

19



OFICINA ESPAÑOLA DE  
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 394 633**

51 Int. Cl.:

**C12Q 1/68**

(2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

96 Fecha de presentación y número de la solicitud europea: **21.12.2006 E 06835676 (5)**

97 Fecha y número de publicación de la solicitud europea: **10.09.2008 EP 1966394**

54 Título: **Estrategias mejoradas para elaboración de perfiles de transcritos usando tecnologías de secuenciación de alto rendimiento**

30 Prioridad:

**22.12.2005 US 752591 P**

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

**04.02.2013**

73 Titular/es:

**KEYGENE N.V. (100.0%)  
AGRO BUSINESS PARK 90  
6708 PW WAGENINGEN, NL**

72 Inventor/es:

**VAN EIJK, MICHAEL JOSEPHUS THERESIA**

74 Agente/Representante:

**CARPINTERO LÓPEZ, Mario**

ES 2 394 633 T3

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

## DESCRIPCIÓN

Estrategias mejoradas para elaboración de perfiles de transcritos usando tecnologías de secuenciación de alto rendimiento

## Campo técnico

- 5 La presente invención se refiere a los campos de la biología molecular y la genética. La invención se refiere a estrategias mejoradas para determinar la secuencia de transcritos basadas en el uso de tecnologías de secuenciación de alto rendimiento. La invención se refiere además a estrategias mejoradas para la elaboración no sesgada de perfiles de transcritos.

## Antecedentes de la invención

- 10 La elaboración de perfiles de transcritos es una de las tecnologías fundamentales usadas en la investigación actual en biotecnología. El principal dominio de aplicación de la elaboración de perfiles de transcritos es el descubrimiento de genes implicados en rasgos complejos. Esto incluye una amplia variedad de fenómenos biológicos tales como el descubrimiento de genes implicados en enfermedades (humanas) con el fin de identificar objetivos para el desarrollo de medicación (descubrimiento de objetivos), descifrar rutas bioquímicas que controlan la síntesis de biomoléculas (industria de la fermentación), analizar rasgos complejos para la cría de plantas y animales (descubrimiento de genes) y muchos otros.

- Un segundo dominio de aplicación sigue la ruta inversa, es decir, usar la elaboración de perfiles de transcritos para la determinación diagnóstica rutinaria de perfiles de transcritos de (un subconjunto seleccionado de) genes con el fin de predecir un fenotipo complejo. Ejemplos de esta categoría son la clasificación molecular, el diagnóstico y la predicción de pronósticos clínicos de cáncer de mama humano (Van de Vijver et al., 2002; N. Engl. J. Med., vol. 347:25:1999-2009; van't Veer et al., 2002, Breast Cancer Res., vol. 5 (1): 57-8; www.agendia.com) y de carcinoma papilar de células renales (Yang et al., 2005). Se describen enfoques para la identificación de genes pertinentes basada en datos de elaboración de perfiles de transcritos recogidos en poblaciones secretoras por Schadt y sus colaboradores (2005, Sci. STKE, vol. 296:pre40). En resumen, la elaboración de perfiles de transcritos es de vital importancia en la investigación en ciencias biológicas.

- Las tecnologías para la elaboración de perfiles de transcritos han evolucionado rápidamente durante los últimos 10 años. Hasta principios de los noventa (poco después de la disponibilidad generalizada de la PCR), la elaboración de perfiles de transcritos se realizaba mediante análisis de bandas northern o ensayos de protección de RNasa. Aunque estas técnicas son bastante específicas y sensibles (especialmente los ensayos de protección de RNasa), las limitaciones de estas tecnologías son que sólo se pueden analizar uno o unos pocos genes a la vez (rendimiento bajo), al mismo tiempo que los procedimientos son tediosos y requieren mucho tiempo. Además, ambos procedimientos requieren el uso de técnicas de marcaje radiactivo, lo que supone un peligro para la salud.

- Con la aparición de la técnica de presentación diferencial (PD) en 1992 (Liang y Pardee, 1992, Science, vol. 257 (5072): 967-71) y muchas modificaciones y mejoras de la PD (p. ej., Ordered Differential Display, Matz et al., 1997, Nucl. Acids, Res., vol. 25(12):2541-2), se dio un paso hacia la elaboración de perfiles de transcritos multiplexados. Las características de la PD son que subconjuntos de genes aleatorios son el objetivo mediante la hibridación poco rigurosa de un cebador de PCR diseñado aleatoriamente con la muestra de ADNc que se va a analizar, dando como resultado la amplificación preferente de transcritos expresados que contienen secuencias con alta homología con el cebador de PCR usado. Posteriormente, los productos de amplificación se resuelven en geles de secuencia, dando como resultado un patrón identificador que representa los subconjuntos de genes transcritos. Aunque los procedimientos de PD tienen un rendimiento superior en comparación con las bandas northern y los ensayos de protección de RNasa, sus limitaciones son la reproducibilidad/solidez bastante baja de estas técnicas. Esto se debe, en parte, a la hibridación inespecífica del cebador de PCR aleatorio usado. En consecuencia, los patrones identificadores generados usando diferentes cebadores aleatorios no se dirigen sistemáticamente a subconjuntos diferentes (complementarios) de transcritos. Una desventaja adicional es que los procedimientos de PD requieren la preparación de geles planos o la detección por electroforesis en gel por capilaridad. Otra limitación más es que no se conoce el origen génico de las bandas observadas en las identificaciones, lo que requiere la escisión de la banda, la elución, la reamplificación y la secuenciación del ADN que se quiere revelar; esta última limitación la comparten otros procedimientos de elaboración de perfiles de transcritos a base de identificaciones. Finalmente, con la detección de 50-100 fragmentos por carril en un gel/rastro de capilaridad, la tecnología es moderadamente multiplexada.

- El procedimiento de ADNc-AFLP (Bachem et al., 1996, Plant J., vol. 9 (5): 745-53) aborda dos de las principales limitaciones de la tecnología de PD, a saber, la reproducibilidad/solidez y la complementariedad de la información obtenida en los identificadores generados con diferentes cebadores de PCR. La solidez y la reproducibilidad del procedimiento de ADNc-AFLP es muy alta debido a que la amplificación de fragmentos de restricción ligados a adaptador usando selectivos AFLP<sup>®</sup> (Keygene N.V., Países Bajos; véanse, p. ej., el documento EP 0 534 858 y Vos P., et al. (1995). AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research, vol. 23, No. 21, p. 4407-4414) cebadores se produce bajo condiciones altamente rigurosas, dando como resultado patrones identificadores altamente reproducibles. Además, el uso de cebadores de AFLP selectivos con diferentes nucleótidos selectivos

garantiza la obtención de identificadores que contienen información complementaria. Por consiguiente, la tecnología de ADNc-AFLP permite la toma de muestras reproducibles de subconjuntos del transcriptoma. Otra ventaja del (ADNc-)AFLP (y la PD) es que no se necesita información de secuencia previa y, por lo tanto, la tecnología se puede aplicar a una amplia variedad de organismos. Las limitaciones del ADNc-AFLP son sus niveles moderados de multiplexación por carril/rastro y el hecho de que no se conoce de forma directa el origen génico de las bandas (véase también la PD).

Las limitaciones en los niveles de multiplexación de los procedimientos de elaboración de perfiles de transcritos anteriormente descritos se han abordado tanto por SAGE (análisis en serie de la expresión génica; Velculescu et al., 1995, *Science*, vol. 270 (5235): 484-7) como por secuenciación masiva de firmas en paralelo (MPSS: Brenner et al., 2000, *Nature Biotechnology*, vol. 18(6):630-4; Meyers et al., 2004, *Nature Biotechnology*, vol. 22(8):1006-11). Al igual que el ADNc-AFLP, ambos procedimientos usan enzimas de restricción de tipo IIS para cortar el ADNc de muestra, seguido por la ligación del adaptador.

En el SAGE, los fragmentos ligados a adaptador se concatenan posteriormente y se secuencian mediante secuenciación de Sanger. A partir del rastro de secuencia de Sanger se extraen marcas de secuencia corta de 14-20 pb, proporcionando información cuantitativa sobre los genes transcritos ("northern digital"). Comparando la frecuencia de las marcas entre muestras, se obtiene información sobre los niveles de expresión relativos entre muestras estudiadas, sin la necesidad de información de secuencia previa. Aunque esto da como resultado la determinación (precisa) de la abundancia relativa del transcrito en diferentes muestras, dadas las marcas de secuencia corta obtenidas es difícil evaluar de qué genes derivan las marcas, a menos que estén disponibles las colecciones de EST grandes o la secuencia completa del genoma del organismo estudiado y puedan someterse las secuencias de marcas a búsquedas de homología tales como el análisis BLAST (Basic Local Alignment Search Tool). Por consiguiente, aunque el SAGE es altamente multiplexado, reproducible y sólido, su valor se limita a organismos con genomas secuenciados. Otra limitación es que el procedimiento no es muy adecuado para procesar muestras grandes (rendimiento bajo) debido a los costes de la secuenciación de Sanger a gran escala.

Al contrario que el SAGE, la MPSS se basa en reacciones de secuenciación en fase sólida. Sin embargo, la MPSS presenta esencialmente las mismas limitaciones que el SAGE, es decir, que se obtienen marcas de secuencia muy cortas (aproximadamente 20 pb), lo que limita en gran medida el seguimiento posterior (identificación génica / conversión del ensayo) de marcas de secuencia interesantes en organismos para los que está disponible una secuencia (del genoma) limitada. En resumen, aunque el SAGE y la MPSS son tecnologías de elaboración de perfiles de transcritos sólidas y altamente multiplexadas que no requieren información de secuencia previa para su aplicación, su valor en la práctica se limita a organismos para los que se ha determinado la secuencia completa del genoma o están disponibles colecciones de EST grandes con el fin de conectar marcas de secuencia a genes. Ambos procedimientos son de bajo rendimiento y técnicamente complejos.

Los puntos fuertes conceptuales son que ambos procedimientos se basan en el muestreo estadístico de colecciones de transcritos (dando lugar a "northern digitales") en combinación con determinación de secuencias precisa, lo que permite estimaciones no sesgadas de niveles de transcripción (relativos) de muchos genes simultáneamente y el hecho de que la elaboración de perfiles de transcritos no presenta hibridación cruzada con sondas sobre soportes sólidos.

En 1995, se introdujeron las micromatrices de expresión génica (Schena et al., 1995, *Science*, vol. 270 (5235): 467-70), lo que supuso un cambio paradigmático en el campo de la elaboración de perfiles de transcritos. Aunque inicialmente se usaban como sondas las llamadas micromatrices "moteadas" que contenían productos de PCR derivados de EST, en los años posteriores se ha centrado la atención en los chips de ADN de oligonucleótidos (Pease et al., 1994, *Proc. Nat. Ac. Sci. USA*, vol. 91(11):5022-6), debido a su mayor solidez y flexibilidad de escala. Actualmente, el mercado de la elaboración de perfiles de transcritos está dominado por chips de ADN de oligonucleótidos de diversos proveedores (p. ej., Affymetrix, Nimblegen, Agilent, etc.). La potencia de los chips de ADN reside en el gran número de secuencias de ADN que se pueden unir / sintetizar sobre su superficie, lo que permite la elaboración de perfiles de transcritos paralela masiva, permitiendo, p. ej., la elaboración de perfiles de transcritos para todos los genes humanos conocidos (= nivel de multiplexación de genes alto). Además, el procedimiento de fabricación de los chips y la hibridación se pueden automatizar y controlar, lo que permite un alto rendimiento y solidez, respectivamente. En consecuencia, los chips de ADN son el estado de la técnica para la elaboración de perfiles de transcritos en el año 2005. Sin embargo, aunque la capacidad de multiplexación, el rendimiento y la solidez son puntos fuertes muy importantes de los chips de ADN, dos importantes limitaciones de la elaboración de perfiles de transcritos basada en chips son que se necesita la información de secuencia para poder construir el chip y que la hibridación cruzada entre secuencias altamente homólogas tales como las derivadas de miembros de familias de genes duplicados puede afectar a la precisión de los resultados. Esta última es muy difícil de controlar/excluir, debido a que es una característica intrínseca de la detección basada en la hibridación. Debido a estos hechos, es difícil realizar la comparación de los resultados obtenidos usando chips de ADN de diferentes proveedores (que reflejan diferentes tecnologías de producción subyacentes y protocolos de aplicación) (Yauk et al., 2005, *Nucleic Acids Research*, vol. 32(15):e124). Dentro de una plataforma, es necesaria la validación de los resultados mediante un procedimiento independiente tal como ensayos de PCR en tiempo real (p. ej., TaqMan, Invader). Por tanto, los chips de ADN no proporcionan datos que se ajusten al concepto de un northern digital, pero son útiles para la determinación de los niveles de expresión relativos si se usa la misma plataforma para todas las

muestras.

Idealmente, una tecnología de elaboración de perfiles de transcritos es altamente multiplexada, es decir, se pueden estudiar muchos genes simultáneamente, de alto rendimiento, muy sólida y reproducible, altamente precisa (no presenta hibridación cruzada) y aplicable sin la necesidad de información de secuencia previa. La invención descrita a continuación proporciona procedimientos que cumplen tales criterios.

## Sumario de la invención

Los presentes inventores han descubierto ahora que este problema se puede resolver con una estrategia diferente y que las tecnologías de secuenciación de alto rendimiento se pueden usar de forma eficaz en la elaboración de perfiles de transcritos.

La invención comprende emplear una tecnología que, preferentemente, divide el transcriptoma en subconjuntos reproducibles. Los subconjuntos se secuencian y se ensamblan en cóntigos que corresponden a transcritos individuales. Mediante la repetición de esta etapa de manera que se proporcione un subconjunto reproducible diferente, se obtienen diferentes conjuntos de cóntigos. Estos cóntigos diferentes se usan para ensamblar las secuencias proyecto de los transcritos. La invención no requiere conocimiento alguno de la secuencia y se puede aplicar a transcritos de cualquier complejidad. La invención también se puede aplicar a una combinación de transcritos, p. ej., derivados de diferentes tejidos del mismo organismo u organismos diferentes. La presente invención proporciona un acceso más ágil, fiable y rápido a cualquier transcrito de interés y, de este modo, permite acelerar el análisis de los transcritos.

La invención también se refiere a la determinación (no sesgada) de niveles relativos de transcritos de genes sin que sea necesaria información de secuencia de estos genes. Con este fin, se determina la frecuencia de una secuencia dentro de una muestra de ADNc mediante secuenciación de colecciones de complejidad reducida de dicha muestra de ADNc y hibridación de la secuencia para determinar el número de veces que se identifica la secuencia en las colecciones. Esto puede repetirse para una segunda muestra de ADNc y se pueden normalizar las frecuencias de las dos muestras de ADNc, en caso necesario, y compararse para determinar los niveles de transcripción relativos.

## Definiciones

En la siguiente descripción y ejemplos se usan una serie de términos. Las definiciones siguientes se proporcionan con el fin de proporcionar una comprensión clara y consistente de la memoria descriptiva y las reivindicaciones, incluido el alcance que se da a tales términos. A menos que se definan de otro modo en el presente documento, todos los términos técnicos y científicos usados tienen el mismo significado que entiende comúnmente un experto en la técnica a la que pertenece la presente invención. Las divulgaciones de todas las publicaciones, solicitudes de patente, patentes y otras referencias se incorporan en el presente documento en su totalidad por referencia.

