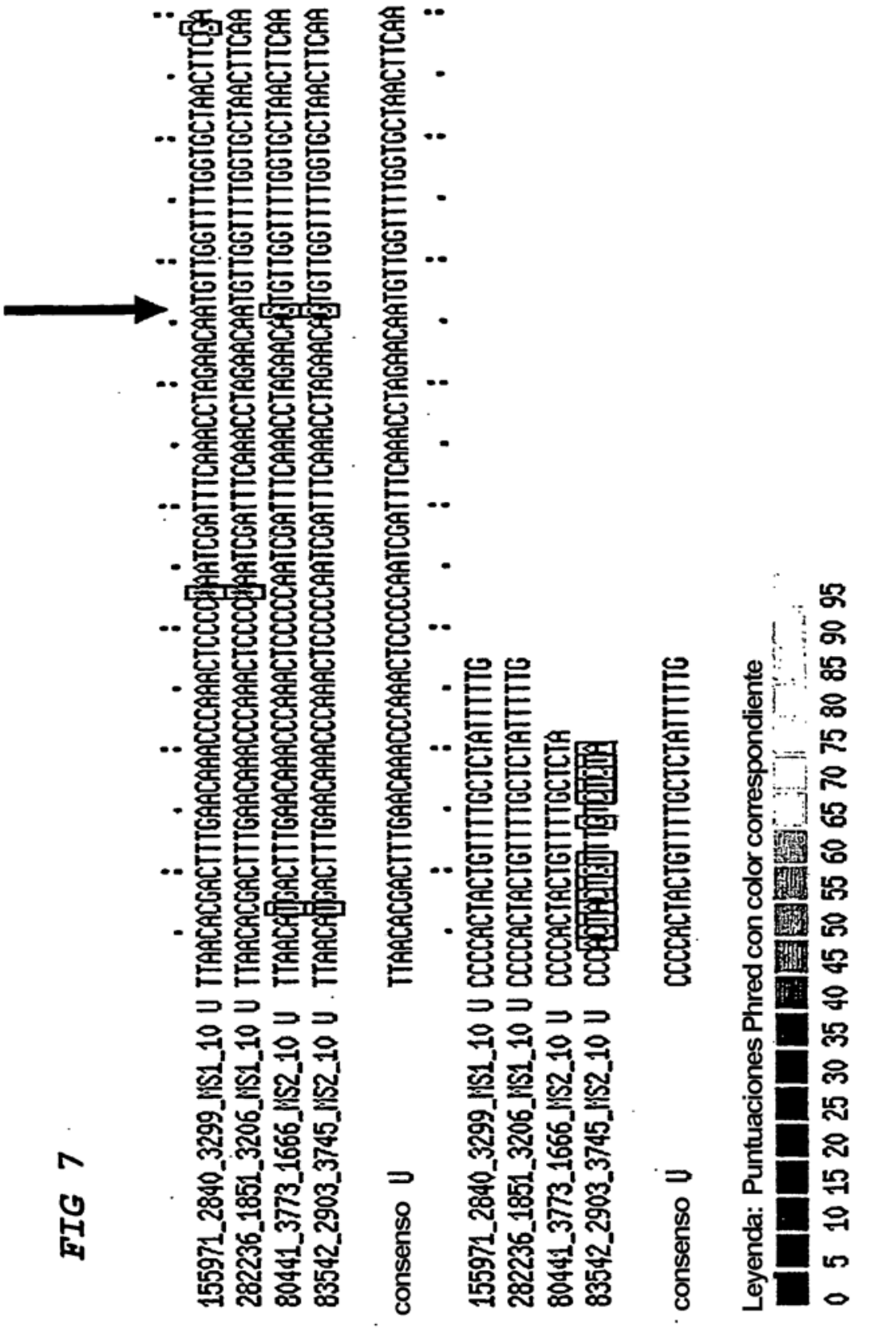

155971_2840_3299_NS1_10 U
282236_1851_3206_NS1_10 U
80441_3773_1666_NS2_10 U
83542_2903_3745_NS2_10 U
consenso U
155971_2840_3299_NS1_10 U
282236_1851_3206_NS1_10 U
80441_3773_1666_NS2_10 U
83542_2903_3745_NS2_10 U
consenso U
Leyenda: Puntuaciones Phred con color correspondiente
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95

FIG 8A.

