

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 714 851**

51 Int. Cl.:

G06F 19/18 (2011.01)

C12Q 1/6827 (2008.01)

C12Q 1/6874 (2008.01)

G06F 19/22 (2011.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **03.11.2016 PCT/CA2016/051277**

87 Fecha y número de publicación internacional: **11.05.2017 WO17075706**

96 Fecha de presentación y número de la solicitud europea: **03.11.2016 E 16836042 (8)**

97 Fecha y número de publicación de la concesión europea: **09.01.2019 EP 3180723**

54 Título: **Procedimiento de alto rendimiento de cribado de una población para la obtención de elementos que comprendan mutaciones en una secuencia objetivo, utilizando un análisis de secuencia sin alineación**

30 Prioridad:

04.11.2015 CA 2911002

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

30.05.2019

73 Titular/es:

**VINELAND RESEARCH AND INNOVATION
CENTRE (100.0%)
4890 Victoria Avenue North
Vineland Station, ON L0R 2E0 , CA**

72 Inventor/es:

**BANKS, TRAVIS WILFRED y
SOMERS, DARYL JOHN**

74 Agente/Representante:

ARPE FERNÁNDEZ, Manuel

ES 2 714 851 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

Procedimiento de alto rendimiento de cribado de una población para la obtención de elementos que comprendan mutaciones en una secuencia objetivo, utilizando un análisis de secuencia sin alineación

CAMPO DE LA INVENCION

[0001] La presente invención se refiere al campo de la genética y la biología molecular. Concretamente, la presente invención se refiere a procedimientos de inversión genética de alto rendimiento para el cribado de elementos de una población que comprende una o más mutaciones en una o más secuencias objetivo, utilizando un análisis de secuencia sin alineación. Asimismo, la invención también facilita kits para su utilización con los procedimientos.

ANTECEDENTES DE LA INVENCION

[0002] La industria agrícola global se enfrenta a muchos retos y presiones que resultan especialmente evidentes en lo tocante a la producción de organismos sésiles: tensiones bióticas y abióticas que amenazan el rendimiento y la calidad de las cosechas; el incremento en los costes de mano de obra, agua y energía; y otras limitaciones adicionales impuestas por las preferencias de los consumidores. De este modo existe una gran demanda de producción de cultivos que sean tolerantes a las condiciones extremas, que requieran pocas o ningunas aportaciones (es decir, un menor uso de agua, fertilizantes y/o pesticidas), y que también resulten atractivos para los consumidores. Las posibilidades de desarrollo de rasgos utilizando sistemas de cultivo tradicionales van paulatinamente siendo más limitadas, debido a la ausencia de diversidad genética en variedades de plantas cultivadas. Es posible la introgresión de características valiosas procedentes de accesiones silvestres, pero este procedimiento podría no ser factible si la característica que interesa se encuentra estrechamente vinculada con aquellas asociadas a características no deseadas (Fitzpatrick et al, Plant Cell. 24:395-414). Puede adoptarse un enfoque transgénico, pero los organismos genéticamente modificados, y especialmente aquellos con los que se obtienen productos comestibles, resultan controvertidos y presentan unos desafíos completamente nuevos en relación con la normativa en materia de seguridad alimentaria y aceptación por parte de los consumidores. La mutagénesis constituye un procedimiento eficaz y eficiente de introducir diversidad genética en plantas cultivadas (Wang et al, Plant Biotechnology Journal 10:761-772).

[0003] La mutagénesis de plantas para la mejora del cultivo constituye una técnica que lleva utilizándose desde hace casi un siglo. Uno de los primeros informes relativos a mutagénesis de plantas dirigidas por humanos fue elaborado por Lewis Stadler quien, en la década de 1920, irradió cebada y maíz con rayos X y luz ultravioleta (Stadler 1932). La mutagénesis química se introdujo como una técnica alternativa que recurría a reactivos alquilantes, como el sulfonato de etilmetano (EMS), que altera el ADN e induce cambios hereditarios en el genoma (Koorneef et al. 1982). Se han conseguido numerosos éxitos utilizando la mutagénesis para el desarrollo de nuevas características y nuevas variedades de cultivos. El Organismo Internacional de Energía Atómica (IAEA), que es un organismo dependiente de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), con sede en Viena, Austria, proporciona una base de datos de variedades registradas obtenidas a partir de mutagénesis (<http://mvgs.iaea.org/>). Un ejemplo popular de variedad mutante la constituye el pomelo de la variedad Ruby Red, presentada en 1970 en Texas, con la que se utilizó la radiación para mejorar el tono rojo y la conservación del color del pomelo. Asimismo, la cebada Golden Promise se obtuvo mediante una mutación que causaba semi-enanismo, un mayor rendimiento y tolerancia a la sal (Forster 2001). Ambas variedades se han utilizado ampliamente en la producción de las variedades de pomelo y cebada de nuestros días.

[0004] En recientes décadas se han desarrollado alternativas a la mutagénesis para inducir cambios en el genoma. Entre estos procedimientos se encuentra la utilización de constructos artificiales, frecuentemente híbridos, de ADN/ARN con otros aminoácidos o ácidos nucleicos (Sargent et al, Oligonucleotides 21(2):55-75, 2011, KeyBase@ by KeyGene), para inducir cambios genéticos en una o varias ubicaciones específicas. Las moléculas híbridas se transfieren a la célula mediante biolística o fusión de membranas, desde donde a continuación migran al núcleo para propiciar un cambio en el genoma. A continuación, se induce a las células afectadas a crecer en plantas completas. Más recientemente se han desarrollado mutagénesis específicas del emplazamiento mediante la utilización de sistemas enzimáticos de diseño. Las nucleasas de dedos de Zinc, las nucleasas efectoras de tipo activador de transcripción y las meganucleasas son todas ellas unas enzimas diseñadas específicamente y que tienen la capacidad de cortar el ADN genómico en el punto deseado. Estas enzimas se expresan en la célula o células de una planta obtenida a partir de una construcción transgénica introducida en el genoma o a partir de un fragmento extra-genómico presente en el interior de la célula. Recientemente se han transferido a células vegetales como proteínas, eliminando la necesidad de introducir ADN ajeno en la célula de la planta. Todas estas enzimas son capaces de cortar el ADN de doble cadena del genoma. De este modo se activa el mecanismo de reparación del ADN de la célula, que permite la incorporación de un nuevo fragmento de ADN mediante recombinación homóloga, o la planta cierra la ruptura, introduciendo algún tipo de cambio, ya sea una inserción, borrado o sustitución a través del proceso de unión de extremos no homólogos. Más recientemente se ha desarrollado un nuevo tipo de nucleasa artificial basada en el mecanismo de inmunidad bacteriana, denominado sistema CRISPR/Cas. La principal ventaja del sistema CRISPR/Cas consiste en la utilización de una molécula de ARN para dotar al complejo de su especificidad de edición del genoma. Esto contrasta con el resto de las anteriores tecnologías de nucleasa en las que la especificidad del genoma se consiguió mediante el diseño de ADN a medida

para la unión de dominios proteicos. En el caso del sistema CRISPR/Cas, la selección del genoma se efectúa mediante el emparejamiento complementario de una molécula de guía de ARN a una región específica, por lo que la personalización de una nueva nucleasa resulta tan sencillo como el diseño de una nueva secuencia de guía de ARN. Al igual que en el caso de las anteriores nucleasas, los constructos que expresan este sistema se sitúan en una célula en la que se expresa y modifica el genoma.

[0005] Se recurre a un cribado genético para identificar y seleccionar aquellos individuos que poseen un fenotipo que reviste interés, en una población mutagenizada o modificada. Los tipos de cribado genético se dividen en cribados genéticos directos y cribados genéticos inversos. Un cribado genético directo identifica en primer lugar a los individuos de una población mutante con un fenotipo de interés específico. Una vez encontrado el fenotipo, se lleva a cabo el trabajo de seguimiento para identificar el cambio genético subyacente que dio lugar a la creación del nuevo rasgo. Por ejemplo, en Nordstrom et al. (Nat. Biotech. 31(4):325-331; 2013) se utilizan k-meros (fragmentos cortos y solapados de ácidos nucleicos) en un cribado genético directo para identificar las mutaciones mediante comparación directa de datos de secuenciación del genoma completo procedentes de individuos mutantes y naturales. Al contrario de un cribado genético directo, un cribado genético inverso analiza el fenotipo de un organismo tras la disrupción de un gen conocido.

[0006] Hasta hace poco, la mutagénesis de los cultivos vegetales se utilizaba exclusivamente en los cribados genéticos directos, en los que se fenotipaban grandes poblaciones mutagenizadas en búsqueda de nuevos rasgos para la mejora del cultivo. Este procedimiento requiere mucho trabajo, tiende a identificar falsos positivos y está limitado por lo que puede medirse visualmente o a través de procedimientos de elevado rendimiento.

[0007] Debido al trabajo asociado a los cribados genéticos directos, la mutagénesis y la crianza basada en mutaciones perdieron popularidad hasta que se tuvieron en cuenta las expectativas aplicación de la biología molecular.

[0008] La Localización en Genomas de Lesiones Locales Inducidas (TILLING) es una técnica de genética inversa mediante la cual se combinan la mutagénesis aleatoria de alta densidad de poblaciones y las técnicas de detección SNP para la identificación de plantas con mutaciones valiosas (McCallum et al. 2000). Se ha recurrido a la desnaturalización de ADN a alta resolución (HRM) en las técnicas TILLING para la detección de mutaciones en poblaciones tratadas con EMS (Gady et al, Plant Methods 5:13), aunque este procedimiento exige mucho trabajo y resulta muy caro.

[0009] La detección SNP en plantas ha avanzado con mucha rapidez desde el año 2000, y en la actualidad es muy habitual utilizar secuenciaciones de nueva generación (NGS) para descubrir polimorfismos SNP en poblaciones de plantas (Missirian et al. 2011, Rigola et al. 2009, Tsai et al., 2011).