Ácido nucleico: un ácido nucleico de acuerdo con la presente invención puede incluir cualquier polímero u oligómero de bases de pirimidina y purina, preferentemente 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 por referencia en su totalidad para todos los fines). La presente invención contempla cualquier desoxirribonucleótido, ribonucleótido o componente de ácido nucleico peptídico, y cualquiera de sus variantes químicas, tales como formas metiladas, hidroximetiladas o glucosiladas de estas bases, y similares. Los polímeros u oligómeros pueden ser heterogéneos u homogéneos en su composición y pueden aislarse a partir de fuentes naturales o pueden producirse artificialmente o sintéticamente. Además, los ácidos nucleicos pueden ser ADN o ARN, o una mezcla de ellos, y pueden existir de forma permanente o transitoria, en forma monocatenaria o bicatenaria, incluidos estados homodúplex, heterodúplex e híbridos.

Reducción de la complejidad: el término reducción de la complejidad se usa para designar un procedimiento 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, complejo) y, preferentemente, es un subconjunto reproducible. En este contexto, reproducible significa que cuando se reduce la complejidad de la misma muestra usando el mismo procedimiento, se obtiene el mismo subconjunto, o al menos uno comparable. El procedimiento usado para la reducción de la complejidad puede ser cualquier procedimiento de reducción de la complejidad conocido en la técnica. Los ejemplos no limitantes de procedimientos de reducción de la complejidad incluyen AFLP® (Keygene N.V., Países Bajos; véase, p. ej., el documento EP 0 534 858), los procedimientos descritos por Dong (véanse, p. ej., los documentos WO 03/012118 y WO 00/24939), enlace indexado (Unrau, et al., 1994, Gene, 145:163-169), los descritos en los documentos US 2005/260628, WO 03/010328 y US 2004/10153, fraccionamiento del genoma (véase, p. ej., el documento WO 2004/022758), análisis en serie de la expresión génica (SAGE; véanse, p. ej. Velculescu et al., 1995, véase anteriormente, y Matsumura et al., 1999, The Plant Journal, vol. 20(6):719-726) y modificaciones del SAGE (véanse, p. ej., Powell, 1998, Nucleic Acids Research, vol. 26(14):3445-3446; y Kenzelmann y Mühlemann, 1999, Nucleic Acids Research, vol. 27(3):917-918), microSAGE (véase, p. ej., Datson et al., 1999, Nucleic Acids Research, vol. 27(5):1300-1307), secuenciación masiva de firmas en paralelo (MPSS; véanse, p. ej., Brenner et al., 2000, Nature Biotechnology, vol. 18:630-634 y Brenner et al., 2000, PNAS, vol. 97(4):1665-1670), colecciones de ADNc autosustraidas (Laveder et al., 2002,

Nucleic Acids Research, vol. 30(9):e38); amplificación múltiple de sondas dependiente de ligación en tiempo real (RT-MLPA; véase, p. ej., Eldering et al., 2003, vol. 31(23):e153), elaboración de perfiles de expresión de alta cobertura (HiCEP; véase, p. ej., Fukumura et al., 2003, Nucleic Acids Research, vol. 31(16):e94), un sistema de micromatrices universal como se divulga en Roth et al., 2004, Nature Biotechnology, vol. 22(4): 418-426, un procedimiento de sustracción de transcriptoma (véase p. ej. Li et al., Nucleic Acids Research, vol. 33(16):e136) y presentación de fragmentos (véase p. ej. Metsis et al., 2004, Nucleic Acids Research, vol. 32(16):e127). Los procedimientos de reducción de la complejidad usados en la presente invención tienen en común que son reproducibles. Reproducibles 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 contraposición con una reducción de la complejidad más aleatoria tal como la microdissección o el uso de ARNm (ADNc) que representa una parte del genoma transcrito en un tejido seleccionado y cuya reproducibilidad depende de la selección del tejido, el tiempo de aislamiento y similares.

Marcaje: el término marcaje se refiere a la adición de una marca a una muestra de ácido nucleico con el fin de poder distinguirla de una segunda muestra o una muestra posterior de ácido nucleico. El marcaje se puede realizar, p. ej., mediante la adición de un identificador de secuencia durante la reducción de la complejidad o mediante cualquier otro procedimiento conocido en la técnica. Tal identificador de secuencia puede ser, p. ej., una secuencia de bases única de longitud variable pero definida que se usa exclusivamente para identificar una muestra de ácido nucleico específica. Ejemplos típicos de éstos son, por ejemplo, las secuencias ZIP. Con el uso de una marca de este tipo, se puede determinar el origen de una muestra tras su procesamiento adicional. En caso de combinar productos procesados que se originan a partir de muestras de ácido nucleico diferentes, se deberían identificar las diferentes muestras de ácido nucleico usando diferentes marcas.

Colección marcada: el término colección marcada se refiere a una colección de ácido nucleico marcado.

Secuenciación: el término secuenciación se refiere a la determinación del orden de los nucleótidos (secuencias de bases) en una muestra de ácido nucleico, p. ej., ADN o ARN.

Alinear y hibridación: el término "alinear" y "hibridación" quiere decir la comparación de dos o más secuencias de nucleótidos basada en la presencia de tramos cortos o largos de nucleótidos idénticos o similares. En la técnica se conocen varios procedimientos para la hibridación de secuencias de nucleótidos, como se explicará en detalle más adelante. En ocasiones, se usan los términos 'ensamblaje' o 'agrupación' como sinónimos, aunque estos términos no son técnicamente idénticos. La hibridación se produce basándose en la comparación de la máxima homología, mientras que el ensamblaje significa preparar un cóntigo basado en un solapamiento.

Rastreo de alto rendimiento: el rastreo de alto rendimiento, abreviado frecuentemente como HTS, es un procedimiento para experimentación científica especialmente pertinente a los ámbitos de la biología y la química. A través de una combinación de robots modernos y otros equipos de laboratorio especializados, permite a un investigador rastrear grandes cantidades de muestras de forma simultánea.

Secuenciación de alto rendimiento: determinación de la secuencia de una secuencia de nucleótidos usando técnicas de alto rendimiento.

Endonucleasa de restricción: una endonucleasa de restricción o enzima de restricción es una enzima que reconoce una secuencia de nucleótidos específica (sitio objetivo) en una molécula de ADN bicatenario y escindiría ambas hebras de la molécula de ADN en cada sitio objetivo.

Fragmentos de restricción: las moléculas de ADN producidas por digestión con endonucleasa de restricción se denominan fragmentos de restricción. Cualquier genoma dado (o ácido nucleico, independientemente de su origen) será digerido por una endonucleasa de restricción concreta en un conjunto discreto de fragmentos de restricción. Los fragmentos de ADN que resultan de la escisión con endonucleasas de restricción se pueden usar adicionalmente en una variedad de técnicas y se pueden detectar, por ejemplo, por electroforesis en gel.

Electroforesis en gel: con el fin de detectar fragmentos de restricción, puede ser necesario un procedimiento analítico para fraccionar moléculas de ADN bicatenario basándose en su tamaño. La técnica usada más comúnmente para lograr dicho fraccionamiento es la electroforesis en gel (por capilaridad). La velocidad a la que se mueven los fragmentos de ADN en geles de este tipo depende de su peso molecular; así, las distancias recorridas disminuyen a medida que aumenta la longitud del fragmento. Los fragmentos de ADN fraccionados por electroforesis en gel se pueden visualizar directamente mediante un procedimiento de tinción, p. ej., tinción con plata o tinción usando bromuro de etidio, si el número de fragmentos incluidos en el patrón es lo suficientemente pequeño. De forma alternativa, el tratamiento adicional de los fragmentos de ADN puede incorporar marcadores detectables en los fragmentos, tales como fluoróforos o marcadores radioactivos.

Ligación: la reacción enzimática catalizada por una enzima ligasa en la que se unen covalentemente dos moléculas de ADN bicatenario se denomina ligación. En general, ambas hebras de ADN se unen covalentemente, pero también se puede impedir la ligación de una de las dos hebras a través de la modificación química o enzimática de uno de los extremos de las hebras. En ese caso, la unión covalente se producirá sólo en una de las dos hebras de ADN.

Oligonucleótido sintético: las moléculas de ADN monocatenario que tienen preferentemente de aproximadamente 10 a aproximadamente 50 bases, que se pueden sintetizar químicamente, se denominan oligonucleótidos sintéticos. En general, estas moléculas de ADN sintéticas se diseñan para que tengan una secuencia de nucleótidos única o deseada, aunque se pueden sintetizar familias de moléculas que tiene secuencias relacionadas y que tienen composiciones de nucleótidos diferentes en posiciones específicas dentro de la secuencia de nucleótidos. El término oligonucleótido sintético se usará para hacer referencia a moléculas de ADN que tienen una secuencia de nucleótidos diseñada o deseada.

Adaptadores: moléculas de ADN bicatenario cortas con un número limitado de pares de bases, p. ej., de aproximadamente 10 a aproximadamente 30 pares de bases de longitud, que se diseñan de forma que se pueden ligar a los extremos de fragmentos de restricción. En general, los adaptadores se componen de dos oligonucleótidos sintéticos, que tienen secuencias de nucleótidos que son parcialmente complementarias entre sí. Al mezclar los dos oligonucleótidos sintéticos en solución bajo condiciones apropiadas, se hibridarán entre sí formando una estructura bicatenaria. Después de la hibridación, un extremo de la molécula adaptadora está diseñado de forma que es compatible con el extremo de un fragmento de restricción y se puede ligar a él; el otro extremo del adaptador se puede diseñar para que no se pueda ligar, pero no es necesariamente el caso (adaptadores doblemente ligados).

Fragmentos de restricción ligados a adaptador: fragmentos de restricción que se han rematado mediante adaptadores como consecuencia de la ligación.

Cebadores: en general, el término cebadores se refiere a una hebra de ADN que puede cebar la síntesis de ADN. La polimerasa de ADN 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 los nucleótidos que se van a ensamblar. Se hará referencia a las moléculas de oligonucleótido sintéticas que se usan en una reacción en cadena de la polimerasa (PCR) como cebadores.

Amplificación de ADN: el término amplificación de ADN se usará normalmente para designar la síntesis *in vitro* de moléculas de ADN bicatenario usando PCR. Cabe destacar que existen otros procedimientos de amplificación y se puede usar en la presente invención sin apartarse de su esencia.

En Volkmuth et al, OMICS, Journal of Integrative biology vol 7, P143-159, 2003 se divulga un procedimiento para el análisis de la expresión cuantitativa en *A. thaliana* basado en ADNc-AFLP como tecnología de reducción de la complejidad seguida de secuenciación de Sanger.

### Descripción detallada de la invención

La presente invención proporciona un procedimiento para determinar una secuencia de nucleótidos de ADNc que comprende las etapas de:

(a) Proporcionar una primera muestra de ADNc;

(b) Realizar una reducción de la complejidad reproducible de la primera muestra de ADNc para obtener una primera colección que comprende las etapas de:

- digerir el ADNc con al menos una endonucleasa de restricción para fragmentarlo en fragmentos de restricción;

- ligar los fragmentos de restricción con al menos un adaptador oligonucleotídico sintético bicatenario que tenga 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 bajo condiciones de hibridación, teniendo dichos uno o más cebadores oligonucleotídicos una secuencia de cebador que incluye una sección de secuencia de nucleótidos complementaria a parte del al menos un adaptador y a parte de la parte restante de la secuencia de reconocimiento de la endonucleasa de restricción; y

- amplificar dichos fragmentos de restricción ligados a adaptador por elongación del uno o más cebadores oligonucleotídicos hibridados;

(c) Marcar la primera colección para obtener una primera colección marcada mediante el uso de al menos un adaptador marcado en la etapa (b);

(d) Realizar, de forma consecutiva o simultáneamente, las etapas (a) y (b) con una segunda muestra y/o una muestra posterior de ADNc, usando una marca diferente para cada muestra de ADNc, para obtener una segunda colección marcada y/o una colección marcada adicional;

(e) Combinar la primera colección marcada y la segunda colección marcada y/o una colección marcada adicional para obtener una colección combinada;

(f) Determinar al menos parte de las secuencias de nucleótidos de la colección marcada mediante secuenciación de alto rendimiento;

(g) Determinar la frecuencia de la secuencia de nucleótidos en la primera muestra de ADNc y la segunda muestra y/o una muestra adicional de ADN; y

5 (h) Comparar la frecuencia de la secuencia de nucleótidos en la primera muestra de ADNc con la frecuencia de la secuencia de nucleótidos en la segunda muestra y/o una muestra posterior de ADNc para obtener niveles de transcripción relativos de la secuencia de nucleótidos en las muestras de ADNc. Hasta el momento, en la técnica de la tecnología de secuenciación, no se ha divulgado o sugerido el uso de esta reducción de la complejidad en combinación con la determinación de secuencia de ADNc de alto rendimiento para representar transcritos.

10 En la etapa (a) del procedimiento, se proporciona ADNc. En la técnica se conoce bien cómo preparar ADNc. Más adelante se expone un procedimiento para la preparación. Sin embargo, se puede usar cualquier procedimiento para la preparación de ADNc.

Habitualmente, el ADNc (ADN complementario) se prepara a partir de ARNm usando la transcriptasa inversa. En ese caso, la transcriptasa inversa sintetiza una hebra de ADN complementaria a un molde de ARN si está provista de un  
15 cebador que forma pares de bases con el ARN y contiene un grupo 3'-OH libre. Un cebador de este tipo puede ser, p. ej., un cebador de oligo-dT que se aparea con la secuencia de poli-A del extremo 3' de la mayoría de las moléculas de ARNm eucariotas. El resto de la hebra de ADNc se puede sintetizar después en presencia de los cuatro desoxirribonucleósidos trifosfato. Posteriormente, se hidroliza la hebra de ARN del híbrido de ARN-ADN resultante, p. ej., elevando el pH. A diferencia del ARN, el ADN es resistente a la hidrólisis alcalina, de modo que la  
20 hebra de ADN permanece intacta. Un cebador alternativo puede ser un cebador aleatorio. El cebado aleatorio de ADNc puede ser beneficioso cuando la transcriptasa inversa no es capaz de transcribir totalmente un molde de ARNm o si existen estructuras secundarias. Otro cebador alternativo puede ser un cebador específico de secuencia.

En la técnica de la biología molecular se conocen bien procedimientos para el aislamiento de ARN a partir de células de un tejido de un organismo o un organismo mismo. Además, muchos kits para síntesis de ADNc comercialmente  
25 disponibles se pueden comprar, tal como, p. ej., de ABgene, Ambion, Applied Biosystems, BioChain, Bio-Rad, Clontech, GE Healthcare, GeneChoice, Invitrogen, Novagen, Qiagen, Roche Applied Science, Stratagene y similares. Procedimientos de este tipo se describen, p. ej., en Sambrook et al. (Sambrook, J., Fritsch, E.F., y Maniatis, T., en *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, NY, Vol. 1, 2, 3 (1989)). El ARN se puede aislar a partir de varias fuentes tales como un cultivo celular, un tejido, etc.

30 En la etapa (b) del procedimiento de acuerdo con la presente invención, se realiza una reducción de la complejidad sobre al menos una porción del ADNc para obtener una primera colección del ADNc que comprende fragmentos de ADNc. En la técnica se conocen muchos procedimientos para reducir la complejidad, como se indica en la sección de definiciones.

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  
35 fragmentos de restricción y seleccionar un grupo de fragmentos de restricción en particular. Opcionalmente, los fragmentos seleccionados se ligan después a secuencias adaptadoras que contienen moldes de cebadores de PCR/secuencias de unión.

En una realización de reducción de la complejidad, se usa una endonucleasa de tipo 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 el saliente que se va a ligar y sólo se liga al fragmento y se amplifica posteriormente el adaptador con el conjunto de nucleótidos compatibles en el saliente. Esta tecnología se representa en la técnica como 'enlazadores de indexación'. Se pueden encontrar ejemplos de este principio, entre  
40 otros, en Unrau y Deugau (1994) *Gene* 145:163-169.

