(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 403 312**

(51) Int. CI.:

C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **13.01.2010 E 10701404 (5)**

(97) Fecha y número de publicación de la concesión europea: **20.03.2013 EP 2379751**

(54) Título: **Nuevas estrategias para la secuenciación del genoma**

(30) Prioridad:

13.01.2009 US 144281 P
17.07.2009 US 226468 P
17.07.2009 NL 2003235

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
17.05.2013

(73) Titular/es:

KEYGENE N.V. (100.0%)
P.O. Box 216
6700 AE Wageningen, NL

(72) Inventor/es:

VAN EIJK, MICHAEL JOSEPHUS THERESIA;
VAN TUNEN, ADRIANUS JOHANNES y
JANSSEN, ANTOINE ANTONIUS ARNOLDUS WILHELMUS

(74) Agente/Representante:

PONTI SALES, Adelaida

DESCRIPCIÓN

Nuevas estrategias para la secuenciación del genoma

Campo técnico de la invención

[0001] La presente invención se refiere a un procedimiento eficaz para la secuenciación del genoma completo de novo. La invención se refiere a la secuenciación de ácidos nucleicos a gran escala y en particular a procedimientos para secuenciar el genoma, o una parte del mismo, de un organismo. La invención se refiere a estrategias mejoradas para determinar la secuencia de, preferiblemente, genomas complejos (es decir grandes) basándose en el uso de tecnologías de secuenciación de alta productividad.

Antecedentes de la invención

[0002] El objetivo de muchos proyectos de secuenciación es determinar, por primera vez, la secuencia del genoma completo de un organismo objetivo (secuenciación del genoma borrador de novo). El tener una secuencia genómica borrador a mano permite la identificación de información genética útil de un organismo, por ejemplo, para identificar el origen de la variedad genética entre especies o individuos de la misma especie. Por lo tanto, hay un deseo general en la materia de llegar a técnicas que permitan la determinación de novo de la secuencia genómica entera de un individuo sea ser humano, animal o planta, con un coste y esfuerzo razonables. Esta búsqueda normalmente se indica como la búsqueda del genoma por 1000 dólares, es decir, determinar la secuencia genómica entera de un individuo por un máximo de 1000 dólares (sin tener en cuenta las fluctuaciones de la moneda). Sin embargo, en la práctica el genoma por 1000 dólares no se basa necesariamente en la secuenciación genómica de novo y estrategia de ensamblado, sino que se puede basar también en un procedimiento de resecuenciación. En este último caso, el genoma resecuenciado no será ensamblado de novo, sino que su ADN secuenciado será comparado con (cartografiado en) una secuencia genómica de referencia existente para el organismo de interés. Por lo tanto, un procedimiento de resecuenciación es técnicamente menos desafiante y menos costoso. Por motivos de claridad, el foco de la presente invención está en las estrategias de secuenciación genómica de novo, capaces de ser aplicadas a organismos para los que se carece de una secuencia genómica de referencia.

[0003] Los esfuerzos actuales están logrando resultados diversos, abundantes y que aumentan rápidamente. No obstante, el objetivo todavía no se ha logrado. Todavía no es económicamente factible secuenciar y ensamblar un genoma completo de una forma directa. Todavía siguen siendo necesarias en la materia estrategias de secuenciación genómica de novo mejoradas.