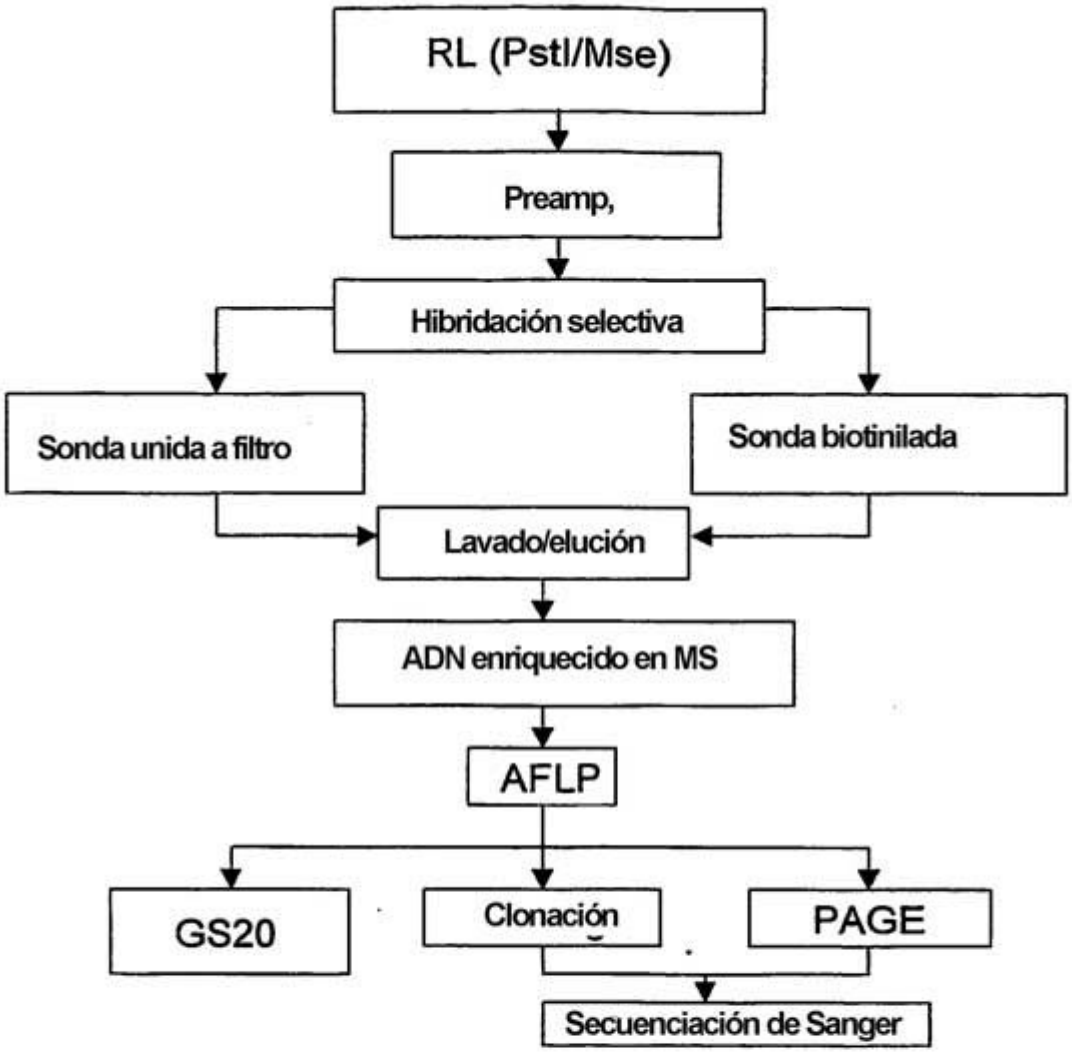
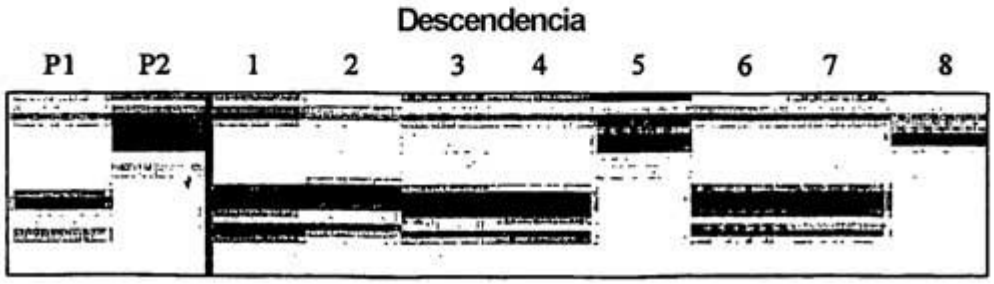

RL (PstI/Mse)
Preamp,
Hibridación selectiva
Sonda unida a filtro
Sonda biotinilada
Lavado/elución
ADN enriquecido en MS
AFLP
GS20
Clonación
PAGE
Secuenciación de Sanger

FIG 8B


Descendencia
P1
P2
1
2
3
4
5
6
7
8