En una realización, el procedimiento de reducción de la complejidad utiliza dos endonucleasas de restricción que tienen diferentes sitios objetivo y frecuencias y dos secuencias adaptadoras diferentes para proporcionar fragmentos de restricción ligados a adaptador, tal como en el AFLP.

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

50 En una realización de la invención, la etapa de reducción de la complejidad comprende eliminar secuencias repetidas desnaturalizando y realineando el ADN y eliminando después los dúplex bicatenarios.

En determinadas realizaciones de la invención, la etapa de reducción de la complejidad comprende hibridar la muestra de ácido nucleico a una perla magnética que está unida a una sonda de oligonucleótidos que contiene una  
55 secuencia deseada. Esta realización puede comprender además la exposición de la muestra hibridada a una nucleasa de ADN monocatenario para eliminar el ADN monocatenario, activando una secuencia adaptadora que contiene una enzima de restricción de 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 la PCR. En esta realización, la secuencia adaptadora puede contener o no una marca o un identificador de secuencia.

5 En determinadas realizaciones de la invención, la reducción de la complejidad utiliza tecnología de presentación diferencial o tecnología READS (Gene Logic).

En determinadas realizaciones de la invención, el procedimiento de reducción de la complejidad comprende exponer la muestra de ADN a una proteína de unión de apareamiento erróneo y digerir la muestra con una exonucleasa de 3' a 5' y después una nucleasa monocatenaria. Esta realización puede incluir o no el uso de una perla magnética unida a la proteína de unión de apareamiento erróneo.

10 En una realización de la presente invención, la reducción de la complejidad comprende el procedimiento CHIP como se describe en otros puntos del presente documento 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 teloméricas, genes de caja MADS, familias de genes de ATP-asa y otras familias de genes.

15 En la etapa (c) se determinan al menos parte de las secuencias de nucleótidos de los fragmentos de ADNc de la primera colección por secuenciación de alto rendimiento. Ejemplos no limitantes de procedimientos de secuenciación de alto rendimiento son los procedimientos divulgados 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, etcétera.  
20 Lo más preferido es que la secuenciación se realice usando el aparato y/o el procedimiento divulgado 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 un solo ciclo y es 100 veces más rápida y más barata que la tecnología competidora basada en la secuenciación de Sanger y los instrumentos de electroforesis por capilaridad disponibles actualmente tales como  
25 MegaBACE (GE Healthcare) o ABI3700 (xl) (Applied Biosystems). La tecnología de secuenciación consiste, a grandes rasgos, en 4 etapas: 1) fragmentar el DNA y ligar adaptadores específicos a una colección de ADN monocatenario (ADNss); 2) alinear el ADNss con perlas y emulsionar las perlas en microrreactores de agua en aceite; 3) depositar las perlas portadoras de DNA en una placa PicoTiter®; y 4) secuenciación simultánea en varios pocillos mediante la generación de una señal lumínica de pirofosfato. El procedimiento se explicará con más detalle  
30 a continuación.

En la etapa (d) se alinean las secuencias de nucleótidos de los fragmentos de ADNc de la primera colección de la etapa (d) para generar cóntigos de la primera colección.

35 Mediante la construcción de cóntigos a partir de secuencias, el procedimiento de ensamblaje será computacionalmente menos complejo y, por lo tanto, más rápido de realizar. Alineando las secuencias de la colección, se pueden construir cóntigos para cada fragmento de restricción del conjunto de fragmentos de restricción para cada combinación de cebador. Esto da lugar a un conjunto de cóntigos, cada uno correspondiente a un fragmento de restricción en particular. Como consecuencia, cada fragmento obtenido a partir de la restricción del ADNc con la al menos una endonucleasa de restricción tiene ahora una secuencia determinada (cóntigo).

40 En la técnica se conocen bien procedimientos de hibridación de secuencias con fines de comparación. Diversos programas y algoritmos de hibridación no limitantes se describen 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. EE.UU. 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 (incorporado en el presente documento por  
45 referencia) presenta una consideración detallada de los procedimientos de hibridación de secuencias y cálculos de homología.

La herramienta de búsqueda de hibridación local básica del NCBI (BLAST) (Altschul et al., 1990) está disponible de varias fuentes, incluido el National Center for Biological Information (NCBI, Bethesda, Md.) y en Internet, para su uso en relación con los programas de análisis de secuencia blastp, blastn, blastx, tblastn y tblastx. Se puede acceder a ella en <<http://www.ncbi.nlm.nih.gov/BLAST/>>. Una descripción de cómo determinar la identidad de secuencias usando este programa está disponible en <[http://www.ncbi.nlm.nih.gov/BLAST/blast\\_help.html](http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html)>. Una aplicación  
50 adicional puede ser en detección de microsatélites (véase Varshney et al. (2005) Trends in Biotechn. 23(1):48-55).

55 En una realización, la hibridación se realiza sobre datos de secuencia que se han recortado para los adaptadores/el cebador y/o los identificadores, pero con secuencias de reconocimiento de enzimas de restricción reconstruidas, es decir, usando sólo los datos de secuencia de los fragmentos que derivan del ADNc. Normalmente, los datos de secuencia obtenidos se usan para identificar el origen del fragmento (es decir, de qué muestra), las secuencias derivadas de la secuencia adaptadora y/o identificadora se eliminan de los datos y se realiza la hibridación en este conjunto recortado.



En la etapa (e), se determina la secuencia de nucleótidos del ADNc, p. ej., ensamblando las secuencias.

Dicho procedimiento es útil, p. ej., para determinar el número de secuencias diferentes presentes en un ADNc o una fracción de complejidad reducida de dicho ADNc, o para descubrir la expresión de determinados genes.

5 En una realización, la etapa (a) comprende las etapas de: i) proporcionar una muestra biológica; ii) aislar el ARNm o el RNA total a partir de la muestra biológica; iii) sintetizar ADNc a partir del ARNm o el RNA total.

En una realización, la secuenciación de alto rendimiento se realiza sobre un soporte sólido tal como un perla (véanse, p. ej., 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)). Un procedimiento de secuenciación de este tipo es particularmente adecuado para la secuenciación barata y eficaz de muchas muestras simultáneamente.

10 En una realización adicional, la secuenciación de alto rendimiento se basa en la secuenciación por síntesis, preferentemente en la pirosecuenciación. La pirosecuenciación se conoce en la técnica y se describe, entre otros, en [www.biotagebio.com](http://www.biotagebio.com); [www.pyrosequencing.com](http://www.pyrosequencing.com), sección technology. Además, la tecnología se aplica, p. ej., 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 Life Sciences). Es una técnica rápida y altamente reproducible, particularmente adecuada para la secuenciación de alto rendimiento.

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

(c1) ligar adaptadores de secuenciación a los fragmentos;

(c2) alinear fragmentos ligados a adaptadores de secuenciación con perlas, alineándose cada perla con un solo fragmento;

20 (c3) emulsionar las perlas en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una sola perla;

(c4) realizar una PCR de la emulsión para amplificar fragmentos ligados a adaptadores de secuencia sobre la superficie de las perlas;

25 (c5) seleccionar / enriquecer perlas que contienen fragmentos ligados a adaptadores de secuenciación amplificados;

(c6) cargar las perlas en pocillos, comprendiendo cada pocillo una sola perla; y

(c7) generar una señal de pirofosfato.

30 En la etapa c1), se ligan adaptadores de secuenciación a los fragmentos de la colección. Dicho adaptador de secuenciación incluye al menos una región "clave" para su hibridación con una perla, una región de cebador de secuenciación y una región de cebador de PCR. Por tanto, se obtienen fragmentos adaptados.

En la etapa (c2), se alinean fragmentos ligados a adaptadores de secuenciación con perlas, alineándose cada perla con un solo fragmento. A los fragmentos ligados a adaptador de secuenciación agrupados se les añaden perlas en exceso con el fin de garantizar la hibridación de un sólo fragmento adaptado por perla para la mayoría de las perlas (distribución de Poisson).

35 En la etapa (c3), se emulsionan las perlas en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una sola perla.

En la etapa c4), se realiza una PCR de la emulsión para amplificar los fragmentos ligados a adaptadores de secuenciación sobre la superficie de las perlas. Los reactivos de PCR están presentes en los microrreactores de agua en aceite, lo que permite que se produzca una reacción de PCR dentro de los microrreactores.

40 En la etapa c5) se seleccionan/enriquecen las perlas que contienen fragmentos ligados a adaptadores de secuenciación amplificados.

En la etapa c6), se cargan las perlas en pocillos, comprendiendo cada pocillo una sola perla. Preferentemente, los pocillos forman parte de una placa PicoTiter™, lo que permite la secuenciación simultánea de una gran cantidad de fragmentos. Después de la adición de perlas portadoras de enzima, se determina la secuencia de los fragmentos usando pirosecuenciación.

45 En la etapa c7), se genera una señal de pirofosfato. En etapas sucesivas, la placa PicoTiter™ y las perlas, así como las perlas de enzima, se someten a diferentes desoxirribonucleótidos en presencia de reactivos de secuenciación convencionales y, tras la incorporación de un desoxirribonucleótido, se genera una señal lumínica que se registra. La incorporación del nucleótido correcto generará una señal de pirosecuenciación que se puede detectar por medios conocidos en la técnica.

En una realización preferida del procedimiento de acuerdo con la presente invención, se realiza la reducción de la complejidad mediante un procedimiento que comprende las etapas de:

- i). Digerir el ADNc con al menos una endonucleasa de restricción para fragmentarlo en fragmentos de restricción;
- 5 ii). Ligar los fragmentos de restricción con al menos un adaptador oligonucleotídico sintético bicatenario que tenga un extremo compatible con uno o ambos extremos de los fragmentos de restricción para producir fragmentos de restricción ligados a adaptador;
- 10 iii). Poner en contacto dichos fragmentos de restricción ligados a adaptador con uno o más cebadores oligonucleotídicos bajo condiciones de hibridación, teniendo dichos uno o más cebadores oligonucleotídicos una secuencia de cebador que incluye una sección de secuencia de nucleótidos complementaria a parte del al menos un adaptador y a parte de la parte restante de la secuencia de reconocimiento de la endonucleasa de restricción;
- y
- iv). Amplificar dichos fragmentos de restricción ligados a adaptador por elongación del uno o más cebadores oligonucleotídicos hibridados.

El procedimiento anterior de reducción de la complejidad también se denomina AFLP® (Keygene N.V., Países Bajos; véanse, p. ej., 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, 4407-4414. El AFLP es un procedimiento altamente reproducible para reducir la complejidad y, por lo tanto, es particularmente adecuado para el procedimiento de acuerdo con la presente invención. El AFLP es un procedimiento para la amplificación selectiva de fragmentos de restricción. El AFLP no requiere ninguna información de secuencia previa y se puede realizar sobre cualquier ADNc de partida.

Por tanto, el AFLP proporciona un subconjunto reproducible de fragmentos ligados a adaptador. Una variante útil de la tecnología de AFLP no usa nucleótidos selectivos (es decir, cebadores +0/+0) y en ocasiones se denomina PCR de enlazador. Esto también permite una reducción de la complejidad muy adecuada, en particular para transcritos y ADNc obtenido a partir de ellos.

En la etapa i), se digiere el ADNc con al menos una endonucleasa de restricción para fragmentarlo en fragmentos de restricción. En determinadas realizaciones, se usan al menos dos endonucleasas de restricción. En otras realizaciones, se pueden usar tres o más endonucleasas de restricción. Las endonucleasas de restricción pueden ser cortadoras frecuentes (es decir, normalmente cortadoras de 4 y 5, es decir, endonucleasas de restricción que tienen una secuencia de reconocimiento de 4 o 5 nucleótidos, respectivamente) o pueden ser cortadoras poco frecuentes (es decir, que normalmente tienen un sitio de reconocimiento de 6 o más nucleótidos, respectivamente), o sus combinaciones. En determinadas realizaciones, se puede usar una combinación de una cortadora poco frecuente y una frecuente. Las endonucleasas de restricción pueden ser de cualquier tipo, incluidos los tipos IIs y IIsa, que cortan el ADNc fuera de su secuencia de reconocimiento, en uno o en ambos lados de la secuencia de reconocimiento.

En la etapa ii), se ligan los fragmentos de restricción 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. Preferentemente, los adaptadores son tales que el sitio de reconocimiento de la endonucleasa no se restablece tras la ligación del adaptador. También es posible emplear dos o más adaptadores diferentes, por ejemplo, en el caso de que se usen dos o más endonucleasas de restricción en la etapa i). Esta etapa de ligación proporciona fragmentos de restricción ligados a adaptador. Los adaptadores pueden ser de extremos romos o pueden contener un saliente, en función de la(s) endonucleasa(s) de restricción que se use(n) en la etapa i).

En determinadas realizaciones, el adaptador puede ser un conjunto de adaptadores conocidos como enlazadores de indexación (Unrau, et al., 1994, Gene, 145:163-169).

En la etapa iii), dichos fragmentos ligados a adaptador se ponen en contacto con uno o más cebadores oligonucleotídicos bajo condiciones de hibridación. El uno o más cebadores oligonucleotídicos tienen una secuencia de cebador que incluye una sección de secuencia de nucleótidos complementaria a parte del al menos un adaptador y a parte de la parte restante de la secuencia de reconocimiento de la endonucleasa de restricción.