[0010] La secuenciación de ADN de nueva generación (NGS) constituye una herramienta muy atractiva para la identificación de mutaciones en poblaciones de individuos. El rápido descenso de sus precios, el rendimiento cada vez más elevado y la completa caracterización del ADN de las secuencias objetivo han llevado a los investigadores a considerar las posibilidades de la NGS como herramienta para el TILLING (Rigola et al, PLoS One 4:e4761; Tsai et al, Plant Physiology 156:1257-1268). No obstante, debido a la tasa de error intrínseca de las tecnologías NGS resulta difícil discernir entre las mutaciones y los errores de secuenciación en grupos de millares de individuos. La tecnología Illumina de secuenciación produce un error de identificación de bases casi en dos ocasiones por cada 1000 bases secuenciadas (Minoche et al, Genome Biology 12:R112). En un esfuerzo por diferenciar entre errores y mutaciones, los investigadores han creado estrategias de agrupación multidimensionales combinadas con generación de códigos de barras de ADN para la secuenciación de miembros de una población en reacciones múltiples e independientes. Seguidamente se determinan los individuos que albergan una mutación mediante decircunvolución de grupos utilizando códigos de barras (Rigola et al, PLoS One 4:e4761; Missirian et al, BMC Bioinformatics 12:287; WO2007037678, concedida a KeyGene N.V.). También se han desarrollado estrategias que utilizan lecturas compuestas a fin de reducir la tasa de error (WO2014/134729).

[0011] Esta información de contexto se facilita a fin de dar a conocer la información que, según cree el solicitante, puede tener cierta relevancia para la presente invención. No debe admitirse ni interpretarse que la información que antecede constituya la técnica anterior a la presente invención.

RESUMEN DE LA INVENCION

[0012] Uno de los objetos de la presente invención consiste en facilitar procedimientos de alto rendimiento para el cribado de aquellos de los elementos de una mutación que comprendan mutaciones en una o más secuencias objetivo recurriendo al análisis de secuencias sin alineación. Según un aspecto de la presente invención, se facilita un procedimiento para la identificación de los elementos de una población que comprende una o más mutaciones en una o más secuencias objetivo, donde dicho elemento de dicha población es distinto, y donde dicho procedimiento comprende las etapas de: (a) agrupación del ADN genómico aislado de cada uno de los elementos de la población en una o más dimensiones; (b) amplificación de la secuencia o secuencias objetivo del ADN genómico agrupado; (c) secuenciación de los productos amplificados para la obtención de lecturas de secuencia de los productos amplificados; (d) identificación de la mutación o mutaciones en función de los análisis de k-meros de las lecturas de secuencia de (c); donde dicho análisis de k-meros comprende (i) la descomposición de dichas lecturas de secuencia en k-meros de secuencia objetivo; (ii) determinación del número de casos en los que se dan dichos k-meros de la secuencia objetivo obtenidos a partir de dichas lecturas de secuenciación descompuestas; (iii) determinación de un recuento de elementos naturales y elementos mutantes para los k-meros de la secuencia objetivo obtenidos a partir de dichas lecturas de secuenciación descompuestas utilizando secuencias de referencia de las secuencias objetivo a fin de identificar todos los posibles k-meros que contienen una base específica en una

posición específica de la secuencia de referencia u objetivo, donde para cada base situada en una posición específica de dicha secuencia o secuencias de referencia, dicho recuento de elementos naturales se determina mediante el recuento de los k-meros identificados a partir de dichas secuencias de referencia que incluyen dicha base en dichos k-meros de la secuencia objetivo procedentes de dichas lecturas de secuenciación descompuestas; donde, para cada base situada en una posición específica de dicha(s) secuencia(s) de referencia, dicho recuento de mutantes se determina mediante el cambio de dicha base de dicha secuencia o secuencias de referencia a una base diferente, a fin de obtener una secuencia o secuencias de referencia alteradas, identificando todos los posibles k-meros que contienen la base modificada y contando el número de dichos k-meros que contienen la base cambiada y que se encuentran en dichos k-meros de la secuencia objetivo procedentes de dichas lecturas de secuenciación descompuestas; (iv) identificación de una mutación de dicha base en las secuencias objetivo existentes en dichas lecturas de secuencia mediante comparación de la proporción entre el recuento de mutantes y el recuento de elementos naturales para el resto de las bases de dichas secuencias objetivo procedentes de dichas lecturas de secuenciación descompuestas; donde una proporción entre el recuento de mutantes y el recuento de elementos naturales correspondientes a dicha base que sea significativamente diferente, desde el punto de vista estadístico, de la proporción entre recuentos de mutantes y silvestres para el resto de las bases de dicha secuencia o secuencias objetivo será un indicador de una mutación de dicha base en dicha secuencia o secuencias objetivo; y (e) la identificación de los elementos individuales de la población que comprenden la mutación o mutaciones identificadas en las secuencias objetivo.

[0013] En determinadas realizaciones, la secuenciación es del tipo pareado, e incluye adicionalmente la fusión de las lecturas pareadas en lecturas compuestas. En determinadas realizaciones, la etapa de amplificación comprende la amplificación de más de una muestra del ADN genómico agrupado, y la agrupación de los productos de amplificación procedentes de cada muestra para crear grupos de productos de amplificación. En determinadas realizaciones, la secuenciación la lleva a cabo un tercero. En algunas realizaciones, la población se mutageniza mediante agentes químicos inductores de mutaciones, radiaciones ionizantes, intercambio de nucleótidos objetivo, o mutagénesis orientada a la región. En ciertas realizaciones, la identificación de los elementos de la población que comprenden una o más de las mutaciones identificadas en las secuencias objetivo se lleva a cabo mediante desnaturalización de ADN a alta resolución (HRM). En determinadas realizaciones, el procedimiento también comprende, tras la fase de identificación de los elementos de la población que comprenden una o más de las mutaciones identificadas en las secuencias objetivo, la realización de análisis de fenotipo de dichos elementos individuales que comprenden la mutación o mutaciones identificadas en las secuencias objetivo.

[0014] En determinadas realizaciones, la población consiste en una población de plantas. En una serie de realizaciones específicas, la planta es un cereal, una semilla oleaginosa, una fruta, verdura, biocombustible, una planta ornamental, una planta con flor, una planta anual o una planta perenne. En unas realizaciones más específicas, la planta se selecciona entre el grupo que comprende la petunia, el tomate, la pimienta, lechuga, patata, cebolla, zanahoria, brócoli, apio, guisante, espinaca, impatiens, pepino, rosa, batata, árboles frutales, berenjena quimbombó, maíz, soja, colza, trigo, avena, arroz, sorgo, algodón y cebada.

DESCRIPCIÓN DETALLADA DE LA INVENCION

[0015] En el presente documento se describe un procedimiento de genética inversa para la identificación de una mutación en una o más secuencias objetivo, y en algunas realizaciones, el aislamiento de un elemento de la población que presente mutaciones en una o más secuencias objetivo. El procedimiento utiliza una metodología sin alineación para identificar las mutaciones en una o más secuencias objetivo. En ciertas realizaciones específicas, el procedimiento utiliza fragmentos cortos y solapados (k-meros, también denominados kmeros) para la identificación de los *loci* polimórficos (como las mutaciones).

Secuencia objetivo

[0016] La secuencia objetivo candidata puede ser una secuencia objetivo previamente identificada, o también puede identificarse mediante el análisis de la literatura científica y/o la experimentación. Normalmente, una secuencia objetivo es una región de un gen en la que tendría efecto una mutación. Por ejemplo, un experto en la materia se daría cuenta inmediatamente de que las mutaciones en secuencias de no codificación, como los intrones, pueden tener poco o nulo efecto. Dicho experto también se daría cuenta de que las mutaciones en regiones de codificación conservadas de los genes tienen una mayor posibilidad de tener un efecto. Normalmente, una secuencia objetivo tiene una longitud superior a 1000 bases, a fin de facilitar la fragmentación durante la preparación de la biblioteca de secuenciación. En los casos en los que la secuencia objetivo es mayor que el amplicón de la PCR más grande posible con la polimerasa de ADN seleccionada, se crean múltiples amplicones de la PCR. En ciertas realizaciones, en las que se precisan múltiples amplicones de la PCR, los amplicones de la PCR pueden tener un solapamiento de 200bp o mayor.

[0017] En aquellas realizaciones en las que se examinan múltiples secuencias objetivo, cada una de las secuencias objetivo puede encontrarse en los mismos o diferentes genes. Por ejemplo, en las realizaciones en las que se examinan dos secuencias objetivo, ambas secuencias objetivo pueden encontrarse en el mismo gen, o la primera secuencia objetivo puede estar en el primer gen y la segunda secuencia objetivo puede estar en un segundo gen. De este modo, en ciertas realizaciones, se criban uno o más genes en busca de mutaciones. En determinadas realizaciones, se criban dos o más genes en busca de mutaciones. En determinadas realizaciones, se criban tres o más genes en busca de mutaciones.

Población

[0018] La población a partir de la cual se aísla el ADN genómico puede ser una población no mutagenizada, organismos mutagenizados y su progenie (incluyendo, entre otros, plantas, células o animales como la *Drosophila* o los ratones. Las plantas pueden ser, por ejemplo, un cereal, una semilla oleaginosa, una fruta, verdura, biocombustible, una planta ornamental, una planta con flor, una planta anual o una planta perenne. Entre los ejemplos de plantas se encuentran, entre otros, la petunia, el tomate (*Solanum lycopersicum*), el pimiento (*Capsicum annuum*), lechuga, patata, cebolla, zanahoria, brócoli, apio, guisante, espinaca, *impatiens*, pepino, rosa, batata, manzana y otros árboles frutales (peral, melocotonero, ciruelo, nectarina), berenjena, quimbombó, maíz, soja, colza, trigo, avena, arroz, sorgo, algodón y cebada. En determinadas realizaciones, la población consiste en una variedad de plantas anuales. En ciertas realizaciones específicas, la población consiste en una población de tomates. En otras realizaciones específicas, la población es una población de pimientos.