[0004] El documento WO03/027311 describe un procedimiento de secuenciación genómica aleatoria (*shotgun*) de mezcla de matrices de clones (CAPPS). El procedimiento usa lecturas de secuencia aleatorias de diferentes clones mezclados (BAC). Basándose en el ensamblado cruzado de las lecturas aleatorias se puede generar un cóntigo de secuencias a partir de una pluralidad de clones y se puede generar un mapa de los clones con respecto a la secuencia. La publicación describe, con más detalle, la generación de una biblioteca de BAC en mezclas multidimensionales, por ejemplo, un formato bidimensional en el que cada mezcla y fila contiene 148 clones de BAC (formato 148 x 148). Usando CAPPS, las mezclas de BAC se secuencian con una cobertura de 4-5X como media, lo que genera una cobertura de 8-10X por BAC, en el caso del esquema de mezcla bidimensional. Los cóntigos se hacen por BAC por separado basándose en las secuencias que son únicas para el BAC basándose en su aparición en una sola fila y una sola mezcla en el caso de un esquema de mezcla bidimensional. Posteriormente, estos BAC se ensamblan en un cóntigo para el genoma. La publicación demuestra la tecnología basada en 5 BAC solo. La publicación deja sin tocar el problema del procesamiento de datos. Sin embargo, una de las desventajas de esta tecnología es que el uso de fragmentos compartidos aleatoriamente requiere una enorme cantidad de lecturas para cubrir un genoma con un nivel de redundancia de secuencia de 8 a 10 veces, haciendo que este procedimiento sea muy laborioso a gran escala. Además, no da una secuencia basada en el mapa físico de BAC.

[0005] El documento US2007/0082358 describe un procedimiento de ensamblado de novo de información de secuencia basado en una biblioteca aislada y amplificada clonalmente de ADN genómico monocatenario para crear la información de secuencia aleatoria del genoma completo combinado con el mapa de restricción óptico del genoma completo usando una enzima de restricción para la creación de un mapa de restricción ordenado.

[0006] El documento US2002/0182830 describe un procedimiento de cartografía de cóntigos de BAC por comparación de subsecuencias. El procedimiento se dirige a evitar las dificultades asociadas con secuencias repetitivas y la generación de cóntigos por creación de puentes entre regiones ricas en repeticiones.

[0007] La determinación de mapas físicos basados en BAC se puede basar en la secuenciación de bibliotecas de BAC (cartografía física basada en secuencias de clones de BAC) usando por ejemplo, el procedimiento descrito en el documento WO2008/00795 de Keygene indicado también como "perfilado genómico completo" o WPG. Brevemente, el WPG se refiere a la generación de un mapa físico de al menos parte de un genoma, que comprende las etapas de generar una biblioteca de cromosomas artificiales a partir de una muestra de ADN, mezclar los clones, digerir los clones mezclados con enzimas de restricción, ligar adaptadores que contienen identificador, amplificar los fragmentos de restricción ligados a adaptador que contiene identificador, correlacionar los amplicones con los clones y ordenar los fragmentos para generar un cóntigo para crear así un mapa físico.

[0008] El documento WO2008/007951 describe un procedimiento para determinar mapas físicos basados en clones de BAC usando la metodología de WGS.

[0009] El Consorcio de secuenciación del genoma del erizo de mar (*Science* (2006) Vol. 314:941-952) describe un procedimiento para secuenciar el genoma completo del erizo de mar usando dos estrategias diferentes, CAPSS y WGS, que se combinan para ensamblar el genoma del erizo de mar.

[0010] A pesar de todos los desarrollos en la secuenciación de alta productividad, la determinación de secuencias genómicas borrador con gran precisión, todavía se considera caro y laborioso. Sigue siendo necesario complementar los procedimientos que existen actualmente para llegar a procedimientos eficaces y económicos para generar secuencias genómicas borrador. En particular, las tecnologías de secuenciación de alta productividad actuales proporcionan lecturas relativamente cortas (hasta 400 nt), dando como resultado cóntigos relativamente cortos que son difíciles de ensamblar en cóntigos más largos y requieren una gran demanda de potencia computacional.

Resumen de la invención

[0011] Los autores de la presente invención han encontrado que la combinación del perfilado genómico completo basado en clones con la secuenciación (de alta productividad) de fragmentos de muestra de ADN (genómico) usando tecnologías de secuenciación de alta productividad, proporciona una estrategia superior para determinar secuencias genómicas borrador con alta precisión y velocidad. Mediante la generación de cóntigos a partir de las lecturas de secuencia y anclaje de estas lecturas al cóntigo de BAC (o YAC o cualquier otro vector de clonación de inserto grande) obtenido por el perfilado genómico completo, se generan cóntigos de longitud y densidad mayores. Por lo tanto, se obtiene una secuencia genómica borrador que es generada por un número reducido de cóntigos, aumentado así la calidad de la misma.