Las condiciones de hibridación estándar son condiciones para hibridación selectiva. Hibridación selectiva se refiere a hibridación, bajo condiciones de hibridación rigurosas, de una secuencia de ácido nucleico con una secuencia objetivo de ácido nucleico específica hasta un grado detectable mayor (p. ej., al menos 2 veces más que el nivel de base) que su hibridación con secuencias de ácido nucleico que no son objetivo y a la exclusión sustancial de ácidos nucleicos que no son objetivo. Los términos "condiciones rigurosas" o "condiciones de hibridación rigurosas" incluyen referencias a las condiciones bajo las cuales una sonda hibridará con su secuencia objetivo, en un grado detectable mayor que con otras secuencias (p. ej., al menos 2 veces más que el nivel de base). Las condiciones rigurosas dependen de la secuencia y serán diferentes en circunstancias diferentes. Controlando la rigurosidad de las condiciones de hibridación y/o de lavado, se pueden identificar secuencias objetivo que son complementarias al 100 % con la sonda (sondeo homólogo). De forma alternativa, se pueden ajustar las condiciones de rigurosidad para

permitir algunos apareamientos erróneos en secuencias, de forma que se detecten grados de similitud inferiores (sondeo heterólogo). En general, una sonda tiene menos de aproximadamente 100 nucleótidos de longitud, opcionalmente, no más de 50 o 25 nucleótidos de longitud. Normalmente, condiciones rigurosas serán aquellas en las que la concentración de sal es de menos de aproximadamente 1,5 M de iones Na, normalmente una concentración de aproximadamente 0,01 a 1,0 M de iones Na (u otras sales) a pH de 7,0 a 8,3 y la temperatura es de al menos aproximadamente 30 °C para sondas cortas (p. ej., de 10 a 50 nucleótidos) y de al menos aproximadamente 60 °C para sondas largas (p. ej., mayores de 50 nucleótidos). Las condiciones rigurosas también se pueden lograr con la adición de agentes desestabilizantes tales como formamida. Las condiciones poco rigurosas ejemplares incluyen hibridación con una solución tamponadora de formamida del 30 al 35 %, NaCl 1 M, SDS al 1 % (dodecilsulfato de sodio) a 37 °C y un lavado en SSC de 1x a 2x (20xSSC = NaCl 3,0 M/citrato trisódico 0,3 M) a de 50 a 55 °C. Las condiciones moderadamente rigurosas ejemplares incluyen hibridación en formamida del 40 al 45 %, NaCl 1 M, SDS al 1 % a 37 °C y un lavado en SSC de 0,5x a 1x a de 55 a 60 °C. Las condiciones muy rigurosas ejemplares incluyen hibridación en formamida al 50 %, NaCl 1 M, SDS al 1 % a 37 °C y un lavado en SSC 0,1x a de 60 a 65 °C. Normalmente, la especificidad es la función de los lavados post hibridación, siendo los factores críticos la fuerza iónica y la temperatura de lavado final. Para híbridos de ADN-ADN, se puede aproximar la T<sub>m</sub> a partir de la ecuación de Meinkoth y Wahl, Anal. Biochem., 138:267-284 (1984):  $T_m = 81,5\text{ }^{\circ}\text{C} + 16,6 (\log M) + 0,41 (\% \text{ de GC}) - 0,61 (\% \text{ de form}) - 500 / L$ ; donde M es la molaridad de cationes monovalentes, % de GC es el porcentaje de nucleótidos de guanosina y citosina en el ADN, % de form es el porcentaje de formamida en la solución de hibridación y L es la longitud del híbrido en pares de bases. La T<sub>m</sub> es la temperatura (a fuerza iónica y pH definidos) a la que el 50 % de una secuencia objetivo complementaria hibrida con una sonda perfectamente apareada. La T<sub>m</sub> se reduce en aproximadamente un 1 °C por cada 1 % de apareamiento erróneo; por tanto, la T<sub>m</sub> y las condiciones de hibridación y/o de lavado se pueden ajustar para hibridar secuencias de la identidad deseada. Por ejemplo, si se buscan secuencias con >90 % de identidad, se puede reducir la T<sub>m</sub> aproximadamente 10 °C. En general, se seleccionan condiciones rigurosas aproximadamente 5 °C más bajas que el punto de fusión térmico (T<sub>m</sub>) para la secuencia específica y su complemento a una fuerza iónica y pH definidos. Sin embargo, las condiciones altamente rigurosas pueden utilizar una hibridación y/o un lavado a 1, 2, 3 o 4 °C menos que el punto de fusión térmica (T<sub>m</sub>); las condiciones moderadamente rigurosas pueden utilizar una hibridación y/o un lavado a 6, 7, 8, 9 o 10 °C menos que el punto de fusión térmica (T<sub>m</sub>); las condiciones poco rigurosas pueden utilizar una hibridación o un lavado a 11, 12, 13, 14, 15 o 20 °C menos que el punto de fusión térmica (T<sub>m</sub>). Usando la ecuación, las composiciones de hibridación y de lavado y la T<sub>m</sub> deseada, los expertos en la técnica entenderán que las variaciones de la rigurosidad de las soluciones de hibridación y/o de lavado se describen de manera inherente. Si el grado deseado de apareamiento erróneo da lugar a una T<sub>m</sub> de menos de 45 °C (solución acuosa) o 32 °C (solución de formamida), se prefiere aumentar la concentración de SSC de forma que se pueda usar una temperatura más alta. Una guía extensa para la hibridación de ácidos nucleicos se encuentra en Tijssen, Laboratory Techniques in Biochemistry and Molecular Biology-Hybridization with Nucleic Acid Probes, parte 1, capítulo 2 "Overview of principles of hybridization and the strategy of nucleic acid probe assays", Elsevier, N.Y. (1993); y en Current Protocols in Molecular Biology, capítulo 2, Ausubel, et al., Ed., Greene Publishing y Wiley-Interscience, Nueva York (1995).

Cuando se emplean dos o más endonucleasas de restricción, es probable que se usen dos o más cebadores oligonucleotídicos en la etapa iii), dependiendo del sitio de reconocimiento de la endonucleasa. El/los cebador(es) oligonucleotídico(s) tiene/tienen una secuencia de cebador que incluye una secuencia de nucleótidos complementaria con el al menos un adaptador y con parte de la parte restante de la secuencia de reconocimiento de la endonucleasa de restricción más, opcionalmente, la parte restante de la secuencia de reconocimiento de la endonucleasa de restricción, como se explica con más detalle en el documento EP 0 534 858 y en Vos et al. ((1995). AFLP: a new technique for DNA fingerprinting, Nucleic Acids Research, vol. 23, n.º 21, 4407-4414). Normalmente, la parte de la secuencia de reconocimiento es la parte que queda después de la restricción de la secuencia con la endonucleasa de restricción. En resumen, por lo tanto, el/los cebador(es) es/son complementario(s) al menos a la parte conocida de los fragmentos de restricción ligados a adaptador.

En la etapa iv), se amplifican dichos fragmentos de restricción ligados a adaptador por elongación del uno o más cebadores oligonucleotídicos hibridados. Preferentemente, la amplificación se lleva a cabo usando PCR, que es una técnica bien conocida en la técnica.

En una realización preferida de la invención, el cebador comprende además una secuencia seleccionada en el extremo 3' de la secuencia del cebador, comprendiendo dicha secuencia seleccionada 1-10 nucleótidos selectivos que son complementarios a una sección situada inmediatamente adyacente a la parte restante de la secuencia de reconocimiento de la endonucleasa de restricción. Normalmente, la parte de la secuencia de reconocimiento es la parte que queda después de la restricción de la secuencia con la endonucleasa de restricción. Preferentemente, en su extremo 3', el/los cebador(es) contiene(n) una secuencia seleccionada. La secuencia seleccionada comprende un conjunto de 1-10 nucleótidos seleccionados anteriormente, preferentemente 1-8 nucleótidos seleccionados, preferentemente 1-5, más preferentemente 1-3. Un cebador ejemplar puede tener la siguiente estructura ilustrativa (para 2 nucleótidos selectivos (AC)) "5'-región adaptadora específica-región específica de secuencia de restricción-AC-3'". Por tanto, este cebador ejemplar contiene 2 nucleótidos selectivos AC que sólo amplificarán fragmentos ligados a adaptador que contienen los TG complementarios como los dos primeros nucleótidos que siguen a la parte conocida de los fragmentos de restricción ligados a adaptador, es decir, que siguen a los restos del sitio de

reconocimiento de la endonucleasa de restricción.

Para una descripción adicional del AFLP, sus ventajas, sus realizaciones, así como las técnicas, enzimas, adaptadores, cebadores y compuestos y herramientas adicionales usados en él, 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.

En una realización, dicho adaptador comprende además una secuencia identificadora. Una secuencia identificadora de este tipo puede ser, p. ej., una secuencia de bases única de longitud variable usada para indicar el origen de la colección obtenida mediante reducción de la complejidad.

La presente invención también se refiere a un procedimiento para determinar la frecuencia de una secuencia de nucleótidos que comprende las etapas de:

- a) Proporcionar ADNc;
- b) Realizar una reducción de la complejidad sobre al menos una porción del ADNc para obtener una primera colección del ADNc que comprende fragmentos de ADNc;
- c) Determinar al menos parte de las secuencias de nucleótidos de los fragmentos de ADNc de la primera colección por secuenciación; y
- d) Determinar la frecuencia de una secuencia de nucleótidos.

En la etapa (a) del procedimiento, se proporciona ADNc. En la técnica se conoce bien cómo preparar ADNc y anteriormente se proporciona un procedimiento adecuado. El ADNc puede derivarse de cualquier fuente, como también se expone anteriormente.

En la etapa (b) del procedimiento, se realiza una reducción de la complejidad sobre al menos una porción del ADNc para obtener una primera colección del ADNc que comprende fragmentos de ADNc. La reducción de la complejidad se puede realizar mediante cualquier procedimiento conocido en la técnica, como se expone anteriormente.

En la etapa (c) del procedimiento de acuerdo con la invención, se determinan al menos parte de las secuencias de nucleótidos de los fragmentos de ADNc de la primera colección por secuenciación. La secuenciación se puede realizar mediante cualquier procedimiento conocido en la técnica, incluido el conocido procedimiento de Sanger (didesoxi). En una realización preferida, se realiza la secuenciación usando secuenciación de alto rendimiento, lo que permite la secuenciación simultánea de varias muestras. Los procedimientos preferidos para la secuenciación de alto rendimiento se exponen anteriormente.

En la etapa (d) del procedimiento de acuerdo con la invención, se determina la frecuencia de una secuencia de nucleótidos. La frecuencia de una secuencia de nucleótidos se puede determinar, p. ej., mediante el procedimiento siguiente. Se puede usar la hibridación de las secuencias de nucleótidos de fragmentos de ADNc para recoger secuencias de nucleótidos derivadas del mismo gen transcrito y para contar estas secuencias de nucleótidos. Queda pendiente establecer si las secuencias de nucleótidos derivan del mismo gen transcrito mediante homología entre las secuencias. Para los propósitos de la presente invención, se supone que las secuencias de nucleótidos derivan del mismo gen transcrito cuando son al menos el 95, 96, 97, 98, 99, 100 % homólogas en una longitud de al menos 10, preferentemente al menos 15, más preferentemente al menos 20, aún más preferentemente al menos 25, 30, 40, 50, 100, 150, 200 nucleótidos. El procedimiento puede apoyarse en interpretaciones estadísticas tales como una prueba de la T para demostrar estadísticamente diferentes frecuencias. También es posible elaborar una simple clasificación basada en el número de secuencias identificadas. Supóngase que en la muestra 1 una secuencia de nucleótidos del gen (desconocido) "X" se mide 10 veces (siendo 10 el número de secuencias de nucleótidos que tienen, p. ej., una homología de secuencia del 98 %) y en la muestra 2 la misma secuencia se mide 20 veces. En este caso, es probable que el nivel de transcripción del gen X en la muestra 2 sea el doble que en la muestra 1, con la condición de que el número total de secuencias determinadas para las muestras 1 y 2 sea el mismo; por lo tanto, la elaboración precisa de perfiles de transcritos, puede requerir normalización entre muestras y/o comparar las frecuencias de secuencias derivadas del gen "X" con las de los llamados genes de mantenimiento, cuyos niveles de transcripción relativos se suponen constantes a través de varias muestras. La clasificación de perfiles de transcripción relativa entre muestras en relación con las características fenotípicas de las muestras proporciona información sobre qué genes influyen en la aparición de diferentes fenotipos. El término fenotipos incluye todo tipo de características de un organismo, p. ej., estado de enfermedad, etcétera.

Para la evaluación estadística del número de secuencias de nucleótidos por gen (es decir, un northern digital) es importante garantizar la secuenciación redundante de los fragmentos de ADNc. Como tal, puede ser útil establecer una complejidad de colección de secuencias antes de realizar el experimento y ajustar según el número de lecturas de secuencia necesario para obtener suficientes secuencias. Por ejemplo, una muestra de ADNc típica comprende 8.000-16.000 transcritos diferentes. En el caso de que se use ADNc-AFLP +0/+1, suponiendo que se usan dos endonucleasas de restricción que reconocen una secuencia de 4 nucleótidos, que se dirigen aproximadamente al 80 % del número total de transcritos, la mezcla de complejidad reducida comprenderá aproximadamente 1.600-3.200

transcritos. Con una secuenciación 20 veces redundante, esto equivale a de 32.000 a 64.000 lecturas necesarias por muestra. Esto es suficiente para poder determinar también los niveles de transcritos de genes que se expresan a niveles relativamente bajos.

5 Un procedimiento muy adecuado para determinar la complejidad de la colección de secuencias se describe en el documento WO 03/010328.

La invención también se refiere a un procedimiento para determinar los niveles de transcripción relativos de una secuencia de nucleótidos en muestras de ADNc que comprende las etapas de:

- a) Determinar la frecuencia de una secuencia de nucleótidos en una primera muestra de ADNc realizando un procedimiento como se define en la reivindicación 2 sobre dicha primera muestra de ADNc;
- 10 b) Determinar la frecuencia de la misma secuencia de nucleótidos en una segunda muestra y/o en una muestra posterior de ADNc realizando un procedimiento como se define en la reivindicación 2 sobre dicha segunda muestra y/o una muestra posterior de ADNc; y
- c) Comparar la frecuencia de la secuencia de nucleótidos en dicha primera muestra de ADNc con la frecuencia de la misma secuencia de nucleótidos en dicha segunda muestra y/o una muestra posterior de ADNc para  
15 obtener niveles de transcripción relativos de la secuencia de nucleótidos.

En la etapa (a) del procedimiento, se determina la frecuencia de una secuencia de nucleótidos en una primera muestra de ADNc realizando un procedimiento como se define en la reivindicación 2 sobre dicha primera muestra de ADNc.

20 En la etapa (b) del procedimiento, se determina la frecuencia de la misma secuencia de nucleótidos en una segunda muestra y/o una muestra posterior de ADNc realizando un procedimiento como se definen en la reivindicación 2 sobre dicha segunda muestra y/o una muestra posterior de ADNc.

En la etapa (c), se compara la frecuencia de la secuencia de nucleótidos en dicha primera muestra de ADNc con la frecuencia de la misma secuencia de nucleótidos en dicha segunda muestra y/o una muestra posterior de ADNc para obtener niveles de transcripción relativos de la secuencia de nucleótidos.

25 Conocer dichos niveles de transcripción relativos puede ser importante para establecer los transcritos importantes para determinados fenotipos, como se analiza anteriormente.

La invención también se refiere a un procedimiento para determinar niveles de transcripción relativos de una secuencia de nucleótidos en muestras de ADNc que comprende las etapas de:

- a) Proporcionar una primera muestra de ADNc;
- 30 b) Realizar una reducción de la complejidad sobre la primera muestra de ADNc para obtener una primera colección;
- c) Marcar la primera colección para obtener una primera colección marcada;
- d) Realizar, de forma consecutiva o simultáneamente, las etapas (a) y (b) con una segunda muestra y/o una muestra posterior de ADNc, preferentemente usando una marca diferente para cada muestra de ADNc, para  
35 obtener una segunda colección marcada y/o una colección marcada posterior;
- e) Combinar la primera colección marcada y la segunda colección marcada y/o una colección marcada posterior para obtener una colección combinada;
- f) Determinar al menos parte de las secuencias de nucleótidos de la colección marcada mediante secuenciación;
- g) Determinar la frecuencia de la secuencia de nucleótidos en la primera muestra de ADNc y la segunda  
40 muestra y/o una muestra posterior de ADN; y
- h) Comparar la frecuencia de la secuencia de nucleótidos en la primera muestra de ADNc con la frecuencia de la secuencia de nucleótidos en la segunda muestra y/o una muestra posterior de ADNc para obtener niveles de transcripción relativos de la secuencia de nucleótidos en las muestras de ADNc.

45 En la etapa (a), se proporciona una primera muestra de ADNc. Una muestra de ADNc se puede obtener como se analiza anteriormente.

En la etapa (b), se realiza una reducción de la complejidad sobre la primera muestra de ADNc para obtener una primera colección. La reducción de la complejidad se puede realizar mediante cualquier técnica, pero preferentemente se realiza por medio de la técnica AFLP<sup>®</sup> de Keygene.

En la etapa (c), se marca la primera colección para obtener una primera colección marcada. El marcaje puede

producirse simultáneamente con la etapa (b) de reducción de la complejidad. Este marcaje simultáneo se puede lograr, p. ej., por AFLP, usando adaptadores que comprenden un único identificador (nucleótido) para cada muestra.

El marcaje está destinado a distinguir entre muestras de origen diferente, p. ej., obtenidas a partir de diferentes líneas de plantas, cuando se combinan dos o más colecciones de complejidad reducida de dos o más muestras de ADNc para obtener una colección combinada. Por tanto, se usan preferentemente diferentes marcas para preparar las colecciones marcadas de la primera muestra de ADNc y la segunda muestra o una muestra posterior de ADNc. Cuando se usan, por ejemplo, cinco muestras de ácido nucleico, se pretende obtener cinco colecciones marcadas diferentes, denotando las cinco marcas diferentes las correspondientes muestras originales.