[0019] Un experto en la materia se daría cuenta fácilmente de que las mutaciones pueden tener lugar de forma espontánea en una población, o que la población podría experimentar la mutagenización a través de medios químicos o físicos. Por ejemplo, un experto se daría cuenta con facilidad de que se puede utilizar como mutágeno el metanosulfonato de etilo (EMS), o que pueden emplearse como mutágenos las radiaciones ionizantes, como los rayos x, los rayos gamma y la radiación por haz de neutrones rápidos. Un experto en la materia se daría cuenta de que la población podría someterse a un intercambio de nucleótidos objetivo o a mutagénesis centrada en regiones. Un experto se daría cuenta también de que los elementos transponibles pueden actuar como mutágenos.

[0020] En determinadas realizaciones de la invención, la población es una población de plantas mutagenizadas mediante EMS o la progenie de las mismas.

[0021] En otras realizaciones, la población es una población de *Solanum lycopersicum* mutagenizada mediante EMS o su progenie.

[0022] En otras realizaciones, la población es una población de *Capsicum annuum* mutagenizada mediante EMS o su progenie.

[0023] En otras realizaciones, la población puede haber sido modificada genéticamente. Un experto en la materia observaría fácilmente las metodologías para la modificación genética de una población. A modo de ejemplo, y entre otros, los procedimientos incluyen procedimientos que utilizan constructos artificiales, procedimientos de mutagénesis específica del emplazamiento, mediante la utilización de sistemas de enzimas de diseño (como las nucleasas de dedos de zinc, las nucleasas efectoras de tipo activador de transcripción, las meganucleasas y las nucleasas basadas en el sistema CRISPR/Cas) o constructos que expresen las enzimas.

Procedimiento:

[0024] La presente invención aporta un procedimiento para la identificación de mutaciones en una o más secuencias objetivo de una población. El procedimiento comprende (a) el agrupamiento del ADN genómico aislado procedente de cada uno de los miembros de la población en una o más dimensiones; (b) amplificación de la secuencia o secuencias objetivo del ADN genómico agrupado; (c) secuenciación de los productos amplificados u obtención de lecturas de las secuencias para los productos amplificados; y (d) identificación de las mutaciones en función de análisis de secuencia sin alineación de los datos de secuenciación. Opcionalmente, los procedimientos también comprenden (e) identificación de los elementos individuales de la población que comprenden la mutación o mutaciones identificadas en las secuencias objetivo. Opcionalmente, los productos de amplificación de la etapa (b) se agrupan con anterioridad a la secuenciación de los productos amplificados. Opcionalmente, en lo que respecta a la etapa (c), la secuenciación es del tipo pareado, e incluye adicionalmente la fusión de las lecturas pareadas en lecturas compuestas.

[0025] En determinadas realizaciones, la invención facilita un procedimiento para el aislamiento de un elemento de una población que presente una o más mutaciones en una o más secuencias objetivo de una población, incluyendo las siguientes etapas:

(a) Agrupamiento del ADN genómico aislado procedente de cada uno de los miembros de la población en una o más dimensiones;

(b) Amplificación de la secuencia o secuencias objetivo en el ADN genómico agrupado;

(c) Secuenciación de los productos amplificados para obtener lecturas de secuencias correspondientes a los productos amplificados;

(d) Identificación de las mutaciones en función del análisis de k-meros de las lecturas de secuencias de (c); donde dicho análisis de k-meros comprende

(i) La descomposición de dichas lecturas de secuencia en k-meros de la secuencia objetivo;

(ii) La determinación del número de ocasiones en que se da cada uno e dichos k-meros de la secuencia objetivo procedentes de dichas lecturas de secuenciación descompuestas;

(iii) Determinación de un recuento de elementos naturales y de un recuento de mutantes para los k-meros de la secuencia objetivo a partir de dichas lecturas de secuenciación descompuestas utilizando las secuencias de referencia de las secuencias objetivo, a fin de identificar todos los posibles k-meros que contengan una base específica en una posición específica de la secuencia de referencia o de la secuencia objetivo,

donde, para cada base de dicha secuencia de referencia, el recuento de elementos naturales viene determinado por el recuento del número de k-meros identificados a partir de dichas secuencias de referencia que incluyen dicha base en dichos k-meros de la secuencia objetivo procedente de las lecturas de secuenciación descompuestas;

donde, para cada base situada en una posición específica de dichas secuencias de referencia, dicho recuento de mutantes se determina cambiando dicha base en dicha secuencia de referencia a una base diferente, a fin de obtener una secuencia de referencia alterada, identificando todos los posibles k-meros que contengan la base cambiada, y contando el número de dichos k-meros que contienen la base cambiada de entre dichos k-meros de la secuencia objetivo procedentes de dichas lecturas de secuenciación descompuestas;

(iv) Identificación de una mutación de dicha base en las secuencias objetivo de dicha lectura de secuencias comparando la proporción entre el recuento de mutantes y el recuento de elementos naturales correspondiente a la base con la proporción entre el recuento de elementos mutantes y naturales para el resto de bases de dichas secuencias objetivo procedentes de dichas lecturas de secuenciación descompuestas; donde una proporción entre el recuento de mutantes y el recuento de elementos naturales correspondiente a dicha base que sea estadísticamente y significativamente diferente a la proporción entre el recuento de mutantes y elementos naturales correspondiente a la demás bases de dichas secuencias objetivo indicaría una mutación de dicha base en dicha secuencia objetivo; y

(e) Identificación de los elementos individuales de la población que comprenden la mutación o mutaciones identificadas en las secuencias objetivo.

Etapa (a): Agrupación del ADN genómico aislado a partir de cada miembro de la población en una o más dimensiones.

[0026] En el estado de la técnica se conocen procedimientos para el aislamiento del ADN genómico. Un experto en la materia observaría fácilmente que la calidad del ADN genómico puede influir en la metodología de cribado, y de este modo, son preferibles aquellos protocolos con los que se obtiene un ADN genómico de alta calidad con un mínimo de contaminación. Además, cualquier experto en la materia apreciaría al momento que en el comercio se encuentran disponibles kits para el aislamiento del ADN genómico (por ejemplo, el Purelink™ Genomic Kit de Invitrogen, o el Wizard® Genomic DNA Purification Kit de Promega).

[0027] Se utilizan cantidades equimolares de ADN genómico procedente de cada muestra en uno o más grupos. El agrupamiento del ADN genómico aislado puede efectuarse en una o más dimensiones (es decir, agrupamiento unidimensional o poli-dimensional) En el caso del agrupamiento unidimensional, cada elemento de la población se encuentra representado una vez (es decir, en un único grupo). En el caso del agrupamiento bidimensional, cada elemento de la población se encuentra representado dos veces (es decir, en dos grupos). En el caso del agrupamiento tridimensional, cada elemento de la población se encuentra representado en tres ocasiones (es decir, en tres grupos). En la técnica se conocen ya estrategias de agrupamiento adecuadas (véase, por ejemplo, Tsai et al., Plant Physiology 156:1257-1268). Cualquier experto en la materia tendría conocimiento de los procedimientos adecuados para la identificación de cada grupo en los procedimientos de agrupamiento multidimensional (dichos procedimientos pueden incluir etiquetas o códigos de barras).

[0028] En una realización, se agrupan cantidades equimolares de ADN genómico procedente de cada uno de los pocillos de una placa de 96 pocillos para crear una placa de grupo. En otra realización, se agrupan cantidades equimolares de ADN genómico procedente de cada uno de los pocillos de una placa de 384 pocillos para crear una placa de grupo. En las realizaciones en las que se recurre al agrupamiento unidimensional, cada elemento de la población se representa una vez (es decir, una única placa de pocillos). En las realizaciones en las que se utiliza el agrupamiento bidimensional, cada uno de los elementos de la población se representa en dos ocasiones (es decir, en un grupo de filas y un grupo de columnas). En las realizaciones en las que se utiliza el agrupamiento tridimensional, cada elemento se representa tres veces (es decir, un grupo de fila, un grupo de columna y un grupo de placa).

[0029] Cualquier experto en la materia observaría que la cantidad de ADN procedente de cada muestra dependerá del número de amplicones necesarios. En determinadas realizaciones, y para reducir el impacto de los errores de la polimerasa del ADN en una fase inicial, en una sola reacción PCR se utilizan 30 copias diploides del genoma de cada uno de los individuos de un pocillo.

[0030] En determinadas realizaciones, se agrupa una cifra mayor de un 50 copias del genoma de cada uno de los individuos de un pocillo. Un experto en la materia podría determinar con facilidad la cantidad de ADN. Por ejemplo, en el caso de la petunia, al menos 40 copias del genoma de cada planta individual equivalen a \square 20 ng. En el caso del tomate, se encuentran presentes al menos 25 copias del genoma en 50 ng de ADN extractado de 6 x 384 individuos. En el del pimiento, se encuentran presentes al menos 6 copias del genoma en 50ng de ADN extractado de 6 x 384 individuos. Para reducir aún más los impactos de los errores de la PCR, pueden llevarse a cabo duplicados de las reacciones de la PCR.

Etapa (b): Amplificación de la secuencia o secuencias objetivo en el ADN genómico agrupado.

[0031] Para cada grupo, el ADN genómico agrupado se utiliza como plantilla para las reacciones en cadena de la polimerasa (PCR) que producen amplicones para una o más secuencias objetivo. Cada reacción PCR preferiblemente amplifica una única región de la secuencia objetivo. A fin de reducir el número de errores de la polimerasa del ADN que se propagan mediante la PCR, pueden llevarse a cabo múltiples reacciones PCR utilizando ADN procedente del grupo, combinándose a continuación. Opcionalmente, las reacciones PCR se purifican (por ejemplo, mediante purificación en columna) antes de su combinación. En determinadas realizaciones, se efectúan de 3 a 12 reacciones PCR utilizando el ADN del grupo y combinándolas después. En ciertas realizaciones, se llevan a cabo 5 reacciones PCR utilizando ADN del grupo combinándolas a continuación.

Cualquier experto en la materia se daría cuenta muy pronto de que los errores de la polimerasa del ADN también pueden minimizarse mediante la utilización de una enzima de alta fidelidad, como Kapa Taq (Kapa Biosystems), Platinum Taq (Invitrogen), PfuUltra (Agilent Technologies) o Phusion (New England Biolabs).