La marca puede ser cualquier marca conocida en la técnica para distinguir muestras de ácido nucleico, pero preferentemente es una secuencia identificadora corta. Una secuencia identificadora de este tipo puede ser, p. ej., una secuencia de bases única de longitud variable usada para indicar el origen de la colección obtenida mediante reducción de la complejidad. La incorporación de una marca de oligonucleótido en un adaptador o cebador es muy conveniente, ya que no se necesitan etapas adicionales para marcar una colección. Una secuencia identificadora de este tipo puede ser de longitud variable, en función del número de muestras de ácido nucleico que se van a comparar. Una longitud de aproximadamente 4 bases ( $4^4 = 256$  secuencias marcadoras diferentes posibles) es suficiente para distinguir entre el origen de un número limitado de muestras (hasta 256), aunque se prefiere que las secuencias marcadoras difieran en más de una base entre las muestras que se van a distinguir. En caso necesario, se puede ajustar la longitud de las secuencias marcadoras en consecuencia.

En la etapa (d), se realizan, de forma consecutiva o simultáneamente, las etapas (a) y (b) con una segunda muestras o una muestra posterior de ADNc, preferentemente usando una marca diferente para cada muestra de ADNc, para obtener una segunda colección marcada o una colección marcada posterior. Las muestras de ADNc pueden ser, p. ej., de origen diferente, p. ej., diferentes líneas de plantas, de forma que se pueden comparar tales perfiles de transcritos de tales líneas de plantas. De forma alternativa, se pueden derivar muestras de ADNc, p. ej., de una sola línea de plantas en distintas etapas del desarrollo con el fin de comparar perfiles de transcritos durante el desarrollo de plantas. También es posible realizar el procedimiento de acuerdo con la presente invención sobre muestras de ADNc completamente independientes simplemente por su eficacia.

En la etapa (e), se combinan la primera colección marcada y la segunda colección marcada y/o una colección marcada posterior para obtener una colección combinada. Esta colección combinada se puede someter a secuenciación simultánea para proporcionar un procedimiento altamente eficaz.

En la etapa (f), se determinan al menos parte de las secuencias de nucleótidos de la colección combinada por secuenciación, preferentemente, secuenciación de alto rendimiento, preferentemente como se describe anteriormente.

En la etapa (g), se determina la frecuencia de la secuencia de nucleótidos en la primera muestra de ADNc y la segunda muestra y/o una muestra posterior de ADN. Las secuencias de nucleótidos de la primera colección se pueden distinguir de las secuencias de nucleótidos de la segunda colección y/o una colección posterior por medio de la marca. En este caso, la hibridación se puede realizar sobre datos de secuencia que se han recortado para los adaptadores/el cebador y/o los identificadores, pero con secuencias de reconocimiento de enzimas de restricción reconstruidas, es decir, usando sólo los datos de secuencia de los fragmentos que derivan del ADNc. Normalmente, los datos de secuencia obtenidos se usan para identificar el origen del fragmento (es decir, de qué muestra), las secuencias derivadas de la secuencia adaptadora y/o identificadora se eliminan de los datos y se realiza la hibridación en este conjunto recortado.

En la etapa (h), se compara la frecuencia de la secuencia de nucleótidos en la primera muestra de ADNc con la frecuencia de la misma secuencia de nucleótidos en la segunda muestra y/o una muestra posterior de ADNc para obtener niveles de transcripción relativos de la secuencia de nucleótidos en las muestras de ADNc.

Debido a la estrategia de marcaje, se puede realizar simultáneamente la determinación de los niveles de transcripción de una secuencia de nucleótidos para diferentes muestras de ADNc, lo que es muy ventajoso. El procedimiento es muy adecuado para la identificación rápida de transcritos implicados en un rasgo fenotípico determinado, como se analiza anteriormente.

En una realización preferida, el marcaje de la primera colección y de la segunda colección o una colección posterior se realiza usando diferentes marcas. Como se analiza anteriormente, se prefiere que cada colección de una muestra de ADNc se identifique mediante su propia marca.

#### Breve descripción de los dibujos

**Figura 1:** Productos de ADNc-AFLP marcados (A/C) de las líneas de pimiento PSP11 y PI 201234. Se cargan dos muestras de ambas líneas por duplicado en un gel de agarosa al 1 %.

M: marcador de 100 pb

- 1: muestra 1 de ADNc-AFLP de PSP11
- 2: muestra 1 de ADNc-AFLP de PSP11
- 3: muestra 2 de ADNc-AFLP de PSP11
- 4: muestra 2 de ADNc-AFLP de PSP11
- 5: ADNc-AFLP de PI 201234 - muestra 1
- 6: ADNc-AFLP de PI 201234 - muestra 1
- 7: ADNc-AFLP de PI 201234 - muestra 2
- 8: ADNc-AFLP de PI 201234 - muestra 2

**Figura 2:** Representación esquemática de productos de amplificación de AFLP +1/+1 de pimiento después de la amplificación con cebadores de AFLP que contienen secuencias marcadoras de 4 pb en 5' prima.

**Figura 3:** Flujo de trabajo de la preparación de la colección de secuencias.

**Figura 4:** Ejemplo de resultado de 13 lecturas de secuencia.

**Figura 5:** Resultados de Blast

**Figura 6:** Presentación de datos no procesados de una regulación por incremento.

**Figura 7:** Presentación de datos no procesados de una regulación por incremento.

### Ejemplos

Se han acumulado un gran número de ejemplos de regulación temporal y espacial de expresiones génicas en plantas superiores usando enfoque tales como la hibridación de bandas northern o aplicaciones de expresión de micromatrices de ADN. La última tecnología permite el control de la expresión de miles de genes simultáneamente. A diferencia de estos procedimientos de análisis, el análisis digital de los perfiles de expresión génica se puede lograr secuenciando directamente transcritos marcados usando tecnologías de secuencia de alto rendimiento. El número de secuencias obtenidas a partir de un transcrito específico en una muestra refleja el nivel de transcripción de esta secuencia en particular. La comparación de estos números entre varias muestras, aunque supone profundizar en la secuenciación, permite la medida precisa de los niveles de transcripción entre estas muestras. Esta tecnología parece ser una potente herramienta para descubrir nuevos marcadores de calidad desconocidos que están relacionados con determinados perfiles de expresión.

Aquí se describe la secuenciación de alto rendimiento de ADNc, cuya complejidad se ha reducido usando la tecnología de AFLP, derivado de la fracción de ARNm de dos líneas de pimiento. Mediante secuenciación directa de fragmentos de ADNc marcados, se pudieron generar perfiles de expresión.

### Procedimientos

#### ***Aislamiento de ARNtot/Poli(A)+ de ARN***

Se aisló ARN total de las líneas de pimiento PSP11 y PI 201234 a partir de material de hoja usando el minikit RNeasy del Rneasy Plant Mini Protocol de QIAGEN (n.º de cat: 74104'). Como entrada se han usado aprox. 100 mg de material de hoja por muestra.

Siguiendo este protocolo se obtuvieron rendimientos de 2,5-3 µg de RNA total por muestra. Posteriormente, se aisló la fracción de ARN poli(A)<sup>+</sup> a partir de 1 µg de las muestras de ARN total usando el minikit de ARNm Oligotex de QIAGEN (n.º de cat: 70022). Se obtuvieron rendimientos de 150-200 ng de ARN poli(A)<sup>+</sup>. Las concentraciones de estas muestras fueron de 5 - 10 ng/ul. Tanto el ARN total como el ARN tpoli(A)<sup>+</sup> se analizaron en un gel de agarosa para comprobar la calidad del ARN.

#### ***Síntesis de ADNc***

Se generó ADNc de acuerdo con el protocolo siguiente:

##### *Síntesis de la primera hebra del ADNc*

Añadir conjuntamente:

10 µl de ARN poli(A)<sup>+</sup> (50 - 100 ng)

5 µl de oligo-dT25 (70 ng/ul)

Posteriormente, añadir:

5 ul de tampón de primera hebra 5x (suministrado con Superscript II RT)

2,5 µl de DTT 0,1 M

1 ul de dNTP 10 mM

5 0,5 ul de Superscript II (200 U/ul)

1 ul de agua MQ hasta un volumen final de 25 ul

Incubar 2 horas a 42 °C

Síntesis de la segunda hebra del ADNc

Añadir conjuntamente:

10 25 ul de mezcla de reacción de primera hebra

8 ul de tampón de segunda hebra 10 x

1,5 ul de dNTP 10 mM

7,5 unidades de ligasa de ADN de *E. coli*

25 unidades de polimerasa de *E. coli*

15 0,8 unidades de RNasa-H (1 U/ul)

Añadir agua MQ hasta un volumen final de 80 ul

Incubar 1 hora a 12 °C

Incubar 1 hora a 22 °C

20 Posteriormente, se purificaron las muestras de ADNc usando el kit de purificación de membrana de PCR Qiaquick de QIAGEN (n.º de cat: 28104). La elución se llevó a cabo usando 30 µl de tampón de elución (Tris-HCl 5 mM, pH 8,5).

**Preparación de moldes de ADNc - AFLP usando cebadores de AFLP marcados**

Se prepararon moldes de AFLP de los ADNc generados de las líneas progenitoras de pimiento PSP11 y PI-201234 usando la combinación de endonucleasas de restricción *TaqI* / *Mse I* como se describe por Zabeau & Vos, 1993: Selective restriction fragment amplification; a general method for DNA fingerprinting. documento EP 0534858-A1, B1; patente de EE. UU. 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 DNA fingerprinting. Nucl. Acids Res., 21, 4407-4414).

Procedimiento de restricción y ligación de ADNc

30 La digestión se realizó en dos etapas; primero con la *TaqI* (mayor temperatura de incubación), después con *MseI* (temperatura de incubación más baja)

La restricción de ADNc con *TaqI* y *MseI* se llevó a cabo como sigue:

Restricción del ADN

Añadir conjuntamente:

250 ng de ADNc

35 10 unidades de *TaqI*

8 µl de tampón RL 5x. El tampón RL 5x es Tris-HAc 50 mM, MgAc 50 mM, KAc 250 mM, DTT 25 mM, 250 ng/µl de BSA; pH 7,5).

Añadir agua MQ hasta un volumen final de 40 µl

Incubar 2 horas a 65 °C

40 Después de la restricción con *TaqI*,



Añadir:

10 unidades de *Mse*I

2 µl de tampón RL 5x

Añadir agua MQ hasta un volumen final de 50 µl

5 Incubar 2 horas a 37 °C

Ligación de adaptadores

A la mezcla de digestión se le añadieron los componentes siguientes:

1 µl de ATP 10mM

1 µl de ligasa de ADN T4

10 1 µl de adaptador de *Taq*I (50 pmol/µl)

CTCGTAGACTGCGTAC/CGGTACGCAGTCT

1 µl de adaptador de *Mse*I (50 pmol/µl)

GACGATGAGTCCTGAG/TACTCAGGAACTCAT

2 µl de tampón RL 5x.

15 Añadir agua MQ hasta un volumen final de 60 µl

Incubar 3 horas a 37 °C

Amplificación de ADNc - AFLP

Después de la restricción-ligación, se usó este producto de restricción/ligación como molde en una etapa de amplificación no selectiva. Estos productos de AFLP no selectivos se usaron posteriormente como molde para amplificación selectiva (+1/+1). Se realizó un control de calidad sobre este producto +1/+1 realizando una amplificación selectiva +2/+3. Los productos de la última amplificación se comprobaron en un gel de secuencia al 4,5 %.

La amplificación no selectiva de ADNc-AFLP se realizó como sigue:

5 µl de mezcla de restricción-ligación no diluida

25 1,5 µl de cebador de *Taq*I (50ng/µl) (CTCGTAGACTGCGTACCGA)

1,5 µl de cebador de *Mse*I (50ng/µl) (GATGAGTCCTGAGTAA)

2 µl de dNTP 5 mM

1 unidad de polimerasa *Taq*

5 µl de tampón de PCR 10x

30 Añadir agua MQ hasta un volumen final de 50 µl

Las amplificaciones de PCR se realizaron usando un PE9700 con un bloque de oro o plata usando las condiciones siguientes: 30 ciclos (30" a 94 °C, 60" a 56 °C y 1 20" a 72 °C).

La amplificación selectiva de ADNc-AFLP usando secuencias marcadas se realizó como sigue:

Para producto de ADNc-AFLP no selectivo derivado para la línea de pimiento PSP11

35 5 µl de producto no selectivo diluido 600 x

1,5 µl de cebador ACAC Tr01 (+A)\* (50 ng/µg) (ACACGTAGACTGCGTACCGAA)

1,5 µl de cebador ACAC M02 (+C)\* (50 ng/µg) (ACACGATGAGTCCTGAGTAAC)

2 µl de dNTP 5 mM

1,5 unidades de polimerasa AmpliTaq-Gold

5 ul de tampón de PCR 10x

Añadir agua MQ hasta un volumen final de 50 µl

Para producto de ADNc-AFLP no selectivo 0/0 derivado para la línea de pimiento PI 201234.

5 ul de producto no selectivo diluido 600 x

5 1,5 ul de cebador AGCT Tr01 (+A)\* (50 ng/µg) (AGCTGTAGACTGCGTACCGAA)

1,5 ul de cebador AGCT M02 (+C)\* (50 ng/µg) (AGCTGATGAGTCCTGAGTAAC)

2 ul de dNTP 5 mM

1,5 unidades de polimerasa AmpliTaq-Gold

5 ul de tampón de PCR 10x

10 Añadir agua MQ hasta un volumen final de 50 ul

Las amplificaciones de PCR se realizaron usando un PE9700 con un bloque de oro usando las condiciones siguientes: 1 ciclo de 12' a 94 °C (comienzo en caliente), 30" a 94 °C, 30" a 65 °C, 60" a 72 °C; 2 3 ciclos - disminución de la temperatura de hibridación 0,7 °C en cada ciclo durante 12 ciclos - fase de temperatura decreciente de 13 ciclos - 30" a 94 °C, 30" a 56 °C, 60" a 72 °C. La calidad de los productos +1/+1 generados se controló en un gel de agarosa al 1 % usando una escalera de 100 pares de bases para comprobar la distribución de longitud de los fragmentos (véase la figura 1).

15 Los cebadores selectivos contienen marcas de 4 pb (subrayadas anteriormente) en sus extremos 5' para distinguir los productos de amplificación que se originan a partir de las correspondientes líneas de pimiento al final del procedimiento de secuenciación. El principio de generar productos de PCR de ADNc - AFLP marcados de acuerdo con este procedimiento se muestra en la figura 2

#### Preparación de colecciones de secuencias y secuenciación de alto rendimiento

Los productos de ADNc AFLP marcados de ambas líneas de pimiento se sometieron a secuenciación de alto rendimiento usando tecnología de secuenciación de 454 Life Sciences / Roche GS20 como se describe por Margulies et al., (Margulies et al., Nature 437, pág. 376-380 y suplementos en línea). En primer lugar, se purificaron los productos de PCR de ADNc - AFLP marcados y se ligaron a un adaptador modificado (CCATCTCATCCCTGCG TGTCCCATCTGTTCCTCCCTGTCTCAGT/CTGAGACAGGGAG-GGAACAGATGG y BIO-TEG-CCTATCCCCTGTGTGCCTTGCCTATCCCCTGTTGCGTGTCTCAGT/P-CTGA-GACACG CAACAGGGGATAGGCAAGGCACACAGGGGATAGG) para facilitar la amplificación por PCR de la emulsión y la posterior secuenciación de fragmentos como se describe por Margulies y sus colaboradores. Los cebadores de PCR de emulsión, los cebadores de secuencia y las condiciones de desarrollo de la secuencia fueron como se describe por Margulies y sus colaboradores. El procedimiento de preparación de la colección de secuencias se muestra en la figura 3. Se realizó un desarrollo de secuencias de alto rendimiento GS20 en los laboratorios de Keygene NV, Wageningen, Países Bajos.

#### Procesamiento de datos de desarrollo de secuencias GS20.

35 Los datos de secuencias resultantes de la mitad del desarrollo de secuencias GS20 (es decir, 1 canal de los 2 canales disponibles en la placa PicoTiter GS20) se procesaron usando una fuente de información bioinformática (Keygene N.V.). Específicamente, se convirtieron lecturas de secuencias automáticas no procesadas en formato FASTA y se inspeccionaron para detectar la presencia de secuencias adaptadoras de AFLP usando un algoritmo BLAST. Tras los apareamientos altamente fiables con las secuencias de cebador de AFLP marcado conocido, se recortaron las secuencias, se restablecieron los sitios de endonucleasa de restricción y se asignaron las marcas apropiadas. Posteriormente, todas las secuencias recortadas mayores de 33 bases se agruparon usando un procedimiento de megaBLAST basado en homologías de secuencia globales. A continuación, se ensamblaron las agrupaciones en uno o más contigs por agrupación, usando un algoritmo de hibridación múltiple CAP3.

#### *Ejemplo de resultado de 13 lecturas de secuencia:*

#### 45 Agrupación 387

Las ID de marcas de la muestra 2 (AGTC) se representan en **NEGRITA**. Las ID de marcas de la muestra 1 (ACAC) están subrayadas. Véase la fig. 4.

En la tabla 1 se muestran estadísticas globales de los desarrollos de las secuencias:

**Tabla 1:** Estadísticas globales del desarrollo de ADNc-AFLP.

Fragmentos de secuencia con muestra identificada	174421
Lecturas de la muestra 1 (PSP11)	50599
Lecturas de la muestra 2 (PI 201234)	123822
proporción de muestras (muestra 2/muestra 1)	2,45
agrupaciones	6712
Agrupaciones presentes tanto en la muestra 1 como en la muestra 2	1433

Interpretación:

Etapa 1) El "factor de normalización de la profundidad de la secuenciación de la muestra" es 2,45 y se define como el total de lecturas obtenidas a partir de la muestra 2 dividido entre el número total de lecturas derivadas de la muestra 1 ( $123822 / 50599 = 2,45$ ). El número de lecturas derivadas de la muestra 2 por cóntigo se dividió entre 2,45 con el fin de comparar los niveles de transcripción con los de la muestra 1.

Etapa 2) Se realizó una segunda etapa de "normalización de genes de mantenimiento" mediante la determinación de la "expresión" de un gen "de mantenimiento" que funciona como estándar interno. Para esto, se seleccionó el gen de la arginina descarboxilasa de *Lycopersicon esculentum*. La secuencia de la arginina descarboxilasa de *Lycopersicon esculentum* se "BLASTEÓ" contra las secuencias de los cóntigos obtenidos usando la hibridación múltiple CAP3 para determinar con qué frecuencia se observaban los transcritos del gen de la arginina descarboxilasa del pimiento en las muestras 1 y 2. Posteriormente, se calculó la proporción en la que se observaban estos transcritos en las muestras 1 y 2, después de aplicar en primer lugar el "factor de normalización de la profundidad de la secuenciación de la muestra" (etapa 1). En este ejemplo, esta proporción (= factor de normalización del gen de mantenimiento) fue de  $17/14 = 1,2$  para muestra 1 / muestra 2 (tabla 1).

*Ejemplo de búsqueda BLAST de un gen de mantenimiento (arginina descarboxilasa de Lycopersicon esculentum) contra el grupo de cóntigos.*

Referencia: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller y David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

Consulta = gi|295349|gb|L16582.1|TOMARGDECA ARNm de arginina descarboxilasa de *Lycopersicon esculentum*, cds completo (2060 letras)

Base de datos: taggedReads.fna

174.421 secuencias; 15.408.192 letras en total. Los resultados están en la fig. 5.

**Tabla 2.** Cálculo del factor de normalización 1,2 de un gen de mantenimiento (muestra 1 / muestra 2) basado en la abundancia del homólogo de pimiento del gen de la arginina descarboxilasa de tomate.

Gen de mantenimiento estándar	Lecturas en el cóntigo de la muestra 1	Lecturas en el cóntigo de la muestra 2 Antes/después de la normalización de la profundidad de la secuenciación de la muestra	Proporción Muestra 1 / Muestra 2 (facto de normalización del gen de mantenimiento)
gi 295349 gb L16582.1 TOMARGDECA ARNm de arginina descarboxilasa de <i>Lycopersicon esculentum</i>	17	35/14	1.2 (17/14)

Etapa 3) Para la elaboración de perfiles de expresión reales sólo se tuvieron en cuenta cóntigos que contienen más de 10 lecturas. El nivel mínimo de 10 lecturas por cóntigo se escogió con el fin de evitar resultados de elaboración de perfiles de transcritos poco precisos debido a una profundidad de secuenciación insuficiente. La tabla 2 muestra los niveles de expresión relativos de ARNm de dos transcritos que se expresan de forma diferencial en PSP11 (muestra 1) frente a PI 201234 (muestra 2), siguiendo el procedimiento en tres etapas explicado anteriormente. Específicamente, la agrupación 2215 representa un transcrito regulado por incremento en la muestra 1 y el agrupamiento 847 representa un transcrito regulado por disminución en la muestra 1; en la tabla 3 se muestran

cálculos de los niveles de transcripción relativos de estos transcritos. Finalmente, la tabla 4 contiene un resumen del número de genes transcritos de forma diferencial en la totalidad del conjunto de datos basado en los principios descritos anteriormente.

5 *Ejemplo de regulación por incremento de la muestra 1 - datos no procesados. Agrupación 2215. Las ID de marcas de la muestra 2 (AGTC) se representan en **NEGRITA**. Las ID de marcas de la muestra 1 (ACAC) están subrayadas en la fig. 6.*

*Ejemplo de regulación por disminución de la muestra 1 - datos no procesados. Agrupación 847. Las ID de marcas de la muestra 2 (AGTC) se representan en **NEGRITA**. Las ID de marcas de la muestra 1 (ACAC) están subrayadas en la fig. 7.*

10 **Tabla 3.** Cálculo de niveles de expresión relativos de transcritos representados por las agrupaciones 2215 y 847, siguiendo la normalización de la profundidad de la secuenciación de la muestra (etapa 1) y la normalización del gen de mantenimiento (etapa).

N.º de agrupación:	2215	847
Lecturas de la muestra 1 - datos no procesados	44	11
Lecturas de la muestra 2 - datos no procesados	26	101
Lecturas de la muestra 1 - normalización de la profundidad de la secuenciación de la muestra	44	11
Lecturas de la muestra 2 - normalización de la profundidad de la secuenciación de la muestra	10,6 (26/2,45)	41,2 (101/2,45)
Lecturas de la muestra 1 - normalización del gen de mantenimiento	37 (44/1,2)	9 (11/1,2)
Lecturas de la muestra 2 - normalización del gen de mantenimiento	10,6	41,2
Proporción de expresión de la muestra 1 frente a la muestra 2	3,5 (37/10,6)	0,2 (9/41,2)

15 **Tabla 4:** Resumen de los niveles de transcripción relativos de transcritos secuenciados a partir de PSP11 y/o PI 201234 y presentes en cóntigos que contienen 10 o más secuencias.

	N.º mínimo de lecturas de ambas muestras
	lecturas > 10
Número total de cóntigos que contienen lecturas de la muestra 1 y/o la muestra 2	113
Genes regulados por disminución (proporción del nivel de expresión <0,5)	20
Genes regulados por incremento (proporción del nivel de expresión >2)	17
Genes expresados por igual (proporción de nivel de expresión > 0,5 y < 2)	76

## REIVINDICACIONES

1. Procedimiento para determinar niveles de transcripción relativos de una secuencia de nucleótidos en muestras de ADNc que comprende las etapas de:

(a) Proporcionar una primera muestra de ADNc;

5 (b) Realizar una reducción de la complejidad reproducible de la primera muestra de ADNc para obtener una primera colección que comprende las etapas de:

- digerir el ADNc con al menos una endonucleasa de restricción para fragmentarlo en fragmentos de restricción;

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

15 - poner en contacto dichos fragmentos de restricción ligados a adaptador con uno o más cebadores oligonucleotídicos bajo condiciones de hibridación, teniendo dichos uno o más cebadores oligonucleotídicos una secuencia de cebador que incluye una sección de secuencia de nucleótidos complementaria a parte del al menos un adaptador y a parte de la parte restante de la secuencia de reconocimiento de la endonucleasa de restricción; y

- amplificar dichos fragmentos de restricción ligados a adaptador por elongación del uno o más cebadores oligonucleotídicos hibridados;

20 (c) Marcar la primera colección para obtener una primera colección marcada mediante el uso de al menos un adaptador marcado en la etapa (b);

(d) Realizar, de forma consecutiva o simultáneamente, las etapas (a) y (b) con una segunda muestra y/o una muestra posterior de ADNc, usando una marca diferente para cada muestra de ADNc, para obtener una segunda colección marcada y/o una colección marcada posterior;

25 (e) Combinar la primera colección marcada y la segunda colección marcada y/o una colección marcada posterior para obtener una colección combinada;

(f) Determinar al menos parte de las secuencias de nucleótidos de la colección marcada mediante secuenciación de alto rendimiento;

(g) Determinar la frecuencia de la secuencia de nucleótidos en la primera muestra de ADNc y la segunda muestra y/o una muestra posterior de ADN; y

30 (h) Comparar la frecuencia de la secuencia de nucleótidos en la primera muestra de ADNc con la frecuencia de la secuencia de nucleótidos en la segunda muestra y/o una muestra posterior de ADNc para obtener niveles de transcripción relativos de la secuencia de nucleótidos en las muestras de ADNc.

2. Procedimiento de acuerdo con cualquiera de las reivindicaciones anteriores, en el que la secuenciación de alto rendimiento se realiza sobre un soporte sólido tal como una perla.

35 3. Procedimiento de acuerdo con cualquiera de las reivindicaciones 1 o 2, en el que la secuenciación de alto rendimiento se basa en la secuenciación por síntesis, preferentemente en la pirosecuenciación.

4. Procedimiento de acuerdo con cualquiera de las reivindicaciones 1 o 2, en el que la secuenciación de alto rendimiento comprende las etapas de:

(c1) ligar adaptadores de secuenciación a los fragmentos;

40 (c2) hibridar fragmentos ligados a adaptadores de secuenciación con perlas, alineándose cada perla con un solo fragmento;

(c3) emulsionar las perlas en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una sola perla;

45 (c4) realizar una PCR de la emulsión para amplificar fragmentos ligados a adaptadores de secuencia sobre la superficie de las perlas;

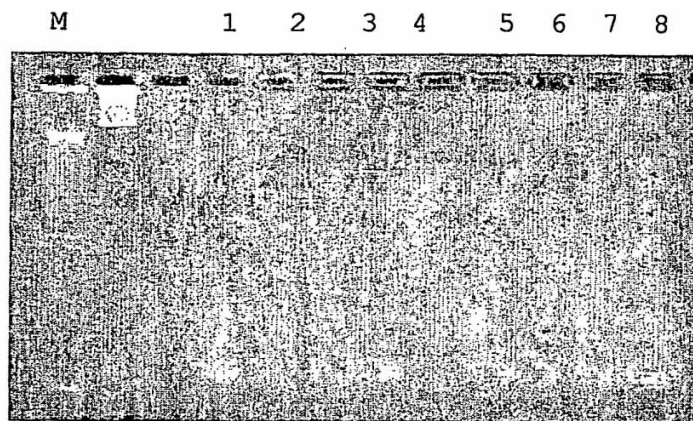
(c5) seleccionar / enriquecer perlas que contienen fragmentos ligados a adaptadores de secuenciación amplificados;

(c6) cargar las perlas en pocillos, comprendiendo cada pocillo una sola perla; y

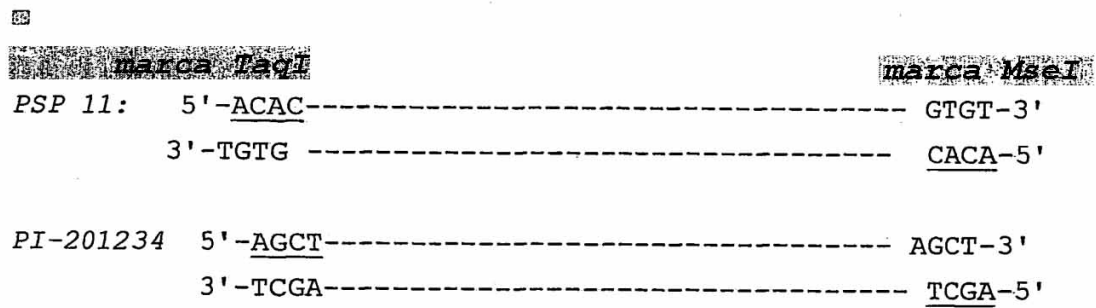
(c7) generar una señal de pirofosfato.

5. Procedimiento de acuerdo con la reivindicación 1, en el que el cebador comprende además una secuencia seleccionada en el extremo 3' de la secuencia del cebador, comprendiendo dicha secuencia seleccionada 1-10 nucleótidos selectivos que son complementarios a una sección situada inmediatamente adyacente a la parte restante de la secuencia de reconocimiento de la endonucleasa de restricción.
- 5
6. Procedimiento de acuerdo con cualquiera de las reivindicaciones 1 o 5, en el que la secuencia seleccionada del extremo 3' de la secuencia del cebador comprende 1-8 nucleótidos selectivos, preferentemente 1-5, más preferentemente 1-3.