[0032] Un experto en la materia vería muy pronto que existen procedimientos para determinar si la reacción PCR ha tenido éxito, así como la cantidad de ADN obtenida. Además, un experto en la materia observaría al momento que existen procedimientos para concentrar y limpiar una muestra de PCR.

[0033] Un experto vería que no todas las polimerasas del ADN comerciales son capaces de polimerizar la misma longitud de amplicón, y que no todas las regiones del ADN pueden amplificarse con la misma eficacia. Se eligen iniciadores que amplifiquen regiones de interés, a fin de maximizar la longitud de la secuencia objetivo y obtener una sola banda robusta al visualizarse en un gel de agarosa. Normalmente, las dimensiones del amplicón varían desde 1000 bp a más de 6500 bp, en función de la longitud de la región que se está amplificando y de la polimerasa del ADN utilizada. En aquellos casos en los que la región que interesa es mayor que lo que puede producirse en un solo producto PCR, la región de interés se amplifica como dos o más productos PCR de menores dimensiones que se solapan. Se generan al menos 200 bp de solapamiento entre amplicones. Esto se lleva a cabo para compensar la baja cobertura de secuenciamiento que suele encontrarse en los extremos 5' y 3' del producto que se está secuenciando. Un experto en la materia se daría cuenta de que las condiciones de la PCR utilizada dependerán de la polimerasa del ADN utilizada, de los cebadores seleccionados y de la calidad del ADN de la plantilla de la PCR.

Opcionalmente, los productos de amplificación de la fase (b) se agrupan antes de secuenciar los productos amplificados.

[0034] Los productos de amplificación de la etapa (b) pueden combinarse opcionalmente en cantidades equimolares con anterioridad a la secuenciación. Por ejemplo, pueden combinarse cantidades equimolares de ADN procedente de cuatro grupos amplicón de 96 pocillos que tengan como objetivo la misma región de la secuencia objetivo para obtener un grupo amplicón de 384 pocillos para una región de la secuencia objetivo. Alternativamente, se utiliza una sola placa de 384 pocillos para obtener el grupo de 384 pocillos. A continuación se pueden combinar cantidades equimolares de varios de estos grupos amplicón de 384 pocillos orientados a diferentes regiones de la secuencia objetivo, o bien pueden entonces combinarse las secuencias objetivo. En una realización, se combinaron ocho grupos amplicón de 384 pocillos. El número de placas de 384 pocillos dependerá del tamaño de la población. En ciertas realizaciones, el número de placas de 384 pocillos varía entre 1 y 15 placas de amplicón de 384 pocillos.

[0035] En determinadas realizaciones, un número suficiente de grupos amplicón orientados a diferentes regiones situadas dentro de la secuencia objetivo se combinan de forma que en el grupo de la biblioteca se represente la secuencia objetivo completa. En otras realizaciones, un número suficiente de grupos amplicón orientados a diferentes secuencias objetivo se combina para obtener el grupo de la biblioteca.

[0036] En determinadas realizaciones, se combinan cantidades equimolares de cuatro grupos amplicón de 96 pocillos orientados a una única región de la secuencia objetivo (o una única secuencia objetivo) para obtener un grupo amplicón de 384 pocillos. En otras realizaciones, se utiliza una única placa de 384 pocillos para obtener el grupo amplicón de 384 pocillos. A continuación se combinan cantidades equimolares de grupos amplicón múltiples de 384 pocillos orientados a diferentes regiones de la secuencia objetivo o a diferentes secuencias objetivo para obtener un grupo de biblioteca. En determinadas realizaciones, se combinan ocho grupos amplicón de 384 pocillos orientados a regiones solapadas de la secuencia objetivo.

[0037] Un experto en la materia sabría inmediatamente cómo concentrar y limpiar el grupo amplicón de 384 pocillos antes de combinar múltiples grupos. En la técnica se conocen procedimientos de preparación de una muestra para su secuenciación, y existen kits disponibles en el comercio (por ejemplo, la tecnología de secuenciación de la siguiente generación de Illumina). En función de la tecnología de secuenciación utilizada, los procedimientos de preparación de una muestra pueden incluir, por ejemplo, la fragmentación aleatoria de la muestra de ADN y el ligado de los adaptadores 5' y 3' con anterioridad a la amplificación de la PCR. En determinadas realizaciones, la dimensión promedio del inserto de la biblioteca de secuenciación se fija a la longitud de lectura de la serie de secuenciaciones, de forma que se maximice el solapamiento entre las lecturas directa e inversa. En determinadas realizaciones, el tamaño medio del inserto de la biblioteca de secuenciación se fija en 100 pares de base.

Etapa (c): Secuenciación de los productos amplificados u obtención de las lecturas de secuencia de los productos amplificados.

[0038] La secuenciación puede llevarse a cabo mediante cualquier procedimiento conocido en la técnica actual. En determinadas realizaciones, la secuenciación se lleva a cabo utilizando procedimientos de secuenciación de alto rendimiento. Dichos procedimientos se conocen en la técnica, e incluyen, por ejemplo, tecnologías de secuenciación de la siguiente generación, como la secuenciación por síntesis de Illumina o la secuenciación por semiconductores de ThermoFisher Scientific. En determinadas realizaciones, la secuenciación la lleva a cabo un tercero.

Opcionalmente, en la etapa (c), la secuenciación es del tipo pareado, e incluye adicionalmente la fusión de las lecturas pareadas en lecturas compuestas.

[0039] En determinadas realizaciones, la secuenciación es del tipo pareado. Las lecturas directas e inversas se combinan opcionalmente en una sola lectura compuesta. Las identificaciones de bases con una posibilidad de error de $> 1/10.000$ o $1/100.000$ se eliminan o enmascaran. En determinadas realizaciones, la secuenciación de tipo pareado la lleva a cabo un tercero, siendo dicho tercero quien obtiene los datos de secuenciación de tipo pareado.

[0040] Cualquier experto en la materia apreciaría que los pares de lectura directa o inversa son reacciones de secuenciación independientes a lo largo de la misma molécula de plantilla. Dicho experto también se daría cuenta de que cuando las identificaciones de base procedentes de lecturas alienadas concuerdan tanto en la dirección directa como en la inversa, la confianza en que la base se identifique correctamente es mayor. Rodrigue et al. (PLoS One 4:34761) demostraron que la combinación de los pares de lectura directa e inversa procedentes de una serie de secuenciaciones tipo pareado de Illumina reduce la tasa de errores de secuenciación en 2 órdenes de magnitud. Con una tasa de error de $1/100.000$ o mejor, pueden secuenciarse de una sola vez las muestras de ADN procedentes de miles de individuos sin que las mutaciones se pierdan en un océano de ruido.

[0041] Un experto en la materia percibiría que se dispone de un software, como SHERA ((Rodrigue et al, PLoS One 4:34761) o PEAR (Zhang et al., Bioinformatics 30(5): 614-620), que podría utilizarse para obtener lecturas compuestas a partir de lecturas de tipo pareado. Las alternativas a SHERA y PEAR incluyen COPE (Liu et al, Bioinformatics 28(22): 2870-2874, FLASH (Magoč and Salzberg, Bioinformatics 27(21): 2957-2963), y PANDASeq (Masella et al, BMC Bioinformatics 13:31).

Etapa (d) Identificación de las mutaciones en función del análisis de secuencia sin alineación de los datos de secuenciación

[0042] Pueden analizarse las secuencias e identificarse las mutaciones utilizando enfoques basados en alineación o enfoques sin alineación, y las secuencias se analizan recurriendo a procedimientos sin alineación. En la técnica se conocen varios procedimientos de análisis de secuencias sin alineación, entre las cuales se incluyen los procedimientos basados en la frecuencia de K-meros. Cualquier experto en la materia se daría cuenta de que los k-meros (también conocidos como kmeros) son fragmentos cortos y solapados generados a partir de una secuencia más larga. Los fragmentos pueden tener cualquier longitud igual o más corta que la longitud de la lectura de secuenciación a partir de la cual se han generado. En una realización, los k-meros tienen de 15 a 31 nucleótidos. En una realización, los k-meros tienen de 15 a 31 nucleótidos. En una realización específica, los k-meros tienen una longitud de 17 nucleótidos.

[0043] Cualquier experto en la materia se daría cuenta de que una lectura de secuenciación procedente del grupo de ADN genómico aislado a partir de miembros de una población incluirá (1) secuencias naturales y (2) secuencias no naturales. Las secuencias no naturales pueden ser el resultado de mutaciones reales o errores de secuenciación.

[0044] Una secuencia de referencia se puede utilizar para identificar a los k-meros resultantes de secuencias de tipo natural en las lecturas de secuenciaciones, y determinar un recuento de elementos naturales para cada base. En ciertas realizaciones, el recuento de elementos naturales viene determinado por la frecuencia de determinación de todos los k-meros identificados a partir de las secuencias de referencia que incluyen la base en k-meros de la secuencia objetivo a partir de las lecturas de secuenciación descompuestas.

[0045] La secuencia de referencia también puede utilizarse para generar secuencias alteradas (es decir, secuencias de un tipo no natural) y k-meros resultantes de secuencias no naturales (es decir, mutantes o errores de secuenciación) en las lecturas de secuenciación y determinar un recuento de mutantes (*es decir, recuento de elementos no naturales*) para cada base. Concretamente, en ciertas realizaciones, para cada base de la secuencia o secuencias de referencia, el recuento de mutantes se determina cambiando la base de las secuencias de referencia a una base diferente, para obtener unas secuencias de referencia alteradas, identificando todos los k-meros que contienen la base cambiada, y determinando la frecuencia de los k-meros que contienen la base cambiada en los k-meros de la secuencia objetivo a partir de las lecturas de secuenciación descompuestas. Este proceso puede repetirse para cada posición de la secuencia de referencia, y se pueden utilizar diferentes bases para obtener la secuencia o secuencias de referencia alteradas.