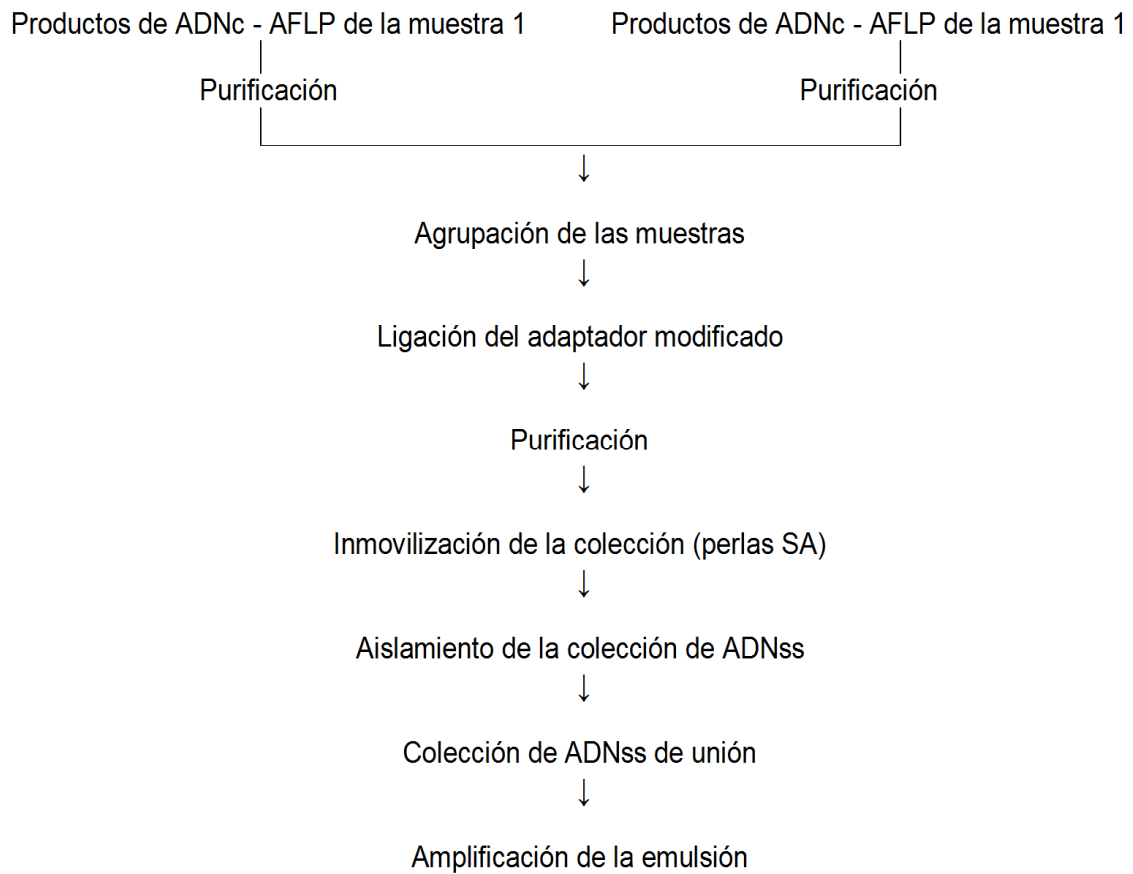
**Fig 1**



**Fig 2**



**Fig 3**





**Fig 4**

```

1.>SMPL-SM_AGTC|TR-4913003|RW-4733936|CF-M02AGCT >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCCCTTTTGTCAAAAAAGGGG

2.>SMPL-SM_AGTC|TR-4932007|RW-4752942|CF-M02AGCT >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGG

3.>SMPL-SM_ACAC|TR-4953247|RW-4774200|CF-M02ACAC >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGGAA

4.>SMPL-SM_AGTC|TR-4966783|RW-4787732|CF-M02AGCT >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGG

5.>SMPL-SM_AGTC|TR-4967968|RW-4788917|CF-M02AGCT >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGG

6.>SMPL-SM_AGTC|TR-4969557|RW-4790520|CF-M02AGCT >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGG

7.>SMPL-SM_AGTC|TR-4970551|RW-4791514|CF-M02AGCT >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACAAAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGG

8.>SMPL-SM_AGTC|TR-4971473|RW-4792435|CF-M02AGCT >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGNAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGG

9.>SMPL-SM_ACAC|TR-4974503|RW-4795464|CF-M02ACAC >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGGAAG

10.>SMPL-SM_ACAC|TR-5053800|RW-4874753|CF-M02ACAC >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGGGAAAG

11.>SMPL-SM_ACAC|TR-5056946|RW-4877898|CF-M02ACAC >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGGGAAAG

12.>SMPL-SM_AGTC|TR-5057240|RW-4878192|CF-M02AGCT >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGG

13>SMPL-SM_ACAC|TR-5057897|RW-4878848|CF-M02ACAC >-< NINGUNO
TTAAACAAGAAGTACGAGACTCTTTTCATACCTTCCTGATTTGTCCGACGAGCAATTGCTCAAAGGAAAAATTGAGTAACCTTTTGTCAAAAAAGGGTAG

```

```

Consulta: 771 830
          cstatagcattgaagaatatgcctctgctgttggtccaagcggtcctctatgtctgtgatcg
          |||| ||||||| |||||||||||||||| || |||| | ||||||||| ||
          cstatggcattgaggaatatgcctctgctgttggtccaggctgtccaatttgtctgtgaccg
          72 13
          Obj:

Consulta: 831 taaggcg 838
          |||||||
          Obj: 12 taaggcg 5

```

Fig 6

Agrupamiento 2215

Lectura - longitud de sec.		ID de la muestra	homología de sec.
0	86nt,	>SMPL-SM_ACAC TR-4898365 RW-4719274 CF-M02ACAC	>-< NINGUNO... *
1	86nt,	>SMPL-SM_ACAC TR-4898518 RW-4719427 CF-M02ACAC	>-< NINGUNO...a +/96%
2	86nt,	>SMPL-SM_ACAC TR-4898650 RW-4719559 CF-M02ACAC	>-< NINGUNO...a +/96%
3	85nt,	>SMPL-SM_ACAC TR-4900482 RW-4721391 CF-M02ACAC	>-< NINGUNO...a +/95%
4	85nt,	>SMPL-SM_AGTC TR-4903275 RW-4724184 CF-M02AGCT	>-< NINGUNO...a +/97%
5	86nt,	>SMPL-SM_ACAC TR-4907918 RW-4728846 CF-M02ACAC	>-< NINGUNO...a +/96%
6	86nt,	>SMPL-SM_ACAC TR-4907931 RW-4728859 CF-M02ACAC	>-< NINGUNO...a +/96%
7	86nt,	>SMPL-SM_AGTC TR-4909659 RW-4730592 CF-M02AGCT	>-< NINGUNO...a +/95%
8	86nt,	>SMPL-SM_ACAC TR-4910385 RW-4731318 CF-M02ACAC	>-< NINGUNO...a +/96%
9	85nt,	>SMPL-SM_AGTC TR-4913022 RW-4733955 CF-M02AGCT	>-< NINGUNO...a +/97%
10	86nt,	>SMPL-SM_ACAC TR-4920018 RW-4740951 CF-M02ACAC	>-< NINGUNO...a +/96%
11	86nt,	>SMPL-SM_AGTC TR-4921404 RW-4742337 CF-M02AGCT	>-< NINGUNO...a +/97%
12	85nt,	>SMPL-SM_AGTC TR-4939338 RW-4760272 CF-M02AGCT	>-< NINGUNO...a +/97%
13	86nt,	>SMPL-SM_ACAC TR-4940227 RW-4761161 CF-M02ACAC	>-< NINGUNO...a +/96%
14	86nt,	>SMPL-SM_AGTC TR-4940529 RW-4761463 CF-M02AGCT	>-< NINGUNO...a +/96%
15	86nt,	>SMPL-SM_ACAC TR-4942223 RW-4763157 CF-M02ACAC	>-< NINGUNO...a +/96%
16	86nt,	>SMPL-SM_ACAC TR-4944416 RW-4765349 CF-M02ACAC	>-< NINGUNO...a +/96%
17	85nt,	>SMPL-SM_ACAC TR-4944867 RW-4765800 CF-M02ACAC	>-< NINGUNO...a +/95%
18	86nt,	>SMPL-SM_ACAC TR-4950174 RW-4771127 CF-M02ACAC	>-< NINGUNO...a +/96%
19	85nt,	>SMPL-SM_ACAC TR-4951211 RW-4772164 CF-M02ACAC	>-< NINGUNO...a +/95%

**Fig 6 Cont**

20	86nt,	>SMPL-SM_ <u>AGTC</u>  TR-4951664 RW-477261 CF-M02AGCT >-< NINGUNO...a +/95%
21	85nt,	>SMPL-SM_ <u>AGTC</u>  TR-4952233 RW-4773186 CF-M02AGCT >-< NINGUNO...a +/97%
22	86nt,	>SMPL-SM_ <u>ACAC</u>  TR-4954586 RW-4775539 CF-M02ACAC >-< NINGUNO...a +/96%
23	86nt,	>SMPL-SM_ <u>ACAC</u>  TR-4955938 RW-4776891 CF-M02ACAC >-< NINGUNO...a +/96%
24	86nt,	>SMPL-SM_ <u>ACAC</u>  TR-4956477 RW-4777430 CF-M02ACAC >-< NINGUNO...a +/96%
25	85nt,	>SMPL-SM_ <u>AGTC</u>  TR-4957577 RW-4778530 CF-M02AGCT >-< NINGUNO...a +/97%
26	86nt,	>SMPL-SM_ <u>ACAC</u>  TR-4963864 RW-4784814 CF-M02ACAC >-< NINGUNO...a +/96%
27	86nt,	>SMPL-SM_ <u>ACAC</u>  TR-4965869 RW-4786818 CF-M02ACAC >-< NINGUNO...a +/96%
28	86nt,	>SMPL-SM_ <u>ACAC</u>  TR-4966323 RW-4787272 CF-M02ACAC >-< NINGUNO...a +/96%
29	86nt,	>SMPL-SM_ <u>ACAC</u>  TR-4973719 RW-4794680 CF-M02ACAC >-< NINGUNO...a +/95%
30	85nt,	>SMPL-SM_ <u>AGTC</u>  TR-4974751 RW-4795712 CF-M02AGCT >-< NINGUNO...a +/97%
31	85nt,	>SMPL-SM_ <u>ACAC</u>  TR-4974769 RW-4795730 CF-M02ACAC >-< NINGUNO...a +/95%
32	85nt,	>SMPL-SM_ <u>ACAC</u>  TR-4975495 RW-4796456 CF-M02ACAC >-< NINGUNO...a +/95%
33	86nt,	>SMPL-SM_ <u>AGTC</u>  TR-4976407 RW-4797368 CF-M02AGCT >-< NINGUNO...a +/95%
34	85nt,	>SMPL-SM_ <u>ACAC</u>  TR-4979742 RW-4800702 CF-M02ACAC >-< NINGUNO...a +/95%
35	85nt,	>SMPL-SM_ <u>ACAC</u>  TR-4984717 RW-4805676 CF-M02ACAC >-< NINGUNO...a +/95%
36	86nt,	>SMPL-SM_ <u>AGTC</u>  TR-4990480 RW-4811451 CF-M02AGCT >-< NINGUNO...a +/95%
37	86nt,	>SMPL-SM_ <u>ACAC</u>  TR-4992560 RW-4813531 CF-M02ACAC >-< NINGUNO...a +/95%
38	85nt,	>SMPL-SM_ <u>ACAC</u>  TR-4993469 RW-4814440 CF-M02ACAC >-< NINGUNO...a +/95%
39	85nt,	>SMPL-SM_ <u>ACAC</u>  TR-4996160 RW-4817131 CF-M02ACAC >-< NINGUNO...a +/95%
40	86nt,	>SMPL-SM_ <u>AGTC</u>  TR-5000664 RW-4821635 CF-M02AGCT >-< NINGUNO...a +/96%
41	85nt,	>SMPL-SM_ <u>ACAC</u>  TR-5003833 RW-4824801 CF-M02ACAC >-< NINGUNO...a +/95%

**Fig 6 Cont**

42	85nt,	>SMPL-SM_ACAC TR-5005452 RW-4826419 CF-M02ACAC>-<NINGUNO...a+/95%
43	86nt,	>SMPL-SM_AGTC TR-5006454 RW-4827421 CF-M02AGCT>-<NINGUNO...a+/95%
44	67nt,	>SMPL-SM_AGTC TR-5008627 RW-4829593 CF-M02AGCT>-<NINGUNO...a+/97%
45	85nt,	>SMPL-SM_ACAC TR-5012256 RW-4833222 CF-M02ACAC>-<NINGUNO...a+/95%
46	85nt,	>SMPL-SM_ACAC TR-5018696 RW-4839659 CF-M02ACAC>-<NINGUNO...a+/95%
47	85nt,	>SMPL-SM_ACAC TR-5019736 RW-4840697 CF-M02ACAC>-<NINGUNO...a+/95%
48	86nt,	>SMPL-SM_ACAC TR-5022201 RW-4843162 CF-M02ACAC>-<NINGUNO...a+/95%
49	86nt,	>SMPL-SM_AGTC TR-5022852 RW-4843813 CF-M02AGCT>-<NINGUNO...a+/95%
50	85nt,	>SMPL-SM_ACAC TR-5023548 RW-4844508 CF-M02ACAC>-<NINGUNO...a+/95%
51	86nt,	>SMPL-SM_AGTC TR-5024453 RW-4845413 CF-M02AGCT>-<NINGUNO...a+/95%
52	86nt,	>SMPL-SM_AGTC TR-5026412 RW-4847372 CF-M02AGCT>-<NINGUNO...a+/95%
53	85nt,	>SMPL-SM_ACAC TR-5027133 RW-4848093 CF-M02ACAC>-<NINGUNO...a+/95%
54	85nt,	>SMPL-SM_ACAC TR-5030214 RW-4851174 CF-M02ACAC>-<NINGUNO...a+/95%
55	86nt,	>SMPL-SM_AGTC TR-5033174 RW-4854133 CF-M02AGCT>-<NINGUNO...a+/95%
56	86nt,	>SMPL-SM_AGTC TR-5036554 RW-4857512 CF-M02AGCT>-<NINGUNO...a+/95%
57	86nt,	>SMPL-SM_AGTC TR-5037426 RW-4858384 CF-M02AGCT>-<NINGUNO...a+/95%
58	85nt,	>SMPL-SM_AGTC TR-5038670 RW-4859628 CF-M02AGCT>-<NINGUNO...a+/96%
59	85nt,	>SMPL-SM_ACAC TR-5043111 RW-4864066 CF-M02ACAC>-<NINGUNO...a+/95%
60	67nt,	>SMPL-SM_AGTC TR-5043543 RW-4864498 CF-M02AGCT>-<NINGUNO...a+/97%
61	85nt,	>SMPL-SM_ACAC TR-5045177 RW-4866131 CF-M02ACAC>-<NINGUNO...a+/95%
62	64nt,	>SMPL-SM_ACAC TR-5045870 RW-4866824 CF-M02ACAC>-<NINGUNO...a+/96%
63	85nt,	>SMPL-SM_ACAC TR-5047816 RW-4868770 CF-M02ACAC>-<NINGUNO...a+/95%

**Fig 6 Cont**

64	86nt,	>SMPL-SM_ <u>AGTC</u>  TR-5047871 RW-4868825 CF-M02AGCT	>--< NINGUNO...a+/95%
65	60nt,	>SMPL-SM_ <u>AGTC</u>  TR-5051602 RW-4872555 CF-M02AGCT	>--< NINGUNO...a+/96%
66	86nt,	>SMPL-SM_ <u>AGTC</u>  TR-5052084 RW-4873037 CF-M02AGCT	>--< NINGUNO...a+/95%
67	85nt,	>SMPL-SM_ <u>ACAC</u>  TR-5053649 RW-4874602 CF-M02ACAC	>--< NINGUNO...a+/95%
68	84nt,	>SMPL-SM_ <u>ACAC</u>  TR-5057753 RW-4878704 CF-M02ACAC	>--< NINGUNO...a+/95%
69	85nt,	>SMPL-SM_ <u>ACAC</u>  TR-5059668 RW-4880618 CF-M02ACAC	>--< NINGUNO...a+/95%