[0046] Una mutación real (es decir, no un error de secuenciación) de la secuencia o secuencias objetivo se identifica (*es decir, que se distingue* de las bases no naturales obtenidas a partir de los errores de secuenciación) comparando la proporción entre elementos mutantes y naturales para la base con la proporción entre mutantes y naturales para el resto de las bases de las secuencias objetivo. Una proporción que sea significativamente diferente de la proporción entre recuentos de elementos mutantes y naturales para el resto de bases de las secuencias objetivo indica una mutación real de dicha base en la secuencia o secuencias objetivo.

[0047] Un experto en la materia observaría inmediatamente las secuencias de referencia apropiadas. Por ejemplo, una secuencia de referencia puede ser la secuencia conocida, incluyendo, entre otras, secuencias obtenidas de las bases de datos públicas, entre ellas la International Nucleotide Sequence Database, el DNA Data Bank of Japan, EMBL, GenBank u otras bases de datos de genomas disponibles públicamente, como Plant GDB. También pueden identificarse las secuencias de referencia a partir de los genomas secuenciados de diversos organismos, incluyendo, entre otros, *Arabidopsis thaliana*, tomate, pimiento, pepino, melón, uva, manzana, melocotón, berenjena, trigo, cebada, maíz, arroz, patata, batata, y rosa. También podría determinarse una secuencia de referencia utilizando la información de secuenciación procedente de los productos de la PCR de los amplicones.

[0048] En determinadas realizaciones, el análisis de los k-meros comprende (i) la descomposición en k-meros de dichas secuencias leídas de las secuencias objetivo; (ii) determinación del número de apariciones de cada uno de

los k-meros obtenidos de las lecturas de secuenciación descompuestas de las secuencias objetivo; (iii) determinación del recuento de elementos naturales y el recuento de mutantes para los k-meros de la secuencia objetivo, utilizando las secuencias de referencia para identificar todos los posibles k-meros que contienen una base particular de la secuencia objetivo o secuencia de referencia. Para cada base de la secuencia de referencia, el recuento de elementos naturales se determina mediante la determinación de la frecuencia de todos los k-meros identificados a partir de la secuencia o secuencias de referencia que incluyen la base en los k-meros de la secuencia objetivo obtenidos a partir de las lecturas de secuenciación descompuestas. Para cada base de la secuencia o secuencias de referencia, el recuento de mutantes se determina cambiando la base en las secuencias de referencia a una base diferente, a fin de obtener una secuencia de referencia alterada, identificando todos los k-meros que contienen la base modificada y determinando la frecuencia de los k-meros que contienen la base cambiada en los k-meros de la secuencia objetivo obtenida a partir de las lecturas de secuenciación descompuestas. Para cada base, una mutación (es decir, no un error de secuenciación) en las secuencias objetivo se identifica comparando la proporción entre el recuento de mutantes y el recuento de elementos naturales correspondientes a la base con la proporción entre el recuento de mutantes y el recuento de elementos naturales para el resto de las bases de la secuencia objetivo. Una proporción que sea significativamente diferente a la proporción entre los recuentos de mutantes y elementos naturales para todas las demás bases de la secuencia objetivo indicará una mutación de dicha base en la secuencia objetivo.

Etapa opcional (e) Identificación de los elementos individuales de la población que comprenden la mutación o mutaciones identificadas en las secuencias objetivo.

[0049] En determinadas realizaciones, el procedimiento incluye igualmente la identificación de los elementos individuales de la población que comprenden la mutación o mutaciones identificadas en las secuencias objetivo. Puede recurrirse a diversos procedimientos para identificar los elementos individuales que comprenden las mutaciones que nos interesan. Por ejemplo, dichos procedimientos pueden incluir la Fusión de Alta Resolución, procedimientos enzimáticos

[0050] (como la digestión de los puntos de desajuste en heterodúplex con endonucleasa CELI), procedimientos basados en hibridación, procedimientos de secuenciación y procedimientos que utilizan la agrupación multidimensional y la identificación mediante códigos de barras (como la agrupación tridimensional).

[0051] En una realización, se utiliza la Fusión de Alta Resolución (HRM) para identificar aquellos elementos de la población que comprenden la mutación o mutaciones identificadas en la secuencia o secuencias objetivo. Cualquier experto en la materia conocerá diversos procedimientos de HRM. Véase, por ejemplo, Erali and Witter (Methods 50(4):250-261). Concretamente, la HRM puede llevarse a cabo utilizando iniciadores que flanqueen la mutación identificada, bien por sí sola o en combinación una sonda de nucleótidos de bloques 3' (como la sonda 'LunaProbe' descrita por Idaho Technology) y el ADN genómico de los individuos de la población, que puede estar agrupado o no.

[0052] En determinadas realizaciones, una vez que se ha detectado la presencia de una mutación en una población, la muestra individual de ADN que contiene la mutación se identifica utilizando la tecnología HRM (De Koeber et al, Molecular Breeding 25: 67-90). En algunas realizaciones, los iniciadores de la PCR que flanquean la mutación que interesa se crean y utilizan para amplificar un producto que contiene el emplazamiento de la mutación en cada una de las muestras de ADN de los grupos de 384 pocillos en los que se ha identificado la mutación que reviste interés. Los iniciadores de la PCR pueden diseñarse de forma que el tamaño del amplicón sea inferior a 75 bp y no se den posiciones del ADN heterocigótico naturales. En determinadas realizaciones, la muestra de ADN aislada que contiene la mutación se identifica mediante un análisis de la curva de fusión. Por ejemplo, pueden utilizarse un scanner de 384 pocillos LightScanner (Idaho Technology) y tinturas HRM LCGreen Plus en el análisis de la curva de fusión. Opcionalmente, la presencia de la mutación se puede confirmar. En determinadas realizaciones, para confirmar la mutación se plantan y siembran las semillas recolectadas de las plantas cuyo ADN se ha aportado a la muestra. Se obtienen los tejidos de dichas plantas y se analiza su ADN mediante secuenciación, de forma que se identifiquen las plantas individuales que presentan la mutación.

[0053] Opcionalmente, la presencia de la mutación puede confirmarse en el individuo identificado mediante otros procedimientos de detección del SNP.

Análisis fenotípico

[0054] La evaluación fenotípica de las plantas puede realizarse para determinar si las mutaciones de interés producen un efecto sobre el rendimiento de la planta en diversas condiciones. Entre los tipos de análisis fenotípico se encuentran, entre otros, la evaluación de las respuestas a la tensión por sequía, el crecimiento a bajas temperaturas, la tolerancia al calor, la resistencia a los patógenos, el rendimiento, los cambios morfológicos (incluyendo, entre otros, la altura y tamaño de la planta y/o el color de las hojas, semillas y/o flores), la modificación del ciclo vital y/o la susceptibilidad a enfermedades.

KITS

[0055] Existen kits que incluyen uno o más de los reactivos necesarios para los procedimientos que se indican en este documento. Por ejemplo, los kits pueden incluir uno o más iniciadores, sondas, polimerasa del ADN e instrucciones de uso.

[0056] Para comprender mejor la invención descrita en este documento, se facilitan los ejemplos siguientes. Se entenderá que los ejemplos pretenden describir realizaciones ilustrativas de la invención, y que no pretenden limitar en modo alguno el alcance de la invención.

5 EJEMPLO 1

[0057] En una realización de la presente invención, el procedimiento de cribado de una población que comprende una mutación en una o más secuencias objetivo comprende las siguientes etapas:

- 10 1) Creación de grupos amplicón estequiométricamente equilibrados para su secuenciación.
 - (a) Agrupamiento de placas – Se ha extraído el ADN de cada uno de los individuos de la población EMS y se ha dispuesto en placas de microtitulación de 96 pocillos. Para cada muestra de una placa de 96 pocillos se han alicuotado cantidades equimolares en un único tubo. Estos son los grupos de placas.
 - 15 (b) PCR de amplicones – Diseño de iniciadores de la PCR con los que se obtenga un producto PCR para las regiones del ADN que se están investigando. Para cada amplicón se llevaron a cabo múltiples reacciones PCR utilizando como plantilla el ADN procedente de un grupo de placas. Esto se hizo para reducir la concentración global de errores de la polimerasa del ADN propagados a través de la PCR para el amplicón. Las replicas de las PCRs se agruparon en un único tubo, visualizándose en gel de agarosa, y se cuantificaron. Esta operación se llevó a cabo para cada grupo de placas. Estas muestras son los grupos amplicón
 - 20 (c) Creación de grupos de placas de 384 pocillos – Se agruparon cantidades equimolares procedentes de 4 grupos amplicón en un solo tubo. Estos son los grupos amplicón de 384 pocillos. Esto se llevó a cabo para que todos los individuos estuviesen representados en nuestras placas HRM de 384 pocillos en un solo grupo. De este modo se reduce la cantidad de trabajo que se necesita durante la fase de genotipado de la identificación/verificación del mutante. Cada grupo amplicón de 384 pocillos se hizo pasar por una columna de limpieza del producto PCR para limpiar y concentrar las muestras. El ADN procedente de la columna se eluyó en el más reducido volumen de agua razonable. Se cuantificaron los grupos amplicón limpios.
 - 25 (d) Elaboración de la biblioteca de grupos a secuenciar – Todos los amplicones utilizados en los mismos grupos de placas de 496 pocillos se combinaron en un solo grupo de bibliotecas, tomando la precaución de alicuotar los amplicones en cantidades equimolares. Cada uno de los grupos de bibliotecas se secuenció en un instrumento de secuenciación de Illumina.
- 30 2. Identificación de la mutación
 - (a) Fusión de lecturas pareadas (PE) – Las lecturas PE representan reacciones de secuenciación independientes de los dos extremos de la misma plantilla. Cuando la plantilla se aproxima a la misma longitud que las lecturas de secuenciación se solapan los extremos pareados. Estas lecturas pueden fusionarse en una única lectura compuesta en la que tendremos una mayor certidumbre en la identificación de bases. De este modo se reducen drásticamente los errores causados por errores de identificación durante la reacción de secuenciación. Se utilizó el software PEAR para crear las lecturas compuestas.
 - 35 (b) Creación de la tabla de k-meros – Las lecturas fusionadas se descompusieron en fragmentos cortos y solapados denominados k-meros. Los fragmentos pueden tener cualquier longitud, pero normalmente funcionan bien los de 15 a 31. La selección de un tamaño de k-mero puede realizarse empíricamente o mediante el examen de la secuencia de referencia. Una longitud de k-mero óptima obtenida de las secuencias de referencia es una longitud en la que todos los k-meros generados a partir de la secuencia de referencia están separados entre sí a una distancia de Hamming de 1, y preferiblemente de 2. El número de veces que se encuentra presente cada uno de los k-meros en los datos de secuenciación de alta calidad se registró en una tabla de búsquedas. Se utilizó el software Jellyfish para crear la tabla de búsquedas. Una vez construida la tabla de búsquedas, las secuencias de referencia se atravesaron una base cada vez. Para cada base, se determinaron todos los k-meros de la secuencia de referencia, y el complemento inverso de dichos k-meros que la incluyen, y se promedió el número de veces que se encontraron dichos k-meros en los datos de secuenciación (a partir de la tabla de búsquedas de k-meros). Esto nos da una medida de la “cobertura” del tipo natural (es decir, el recuento de elementos naturales). A continuación, y paso a paso, se cambia cada una de las G de la referencia a A o C a T y se cuenta el número de veces que todos los k-meros que incluyen el cambio se encuentran presentes en los datos de secuenciación (a partir de la tabla de búsquedas). El promedio de este valor da una medida de la mutación (es decir, el recuento de mutantes). Esta etapa puede ampliarse a fin de cambiar cada una de las bases de la referencia a cada uno de los tres nucleótidos restantes.
 - 45 (c) Selección de la mutación – Se identificó una mutación real (es decir, no un error de secuenciación) de la secuencia objetivo (es decir, distinguiéndola de las bases del tipo no natural derivadas de errores de secuenciación) comparando la proporción entre el recuento de mutantes y elementos naturales correspondiente a la base con la proporción entre el recuento de mutantes y el recuento de elementos naturales para el resto de las bases de las secuencias objetivo. Una proporción que sea significativamente diferente a la proporción entre recuentos de mutantes y elementos naturales para el resto de bases de la secuencia objetivo indicará la existencia de una mutación real de dicha base en la secuencia objetivo.
 - 50 3) La identificación del elemento de la población que comprende la mutación -HRM se utilizó para el análisis genotípico de la población a fin de encontrar plantas que contengan las mutaciones de mayor interés identificadas en la fase anterior. La profundidad de las búsquedas se limita mediante la identificación de las 384 bibliotecas que contienen la mutación.
 - 55 60 65

EJEMPLO 2

5 **[0058]** En el siguiente ejemplo se amplificaron mediante PCR seis genes procedentes de una población EMS de tomate obtenida de 8 grupos consistentes en ADN. Cada grupo contenía ADN de 384 individuos. Para cada uno de los 8 grupos se mezclaron los tres amplicones en cantidades equimolares. A continuación los 8 grupos amplicón se utilizaron como plantilla para la construcción de bibliotecas Illumina Nextera y se secuenciaron en un instrumento Illumina MiSeq utilizando un kit de reacción de 600 ciclos versión 3. Las lecturas por pares se fusionaron utilizando el software PEAR y las lecturas fusionadas y no fusionadas resultantes se utilizaron en posteriores análisis.

10 **[0059]** Mediante el software Jellyfish (Marcais et al, 2011) y los resultados de PEAR, se seleccionó un tamaño de k-mero de 15, se creó una tabla que indicaba cada uno de los k-meros que aparecían en los datos de secuenciación, así como el número de veces que se había encontrado ese k-mero. Tan sólo se utilizaron los k-meros todas cuyas bases tenían una puntuación de calidad > 30.

15 **[0060]** A continuación, para cada base de ADN de las secuencias de referencia se generó la totalidad de los 15-meros que era posible que incluyeran dicha posición. Utilizando estos 15-meros y la tabla generada con Jellyfish, el recuento total de todos los 15-meros que incorporaban la posición actual en los datos de secuenciación se sumaron y dividieron por el número de 15-meros de intersección. Este valor representaba el recuento de elementos 'naturales' correspondiente a esta posición. La identidad de los elementos naturales de esa base se modificó posteriormente (se "mutó") a cada una de las otras tres bases de ADN. Para cada una de estas otras tres bases se repitió el mismo procedimiento descrito anteriormente para determinar el recuento total correspondiente a todos los 15-meros de intersección correspondientes a las bases 'mutadas' que se habían generado utilizando la secuencia modificada. El más elevado de estos valores representaba el "recuento de mutantes" para esa posición de la referencia. Se calculó un recuento normalizado de mutantes para cada posición, comparándose los valores obtenidos para las 8 bibliotecas. Las bibliotecas cuya proporción difería significativamente de las otras indicaba que un miembro del grupo de la población a partir del cual se había generado la biblioteca comprende una mutación en dicha posición. En este ejemplo, la proporción utilizada era el cociente entre el recuento de mutantes (mt) y el recuento de elementos naturales (wt) correspondiente a la biblioteca con el mayor recuento de mutantes en dicha posición, dividido por el cociente de la suma de los recuentos de mutantes del resto de las bibliotecas en dicha posición (otro-mt) dividida entre los recuentos de elementos naturales del resto de las bibliotecas en dicha posición (otro-wt) $[(mt/wt)/(otro-mt/otro-wt)]$. Se aplicó a estos datos una distribución de Poisson, utilizándose para calcular los valores p para cada una de dichas proporciones. Se asignó empíricamente un valor p de corte, utilizándose como parámetro para seleccionar cuáles de las mutaciones putativas eran reales. El elemento individual de la población que contenía la mutación identificada se identificó utilizando una fusión de ADN de alta resolución. Mediante esta técnica, se identificaron los elementos individuales que comprendían una de las mutaciones que se indican a continuación.

35

Gen	Posición de referencia	Mutación	Proporción	Valor P
A1	6049	G->A	41,34	4,26e-15
A2	3075	G->A	95,05	1,19e-61
A2	3405	G->A	114,43	1,05e-81
A2	4056	G->A	61,01	4,31e-30
A3	2326	G->A	95,72	1,19e-61
A3	3509	G->A	22,13	1,11e-04
A4	1451	G->A	46,03	1,59e-18
A5	601	G->A	64,92	1,32e-32
A5	1153	G->A	55,53	2,96e-25
A5	1353	G->A	15,98	2,90e-02
A6	865	G->A	88,91	1,17e-54
A6	869	C->T	37,02	1,56e-12

A6	8528	G->A	38,75	3,70e-13
Tabla 1 – Mutaciones confirmadas. Utilizando la metodología descrita, 13 mutaciones previamente identificadas en estos genes utilizando un procedimiento alternativo.				

EJEMPLO 3

5 **[0061]** En el siguiente ejemplo se amplificaron mediante PCR 4 amplicones procedentes de 3 genes obtenidos a partir de una población de pimiento EMS procedente de 8 grupos consistentes en ADN. Cada grupo contenía ADN de 384 individuos. Para cada uno de los 9 grupos se mezclaron los cuatro amplicones en cantidades equimolares. A continuación los 9 grupos amplicón se utilizaron como plantilla para la construcción de bibliotecas Illumina Nextera y se secuenciaron en un instrumento Illumina MiSeq utilizando un kit de reacción de 600 ciclos versión 3. Utilizando el software Jellyfish (Marcais et al, 2011) con un tamaño de k-mero de 17 se creó una tabla de cada k-mero que aparecía en los datos de secuenciación, así como el número de veces que se había encontrado cada k-mero.

10 **[0062]** A continuación, para cada base de ADN de las secuencias de referencia se generó la totalidad de los 17-meros posibles que incluían dicha posición. Utilizando estos 17-meros y la tabla generada con Jellyfish, el recuento total correspondiente a los 17-meros que incorporan la posición actual en los datos de secuenciación se sumó y dividió por el número de 17-meros de intersección. Este valor representaba el recuento de elementos 'naturales' correspondiente a esta posición. La identidad de los elementos naturales de esa base se modificó posteriormente (se "mutó") a cada una de las otras tres bases de ADN. Para cada una de estas otras tres bases se repitió el mismo procedimiento descrito anteriormente para determinar el recuento total correspondiente a todos los 17-meros de intersección correspondientes a las bases 'mutadas' que se habían generado utilizando la secuencia modificada. El más elevado de estos valores representaba el "recuento de mutantes" para esa posición de la referencia. Se calculó un recuento normalizado de mutantes para cada posición, comparándose los valores obtenidos para las 9 bibliotecas. Las bibliotecas cuya proporción difería significativamente de las otras indicaban que un miembro del grupo de la población a partir del cual se había generado la biblioteca comprende una mutación en dicha posición. En este ejemplo, la proporción utilizada era el cociente entre el recuento de mutantes (mt) y el recuento de elementos naturales (wt) correspondiente a la biblioteca con el mayor recuento de mutantes en dicha posición, dividido por el cociente de la suma de los recuentos de mutantes del resto de las bibliotecas en dicha posición (otro-mt) dividido por los recuentos de elementos naturales del resto de las bibliotecas en dicha posición (otro-wt) $[(mt/wt)/(otro-mt/otro-wt)]$. Se aplicó a estos datos una distribución de Poisson, utilizándose para calcular los valores p para cada una de dichas proporciones. Se asignó empíricamente un valor p de corte, utilizándose como parámetro para seleccionar cuáles de las mutaciones putativas eran reales. El elemento individual de la población que contenía la mutación identificada se identificó utilizando una fusión de ADN de alta resolución. Mediante esta técnica, se identificaron los elementos individuales que comprendían una de las mutaciones que se indican a continuación.