**Fig 7***Agrupamiento 847*

<u>Lectura - longitud de sec.</u>	<u>ID de la muestra</u>	<u>homología de sec.</u>
0 94nt,	>SMPL-SM_AGTC TR-4883337 RW-4704221 CF-M02AGCT	>-<NINGUNO...a +/100%
1 90nt,	>SMPL-SM_ACAC TR-4883909 RW-4704793 CF-M02ACAC	>-<NINGUNO...a +/100%
2 94nt,	>SMPL-SM_AGTC TR-4886846 RW-4707756 CF-M02AGCT	>-<NINGUNO...a +/100%
3 94nt,	>SMPL-SM_AGTC TR-4888343 RW-4709253 CF-M02AGCT	>-<NINGUNO...a +/100%
4 94nt,	>SMPL-SM_AGTC TR-4888697 RW-4709607 CF-M02AGCT	>-<NINGUNO...a +/100%
5 94nt,	>SMPL-SM_AGTC TR-4892791 RW-4713700 CF-M02AGCT	>-<NINGUNO...a +/100%
6 94nt,	>SMPL-SM_AGTC TR-4894880 RW-4715789 CF-M02AGCT	>-<NINGUNO...a +/100%
7 87nt,	>SMPL-SM_AGTC TR-4897114 RW-4718023 CF-M02AGCT	>-<NINGUNO...a +/100%
8 90nt,	>SMPL-SM_ACAC TR-4897371 RW-4718280 CF-M02ACAC	>-<NINGUNO...a +/100%
9 94nt,	>SMPL-SM_AGTC TR-4901953 RW-4722862 CF-M02AGCT	>-<NINGUNO...a +/100%
10 94nt,	>SMPL-SM_AGTC TR-4903361 RW-4724270 CF-M02AGCT	>-<NINGUNO...a +/100%
11 94nt,	>SMPL-SM_AGTC TR-4904140 RW-4725049 CF-M02AGCT	>-<NINGUNO...a +/98%
12 94nt,	>SMPL-SM_AGTC TR-4911001 RW-4731934 CF-M02AGCT	>-<NINGUNO...a +/100%
13 90nt,	>SMPL-SM_ACAC TR-4913392 RW-4734325 CF-M02ACAC	>-<NINGUNO...a +/100%
14 87nt,	>SMPL-SM_AGTC TR-4915105 RW-4736038 CF-M02AGCT	>-<NINGUNO...a +/100%
15 94nt,	>SMPL-SM_AGTC TR-4915956 RW-4736889 CF-M02AGCT	>-<NINGUNO...a +/100%
16 90nt,	>SMPL-SM_ACAC TR-4916571 RW-4737504 CF-M02ACAC	>-<NINGUNO...a +/100%
17 94nt,	>SMPL-SM_AGTC TR-4917457 RW-4738390 CF-M02AGCT	>-<NINGUNO...a +/100%

**Fig 7 Cont**

```

18 94nt, >SMPL-SM_AGTC|TR-4923368|RW-4744301|CF-M02AGCT >-< NINGUNO...a +/100%
19 94nt, >SMPL-SM_AGTC|TR-4924721|RW-4745654|CF-M02AGCT >-< NINGUNO...a +/100%
20 94nt, >SMPL-SM_AGTC|TR-4925652|RW-4746585|CF-M02AGCT >-< NINGUNO...a +/100%
21 94nt, >SMPL-SM_AGTC|TR-4925964|RW-4746897|CF-M02AGCT >-< NINGUNO...a +/100%
22 94nt, >SMPL-SM_AGTC|TR-4927490|RW-4748423|CF-M02AGCT >-< NINGUNO...a +/100%
23 90nt, >SMPL-SM_ACAC|TR-4929253|RW-4750188|CF-M02ACAC >-< NINGUNO...a +/100%
24 93nt, >SMPL-SM_AGTC|TR-4930011|RW-4750946|CF-M02AGCT >-< NINGUNO...a +/100%
25 94nt, >SMPL-SM_AGTC|TR-4930912|RW-4751847|CF-M02AGCT >-< NINGUNO...a +/100%
26 95nt, >SMPL-SM_AGTC|TR-4930919|RW-4751854|CF-M02AGCT >-< NINGUNO...a +/98%
27 95nt, >SMPL-SM_AGTC|TR-4932026|RW-4752961|CF-M02AGCT >-< NINGUNO...a +/98%
28 94nt, >SMPL-SM_AGTC|TR-4933349|RW-4754284|CF-M02AGCT >-< NINGUNO...a +/100%
29 94nt, >SMPL-SM_AGTC|TR-4939421|RW-4760355|CF-M02AGCT >-< NINGUNO...a +/100%
30 94nt, >SMPL-SM_AGTC|TR-4939767|RW-4760701|CF-M02AGCT >-< NINGUNO...a +/100%
31 94nt, >SMPL-SM_AGTC|TR-4940033|RW-4760967|CF-M02AGCT >-< NINGUNO...a +/100%
32 94nt, >SMPL-SM_AGTC|TR-4941722|RW-4762656|CF-M02AGCT >-< NINGUNO...a +/98%
33 93nt, >SMPL-SM_AGTC|TR-4942036|RW-4762970|CF-M02AGCT >-< NINGUNO...a +/100%
34 94nt, >SMPL-SM_AGTC|TR-4944792|RW-4765725|CF-M02AGCT >-< NINGUNO...a +/100%
35 94nt, >SMPL-SM_AGTC|TR-4947138|RW-4768071|CF-M02AGCT >-< NINGUNO...a +/100%
36 94nt, >SMPL-SM_AGTC|TR-4951642|RW-4772595|CF-M02AGCT >-< NINGUNO...a +/100%
37 94nt, >SMPL-SM_AGTC|TR-4951751|RW-4772704|CF-M02AGCT >-< NINGUNO...a +/100%
38 94nt, >SMPL-SM_AGTC|TR-4953697|RW-4774650|CF-M02AGCT >-< NINGUNO...a +/100%

```



**Fig 7 Cont**

```

39 95nt, >SMPL-SM_AGTC|TR-4956808|RW-4777761|CF-M02AGCT >--< NINGUNO...a +/98%
40 94nt, >SMPL-SM_AGTC|TR-4957145|RW-4778098|CF-M02AGCT >--< NINGUNO...a +/100%
41 94nt, >SMPL-SM_AGTC|TR-4959060|RW-4780013|CF-M02AGCT >--< NINGUNO...a +/100%
42 94nt, >SMPL-SM_AGTC|TR-4960682|RW-4781634|CF-M02AGCT >--< NINGUNO...a +/100%
43 94nt, >SMPL-SM_AGTC|TR-4960814|RW-4781766|CF-M02AGCT >--< NINGUNO...a +/100%
44 94nt, >SMPL-SM_AGTC|TR-4963272|RW-4784222|CF-M02AGCT >--< NINGUNO...a +/100%
45 95nt, >SMPL-SM_AGTC|TR-4963923|RW-4784873|CF-M02AGCT >--< NINGUNO...a +/98%
46 94nt, >SMPL-SM_AGTC|TR-4964337|RW-4785287|CF-M02AGCT >--< NINGUNO...a +/100%
47 94nt, >SMPL-SM_AGTC|TR-4964661|RW-4785610|CF-M02AGCT >--< NINGUNO...a +/100%
48 94nt, >SMPL-SM_AGTC|TR-4964984|RW-4785933|CF-M02AGCT >--< NINGUNO...a +/100%
49 93nt, >SMPL-SM_AGTC|TR-4965578|RW-4786527|CF-M02AGCT >--< NINGUNO...a +/100%
50 94nt, >SMPL-SM_AGTC|TR-4967242|RW-4788191|CF-M02AGCT >--< NINGUNO...a +/98%
51 94nt, >SMPL-SM_AGTC|TR-4967410|RW-4788359|CF-M02AGCT >--< NINGUNO...a +/100%
52 93nt, >SMPL-SM_AGTC|TR-4967805|RW-4788754|CF-M02AGCT >--< NINGUNO...a +/100%
53 94nt, >SMPL-SM_AGTC|TR-4968426|RW-4789381|CF-M02AGCT >--< NINGUNO...a +/100%
54 94nt, >SMPL-SM_AGTC|TR-4969295|RW-4790258|CF-M02AGCT >--< NINGUNO...a +/100%
55 94nt, >SMPL-SM_AGTC|TR-4971157|RW-4792120|CF-M02AGCT >--< NINGUNO...a +/100%
56 93nt, >SMPL-SM_AGTC|TR-4973058|RW-4794019|CF-M02AGCT >--< NINGUNO...a +/100%
57 90nt, >SMPL-SM_ACAC|TR-4974092|RW-4795053|CF-M02ACAC >--< NINGUNO...a +/100%
58 94nt, >SMPL-SM_AGTC|TR-4979083|RW-4800043|CF-M02AGCT >--< NINGUNO...a +/100%
59 94nt, >SMPL-SM_AGTC|TR-4981582|RW-4802542|CF-M02AGCT >--< NINGUNO...a +/100%

```

**Fig 7 Cont**

```

60 94nt, >SMPL-SM_AGTC|TR-4982384|RW-4803344|CF-M02AGCT >-<NINGUNO...a +/100%
61 94nt, >SMPL-SM_AGTC|TR-4983869|RW-4804828|CF-M02AGCT >-<NINGUNO...a +/100%
62 94nt, >SMPL-SM_AGTC|TR-4984183|RW-4805142|CF-M02AGCT >-<NINGUNO...a +/100%
63 94nt, >SMPL-SM_AGTC|TR-4985610|RW-4806569|CF-M02AGCT >-<NINGUNO...a +/100%
64 94nt, >SMPL-SM_AGTC|TR-4992392|RW-4813363|CF-M02AGCT >-<NINGUNO...a +/100%
65 94nt, >SMPL-SM_AGTC|TR-4995878|RW-4816849|CF-M02AGCT >-<NINGUNO...a +/100%
66 94nt, >SMPL-SM_AGTC|TR-4997121|RW-4818092|CF-M02AGCT >-<NINGUNO...a +/100%
67 94nt, >SMPL-SM_AGTC|TR-4999529|RW-4820500|CF-M02AGCT >-<NINGUNO...a +/100%
68 94nt, >SMPL-SM_AGTC|TR-4999642|RW-4820613|CF-M02AGCT >-<NINGUNO...a +/100%
69 94nt, >SMPL-SM_AGTC|TR-5000307|RW-4821278|CF-M02AGCT >-<NINGUNO...a +/100%
70 94nt, >SMPL-SM_AGTC|TR-5004064|RW-4825032|CF-M02AGCT >-<NINGUNO...a +/100%
71 94nt, >SMPL-SM_AGTC|TR-5004113|RW-4825081|CF-M02AGCT >-<NINGUNO...a +/100%
72 94nt, >SMPL-SM_AGTC|TR-5005004|RW-4825972|CF-M02AGCT >-<NINGUNO...a +/100%
73 94nt, >SMPL-SM_AGTC|TR-5006293|RW-4827260|CF-M02AGCT >-<NINGUNO...a +/98%
74 94nt, >SMPL-SM_AGTC|TR-5006419|RW-4827386|CF-M02AGCT >-<NINGUNO...a +/100%
75 94nt, >SMPL-SM_AGTC|TR-5006904|RW-4827871|CF-M02AGCT >-<NINGUNO...a +/100%
76 90nt, >SMPL-SM_ACAC|TR-5010125|RW-4831091|CF-M02ACAC >-<NINGUNO...a +/100%
77 94nt, >SMPL-SM_AGTC|TR-5010746|RW-4831712|CF-M02AGCT >-<NINGUNO...a +/100%
78 76nt, >SMPL-SM_ACAC|TR-5011064|RW-4832030|CF-M02ACAC >-< Tr01ACA... at +/97%
79 94nt, >SMPL-SM_AGTC|TR-5013823|RW-4834788|CF-M02AGCT >-<NINGUNO...a +/98%
80 94nt, >SMPL-SM_AGTC|TR-5013955|RW-4834920|CF-M02AGCT >-<NINGUNO...a +/100%

```

**Fig 7 Cont**

81	94nt,	>SMPL-SM_AGTC TR-5018602 RW-4839565 CF-M02AGCT	>-<NINGUNO...a	+ /100%
82	94nt,	>SMPL-SM_AGTC TR-5023027 RW-4843988 CF-M02AGCT	>-<NINGUNO...a	+ /100%
83	94nt,	>SMPL-SM_AGTC TR-5024523 RW-4845483 CF-M02AGCT	>-<NINGUNO...a	+ /100%
84	94nt,	>SMPL-SM_AGTC TR-5028089 RW-4849049 CF-M02AGCT	>-<NINGUNO...a	+ /100%
85	94nt,	>SMPL-SM_AGTC TR-5028414 RW-4849374 CF-M02AGCT	>-<NINGUNO...a	+ /100%
86	94nt,	>SMPL-SM_AGTC TR-5029518 RW-4850478 CF-M02AGCT	>-<NINGUNO...a	+ /100%
87	94nt,	>SMPL-SM_AGTC TR-5030111 RW-4851071 CF-M02AGCT	>-<NINGUNO...a	+ /100%
88	94nt,	>SMPL-SM_AGTC TR-5031539 RW-4852499 CF-M02AGCT	>-<NINGUNO...a	+ /100%
89	94nt,	>SMPL-SM_AGTC TR-5033696 RW-4854655 CF-M02AGCT	>-<NINGUNO...a	+ /100%
90	94nt,	>SMPL-SM_AGTC TR-5033962 RW-4854921 CF-M02AGCT	>-<NINGUNO...a	+ /100%
91	94nt,	>SMPL-SM_AGTC TR-5036809 RW-4857767 CF-M02AGCT	>-<NINGUNO...a	+ /100%
92	94nt,	>SMPL-SM_AGTC TR-5041108 RW-4862064 CF-M02AGCT	>-<NINGUNO...a	+ /100%
93	94nt,	>SMPL-SM_AGTC TR-5044884 RW-4865838 CF-M02AGCT	>-<NINGUNO...a	+ /100%
94	94nt,	>SMPL-SM_AGTC TR-5047474 RW-4868428 CF-M02AGCT	>-<NINGUNO...a	+ /100%
95	94nt,	>SMPL-SM_AGTC TR-5049520 RW-4870474 CF-M02AGCT	>-<NINGUNO...a	+ /100%
96	94nt,	>SMPL-SM_AGTC TR-5050268 RW-4871222 CF-M02AGCT	>-<NINGUNO...a	+ /100%
97	93nt,	>SMPL-SM_AGTC TR-5050874 RW-4871828 CF-M02AGCT	>-<NINGUNO...a	+ /100%
98	94nt,	>SMPL-SM_AGTC TR-5051012 RW-4871966 CF-M02AGCT	>-<NINGUNO...a	+ /100%
99	94nt,	>SMPL-SM_AGTC TR-5052845 RW-4873798 CF-M02AGCT	>-<NINGUNO...a	+ /100%
100	90nt,	>SMPL-SM_ACAC TR-5053070 RW-4874023 CF-M02ACAC	>-<NINGUNO...a	+ /100%
101	93nt,	>SMPL-SM_AGTC TR-5053369 RW-4874322 CF-M02AGCT	>-<NINGUNO...a	+ /100%

**Fig 7 Cont**

```

102 93nt, >SMPL-SM_AGTC|TR-5053925|RW-4874878|CF-M02AGCT >-<NINGUNO...a +/100%
103 96nt, >SMPL-SM_AGTC|TR-5054104|RW-4875057|CF-M02AGCT >-<NIGNUNO...*
104 94nt, >SMPL-SM_AGTC|TR-5054107|RW-4875060|CF-M02AGCT >-<NINGUNO...a +/100%
105 95nt, >SMPL-SM_ACAC|TR-5054781|RW-4875734|CF-M02ACAC >-<NINGUNO...a +/97%
106 96nt, >SMPL-SM_AGTC|TR-5055060|RW-4876013|CF-M02AGCT >-<NINGUNO...a +/97%
107 94nt, >SMPL-SM_AGTC|TR-5055349|RW-4876302|CF-M02AGCT >-<NINGUNO...a +/100%
108 96nt, >SMPL-SM_ACAC|TR-5056743|RW-4877695|CF-M02ACAC >-<NINGUNO...a +/97%
109 96nt, >SMPL-SM_AGTC|TR-5057451|RW-4878403|CF-M02AGCT >-<NINGUNO...a +/97%
110 94nt, >SMPL-SM_AGTC|TR-5059058|RW-4880008|CF-M02AGCT >-<NINGUNO...a +/100%
111 94nt, >SMPL-SM_AGTC|TR-5060243|RW-4881193|CF-M02AGCT >-<NINGUNO...a +/98%

```