15

20

25

30

35

ES 2 714 851 T3

Tabla 2 – Mutaciones confirmadas. Utilizando la metodología descrita, la totalidad de las 24 mutaciones previamente identificadas en estos genes utilizando un procedimiento alternativo

Gen	Posición de referencia	Mutación	Proporción	Valor P
B	1140	G->A	199,24	1,32e-185
D	687	G->A	104,12	6,72e-71
B	3204	C->T	79,28	5,80e-46
B	3070	C->T	95,47	1,37e-61
A	4589	C->T	116,16	7,97e-84
A	18659	G->A	69,59	7,00e-37
D	1591	C->T	48,54	6,20e-20
A	1311	G->A	121,05	2,27e-89
D	1255	C->T	75,10	2,91e-42
B	3744	G->A	74,25	2,38e-41
A	5667	G->A	47,17	3,27e-19
B	6096	G->A	58,61	1,32e-27
A	18214	G->A	113,13	1,53e-80
A	5566	G->A	86,64	1,21e-52
D	1063	C->T	32,01	1,45e-09
B	7948	C->T	41,20	4,50e-15
D	825	G->A	39,78	9,00e-14
A	1061	C->T	38,46	3,88e-13
A	19666	C->T	34,07	1,04e-10
B	3421	C->T	47,78	3,27e-19
A	3009	C->T	49,63	1,15e-20
B	7885	C->T	30,95	1,80e-08
D	1228	C->T	28,03	1,97e-07
D	1065	G->A	34,67	1,04e-10

5 EJEMPLO 4

[0063] En el siguiente ejemplo, la metodología de la invención para el descubrimiento de mutantes sin alineación se aplicó a un conjunto de datos publicado para someterlo a TILLING mediante secuenciación en una población de arroz. (Tsai et al, Plant Physiology 156(3):1257-1268). De acuerdo con la publicación, el ADN de 768 plantas de arroz se alicuotó en una estrategia de agrupamiento tridimensional; 16 grupos de filas de 48 individuos cada uno, 16 grupos de columnas de 48 individuos cada uno, y 12 grupos dimensionales de 64 individuos cada uno. Hemos

tratado de recapitular los resultados de la transición del nucleótido para uno de los genes objetivo definidos en la publicación original. Se descargaron cuarenta y ocho bibliotecas consistentes en 247.895.886 lecturas de pares de base realizadas mediante Illumina single-end 79 del Genbank Short Read Archive con el identificador 'SRP006801'. Utilizando el software Jellyfish (Marcais et al, 2011) con un tamaño de k-mero de 17 se creó una tabla de cada k-mero que apareció en los datos de secuenciación, así como el número de veces que se encontró dicho k-mero. Se descartaron del análisis todos los k-meros con un recuento <2 , ya que representaban errores de secuenciación, y no una genuina secuencia de mutación. Posteriormente, para cada base de ADN de la secuencia de referencia, OS01G63510, se generaron todos los 17-meros posibles que incluían dicha posición. Utilizando estos 17-meros y la tabla generada con Jellyfish, el recuento total de todos los 17-meros que incorporan la posición actual procedentes de los datos de secuenciación se sumó y dividió por el número de 17-meros de intersección. Este valor representaba el recuento de elementos "naturales" para esta posición. A continuación se cambió ("mutó") la identificación del tipo natural de dicha base a cada una de las otras tres bases de ADN. Para cada una de estas otras tres bases se llevó a cabo el mismo procedimiento que se ha descrito anteriormente para determinar el recuento total de todos los 17-meros de intersección correspondientes a las bases "mutadas" que habían sido generadas utilizando la secuencia alterada. El más elevado de estos valores representó el "recuento de mutantes" correspondiente a dicha posición de la referencia.

[0064] De acuerdo con Tsai et al., una mutación real debería generar importantes recuentos de la base de mutantes exactamente en 1 biblioteca de fila, 1 biblioteca de placa y 1 biblioteca dimensional. Esto es exactamente lo que se observó para las mutaciones de transición en las posiciones 530 y 777 de la secuencia de referencia que estábamos tratando de identificar. Cada una de estas posiciones tan sólo tenía 3 recuentos de alelos mutantes distintos de cero con un recuento distinto de cero en el grupo de la fila, el grupo de la placa y el grupo dimensional. El análisis se llevó a cabo en menos de una hora.

Referencias

[0065]

De Koeyer, D.L., Douglass, K., Murphy, A.M., Whitney, S., Nolan, L., Song, Y., and De Jong, W.S. (2010). "Application of high-resolution DNA melting for genotyping and variant scanning of diploid and autotetraploid potato.", *Molecular Breeding*, 25(1), pp. 67-90.

Erali M, Wittwer CT. High Resolution Melting Analysis for Gene Scanning. *Methods* (San Diego, Calif). 2010;50(4):250-261.

Fitzpatrick TB, Basset GJ, Borel P, Carrari F, DellaPenna D, Fraser PD, Hellmann H, Osorio S, Rothan C, Valpuesta V, Caris-Veyrat C, Fernie AR. Vitamin deficiencies in humans: can plant science help? *Plant Cell*. 2012 Feb;24(2):395-414.

Forster, B.P. (2001) Mutation genetics of salt tolerance in barley: A assessment of Golden Promise and other semi-dwarf mutants. *Euphytica* 120(3): 317-328.

Gady AL, Hermans FW, Van de Wal MH, van Loo EN, Visser RG, Bachem CW. Implementation of two high through-put techniques in a novel application: detecting point mutations in large EMS mutated plant populations. *Plant Methods*. 2009 Oct 7;5:13.

Koornneef, M., Dellaert, L.W., and van der Veen, J.H. (1982) EMS-and radiation-induced mutation frequencies at individual loci in *Arabidopsis thaliana* (L) Heynh. *Mutat Res* 93:109-123.

Liu B, et al. COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics*. 2012;28:2870-2874.

Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011;27:2957-2963.

Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011 Mar 15;27(6):764-70.

Masella AP, et al. PANDAsseq: paired-end assembler for Illumina sequences. *BMC Bioinformatics*. 2012;13:31.

McCallum, C., Comai, L., Greene, E.A., and Henikof, S. (2000) Targeting induced local lesions in genomes (TILLING) for plant functional genomics. *Plant Physiol* 123(2):439-442.

Minoche, A.E., Dohm, J.C., Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*. 12: R112.

- Missirian V, Comai L, Filkov V. Statistical mutation calling from sequenced overlapping DNA pools in TILLING experiments. *BMC Bioinformatics*. 2011 Jul 14;12:287.
- 5 Nordström KJ, Albani MC, James GV, Gutjahr C, Hartwig B, Turck F, Paszkowski U, Coupland G, Schneeberger K. Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nat Biotechnol*. 2013 Apr;31(4):325-30.
- Rigola D, van Oeveren J, Janssen A, et al. High-Throughput Detection of Induced Mutations and Natural Variation Using KeyPoint™ Technology. Herrera-Estrella A, ed. *PLoS ONE*. 2009;4(3):e4761.
- 10 Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, Chisholm SW. Unlocking short read sequencing for metagenomics. *PLoS One*. 2010 Jul 28;5(7):e11840
- 15 Sargent RG, Kim S, Gruenert DC. Oligo/polynucleotide-based gene modification: strategies and therapeutic potential. *Oligonucleotides*. 2011 Mar-Apr;21(2):55-75.
- Stadler, L.J. (1932) On the genetic nature of induced mutations in plants. *Proc VI Congress of Genetics* 1:274-294 (www.agris.fao.org).
- 20 Tsai, H., Howell, T., Nitcher, R., Missirian, V., Watson, B., Ngo, K.J., Lieberman, M., Fass, J., Uauy, C., Tran, R.K., Khan, A.A., Filkov, V., Tai, T.H., Dubcovsky, J., Comai, L. (2011). Discovery of rare mutations in populations: TILLING by sequencing. *Plant Physiology*. 156:1257-1268.
- 25 Wang, T. L., Uauy, C., Robson, F. and Till, B. (2012), TILLING in extremis. *Plant Biotechnology Journal*, 10: 761-772.
- Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End read mergeR. *Bioinformatics*. 2014;30(5):614-620.

REIVINDICACIONES

1. Procedimiento de identificación de los miembros de una población que comprenda una o más mutaciones en una o más secuencias objetivo, donde cada uno de dichos miembros de dicha población es distinto, incluyendo dicho procedimiento las siguientes etapas:
- 5 (a) Agrupamiento del ADN genómico aislado procedente de cada uno de los miembros de la población en una o más dimensiones;
- (b) Amplificación de la secuencia o secuencias objetivo en el ADN genómico agrupado;
- 10 (c) Secuenciación de los productos amplificados para obtener lecturas de secuencias correspondientes a los productos amplificados;
- (d) Identificación de las mutaciones en función del análisis de k-meros de las lecturas de secuencia de (c); donde dicho análisis de k-meros comprende
- (i) La descomposición de dichas lecturas de secuencias en k-meros de la secuencia objetivo;
- 15 (ii) La determinación del número de ocasiones en que se da cada uno de dichos k-meros de la secuencia objetivo procedentes de dichas lecturas de secuenciación descompuestas;
- (iii) Determinación de un recuento de tipo natural y de un recuento de mutantes para los k-meros de la secuencia objetivo a partir de dichas lecturas de secuenciación descompuestas utilizando las secuencias de referencia de las secuencias objetivo, a fin de identificar todos los posibles k-meros que contengan una base específica en una posición específica de la secuencia de referencia o de la secuencia objetivo,
- 20 donde, para cada base de dicha secuencia de referencia que se encuentra en una posición específica, el recuento de elementos naturales viene determinado por el recuento del número de k-meros identificados a partir de dichas secuencias de referencia que incluyen dicha base en dichos k-meros de la secuencia objetivo procedente de las lecturas de secuenciación descompuestas;
- donde, para cada base situada en una posición específica de dichas secuencias de referencia, dicho recuento de mutantes se determina cambiando dicha base en dicha secuencia de referencia a una base diferente, a fin de obtener una secuencia de referencia alterada, identificando todos los posibles k-meros que contengan la base cambiada, y contando el número de dichos k-meros que contienen la base cambiada en dichos k-meros de la secuencia objetivo procedentes de dichas lecturas de secuenciación descompuestas;
- 25 (iv) Identificación de una mutación de dicha base en las secuencias objetivo de dicha lectura de secuencias comparando la proporción entre el recuento de mutantes y el recuento de elementos naturales correspondiente a la base con la proporción entre el recuento de elementos mutantes y naturales para el resto de bases de dichas secuencias objetivo procedentes de dichas lecturas de secuenciación descompuestas; donde una proporción entre el recuento de mutantes y el recuento de elementos naturales correspondiente a dicha base que sea estadísticamente y significativamente diferente a la proporción entre el recuento de mutantes y elementos naturales correspondiente a la demás bases de dichas secuencias objetivo indicaría una mutación de dicha base en dicha secuencia objetivo; y
- 30 (e) Identificación de los elementos individuales de la población que comprenden la mutación o mutaciones identificadas en las secuencias objetivo.
- 40 2. Procedimiento de la reivindicación 1, donde dicha secuenciación es una secuenciación de tipo pareado, y que adicionalmente comprende la fusión de las lecturas pareadas en lecturas compuestas.
3. Procedimiento según la reivindicación 1 o 2, donde dicha etapa de amplificación comprende la amplificación de más de una muestra del ADN genómico agrupado, y el agrupamiento de los productos de amplificación procedentes de cada muestra para crear grupos de productos de amplificación.
- 45 4. Procedimiento según cualquiera de las reivindicaciones 1 a 3, donde dicha población está mutagenizada por productos químicos inductores de la mutación, radiaciones ionizantes, intercambio de nucleótidos objetivo, o mutagénesis orientada a la región.
- 50 5. Procedimiento según cualquiera de las reivindicaciones 1 a 3, donde la identificación de los elementos de la población que comprenden una o más de las mutaciones identificadas en las secuencias objetivo se lleva a cabo mediante desnaturalización de ADN a alta resolución (HRM).
- 55 6. Procedimiento según cualquiera de las reivindicaciones 1 a 5, que incluye adicionalmente, tras la etapa de identificación de los elementos individuales de la población que comprenden la mutación o mutaciones identificadas, la realización de análisis fenotípicos de dichos elementos individuales que comprenden la mutación o mutaciones identificadas en las secuencias objetivo.
- 60 7. Procedimiento según cualquiera de las reivindicaciones 1 a 6, donde dicha población es una población de plantas.
8. Procedimiento según la reivindicación 7, donde dicha planta es un cereal, una semilla oleaginosa, una fruta, verdura, biocombustible, una planta ornamental, una planta con flor, una planta anual o una planta perenne.
- 65 9. Procedimiento de la reivindicación 7, donde dicha planta se selecciona entre el grupo que comprende la petunia,

ES 2 714 851 T3

el tomate, la pimienta, lechuga, patata, cebolla, zanahoria, brócoli, apio, guisante, espinaca, impatiens, pepino, rosa, batata, árboles frutales, berenjena quimbombó, maíz, soja, colza, trigo, avena, arroz, sorgo, algodón y cebada.

REFERENCIAS CITADAS EN LA DESCRIPCIÓN

La lista de referencias citada por el solicitante lo es solamente para utilidad del lector, no formando parte de los documentos de patente europeos. Aún cuando las referencias han sido cuidadosamente recopiladas, no pueden excluirse errores u omisiones y la OEP rechaza toda responsabilidad a este respecto.

Documentos de patente citado en la descripción

- WO 2007037678 A, KeyGene N.V. [0010]
- WO 2014134729 A [0010]

10 Bibliografía no de patentes citada en la descripción

- FITZPATRICK et al.** *Plant Cell.*, vol. 24, 395-414 [0002]
- **WANG et al.** *Plant Biotechnology Journal*, vol. 10, 761-772 [0002]
 - **SARGENT et al.** *Oligonucleotides*, 2011, vol. 21 (2), 55-75 [0004]
 - **NORDSTROM et al.** *Nat. Biotech.*, 2013, vol. 31 (4), 325-331 [0005]
 - **GADY et al.** *Plant Methods*, vol. 5, 13 [0008]
 - **RIGOLA et al.** *PLoS One*, vol. 4, e4761 [0010]
 - **TSAI et al.** *Plant Physiology*, vol. 156, 1257-1268 [0010]
 - **MINOCHE et al.** *Genome Biology*, vol. 12, R112 [0010]
 - **MISSIRIAN et al.** *BMC Bioinformatics*, vol. 12, 287 [0010]
 - **TSAI et al.** *Plant Physiology*, vol. 156, 1257-1268 [0027]
 - **RODRIGUE et al.** *PLoS One*, vol. 4, 34761 [0041]
 - **ZHANG et al.** *Bioinformatics*, vol. 30 (5), 614-620 [0041]
 - **LIU et al.** *Bioinformatics*, vol. 28 (22), 2870-2874 [0041]
 - **MASELLA et al.** *BMC Bioinformatics*, vol. 13, 31 [0041]
 - **ERALI ; WITTER.** *Methods*, vol. 50 (4), 250-261 [0051]
 - **DE KOEYER et al.** *Molecular Breeding*, vol. 25, 67-90 [0052]
 - **TSAI et al.** *Plant Physiology*, vol. 156 (3), 1257-1268 [0063]
 - **DE KOEYER, D.L. ; DOUGLASS, K. ; MURPHY, A.M. ; WHITNEY, S. ; NOLAN, L. ; SONG, Y. ; DE JONG, W.S.** Application of high-resolution DNA melting for genotyping and variant scanning of diploid and autotetraploid potato. *Molecular Breeding*, 2010, vol. 25 (1), 67-90 [0065]
 - **ERALI M ; WITTWER CT.** High Resolution Melting Analysis for Gene Scanning. *Methods*, 2010, vol. 50 (4), 250-261 [0065]
 - **NORDSTRÖM KJ ; ALBANI MC ; JAMES GV ; FITZPATRICK TB ; BASSET GJ ; BOREL P ; CARRARI F ; DELLAPENNA D ; FRASER PD ; HELLMANN H ; OSORIO S ; ROTHAN C ; VALPUESTA V.** Vitamin deficiencies in humans: can plant science help?. *Plant Cell*, February 2012, vol. 24 (2), 395-414 [0065]
 - **FORSTER, B.P.** Mutation genetics of salt tolerance in barley: A assessment of Golden Promise and other semi-dwarf mutants. *Euphytica*, 2001, vol. 120 (3), 317-328 [0065]
 - **GADY AL ; HERMANS FW ; VAN DE WAL MH ; VAN LOO EN ; VISSER RG ; BACHEM CW.** Implementation of two high through-put techniques in a novel application: detecting point mutations in large EMS mutated plant populations. *Plant Methods*, 07 October 2009, vol. 5, 13 [0065]
 - **KOORNNEEF, M. ; DELLAERT, L.W. ; VAN DER VEEN, J.H.** EMS-and radiation-induced mutation frequencies at individual loci in *Arabidopsis thaliana* (L) Heynh. *Mutat Res*, 1982, vol. 93, 109-123 [0065]
 - **LIU B et al.** COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics*, 2012, vol. 28, 2870-2874 [0065]
 - **MARÇAIS G ; KINGSFORD C.** A FAST. lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 15 March 2011, vol. 27 (6), 764-70 [0065]
 - **MASELLA AP et al.** PANDAseq: paired-end assembler for Illumina sequences. *BMC Bioinformatics*, 2012, vol. 13, 31 [0065]
 - **MCCALLUM, C. ; COMAI, L. ; GREENE, E.A. ; HENIKOF, S.** Targeting induced local lesions in genomes (TILLING) for plant functional genomics. *Plant Physiol*, 2000, vol. 123 (2), 439-442 [0065]
 - **MINOCHE, A.E. ; DOHM, J.C. ; HIMMELBAUER, H.** Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*, 2011, vol. 12, R112 [0065]
 - **MISSIRIAN V ; COMAI L ; FILKOV V.** Statistical mutation calling from sequenced overlapping DNA pools in TILLING experiments. *BMC Bioinformatics*, 14 July 2011, vol. 12, 287 [0065]
 - **SARGENT RG ; KIM S ; GRUENERT DC.** Oligo/polynucleotide-based gene modification: strategies and therapeutic potential. *Oligonucleotides*, March 2011, vol. 21 (2), 55-75

; **GUTJAHR C ; HARTWIG B ; TURCK F ; PASZKOWSKI U ; COUPLAND G ; SCHNEEBERGER K.** Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nat Biotechnol.*, April 2013, vol. 31 (4), 325-30 [0065]
• High-Throughput Detection of Induced Mutations and Natural Variation Using KeyPoint™ Technology.
RIGOLA D ; VAN OEVEREN J ; JANSSEN A et al. PLoS ONE. 2009, vol. 4, e4761 [0065]
• **RODRIGUE S ; MATERNA AC ; TIMBERLAKE SC ; BLACKBURN MC ; MALMSTROM RR ; ALM EJ ; CHISHOLM SW.** Unlocking short read sequencing for metagenomics. *PLoS One*, 28 July 2010, vol. 5 (7), e11840 [0065]

[0065]
• **STADLER, L.J.** On the genetic nature of induced mutations in plants. *Proc VI Congress of Genetics*, 1932, vol. 1, 274-294, www.agris.fao.org [0065]
• **TSAI, H.; HOWELL, T. ; NITCHER, R. ; MISSIRIAN, V. ; WATSON, B. ; NGO, K.J. ; LIEBERMAN, M. ; FASS, J. ; UAUY, C. ; TRAN, R.K.** Discovery of rare mutations in populations: TILLING by sequencing. *Plant Physiology*, 2011, vol. 156, 1257-1268 [0065]
• **WANG, T. L. ; UAUY, C. ; ROBSON, F. ; TILL, B.** TILLING in extremis. *Plant Biotechnology Journal*, 2012, vol. 10, 761-772 [0065]
• **ZHANG J ; KOBERT K ; FLOURI T ; STAMATAKIS A.** PEAR: a fast and accurate Illumina Paired-End read mergeR. *Bioinformatics*, 2014, vol. 30 (5), 614-620 [0065]