Definiciones

[0012] Agrupación: con el término "agrupación" se entiende la comparación de dos o más secuencias de nucleótidos basada en la presencia de tramos cortos o largos de nucleótidos idénticos o similares y la agrupación de las secuencias con un determinado nivel mínimo de homología de secuencia basado en la presencia de tramos cortos (o más largos) de secuencias idénticas o similares.

[0013] Alineamiento: colocación de múltiples secuencias en una presentación tabular para maximizar la posibilidad de obtener regiones de identidad de secuencia a lo largo de varias secuencias en el alineamiento, p. ej., mediante la introducción de huecos. Se conocen en la materia varios procedimientos de alineamiento de secuencias de nucleótidos, como se explicará con más detalle a continuación.

[0014] AFLP: AFLP se refiere a un procedimiento para la amplificación selectiva de ácidos nucleicos basada en digerir un ácido nucleico con una o más endonucleasas de restricción para dar fragmentos de restricción, ligar adaptadores a los fragmentos de restricción y amplificar los fragmentos de restricción ligados a adaptador con al menos un cebador que es complementario (parte) del adaptador, complementario (parte) del resto de la endonucleasa de restricción, y que puede contener además al menos un nucleótido seleccionado aleatoriamente entre A, C, T o G (o U según sea el caso) en el extremo 3' del cebador. El AFLP no requiere ninguna información de secuencia previa y se puede realizar con cualquier ADN de partida. En general, el AFLP comprende las etapas de:

(a) digerir un ácido nucleico, en particular un ADN o ADNc, con una o más endonucleasas de restricción específicas, para fragmentar el ADN en una serie correspondiente de fragmentos de restricción;
(b) ligar los fragmentos de restricción así obtenidos con un adaptador oligonucleótido sintético bicatenario, uno

de cuyos extremos es compatible con uno o ambos extremos de los fragmentos de restricción, para así producir fragmentos de restricción ligados a adaptador del ADN de partida;
(c) poner en contacto los fragmentos de restricción ligados a adaptador en condiciones de hibridación, con uno o más cebadores oligonucleótidos que son dirigidos hacia los adaptadores y que pueden contener nucleótidos selectivos en su extremo 3';
(d) amplificar el fragmento de restricción ligado a adaptador hibridado con los cebadores por la PCR o una técnica similar para así producir extensión adicional de los cebadores hibridados junto con los fragmentos de restricción del ADN de partida con los que han hibridado los cebadores; y
(e) detectar, identificar o recuperar el fragmento de ADN amplificado o extendido así obtenido.

[0015] El AFLP proporciona por lo tanto, un subconjunto reproducible de fragmentos ligados a adaptador. El AFLP se describe en los documentos EP 534858, US 6045994 y en Vos y col. 1995, "AFLP: a new technique for DNA fingerprinting". *Nucleic Acids Research* 23(21):4407-4414. Se hace referencia a esta publicación para más detalles relacionados con el AFLP. El AFLP se usa habitualmente como una técnica de reducción de la complejidad eficaz, fuerte y reproducible.

[0016] Base selectiva o nucleótido selectivo: Localizado en el extremo 3' del cebador que contiene una parte que es complementaria del adaptador y una parte que es complementaria del resto del sitio de restricción, la base selectiva se selecciona aleatoriamente entre A, C, T o G (o U según sea el caso). Mediante la extensión de un cebador con una base selectiva, la posterior amplificación dará solo un subconjunto reproducible de fragmentos de restricción ligados a adaptador, es decir, solo los fragmentos que pueden ser amplificados usando el cebador que lleva la base selectiva. Los nucleótidos selectivos se pueden añadir al extremo 3' del cebador en un número que varía entre 1 y 10. Típicamente, 1-4 es suficiente. Ambos cebadores pueden contener un número distinto de bases selectivas. Con cada base selectiva añadida, el subconjunto reduce la cantidad de fragmentos de restricción ligados a adaptador amplificados en el subconjunto en un factor de aproximadamente 4. Típicamente, el número de bases selectivas usadas en el AFLP está indicado por +N+M, en donde un cebador lleva N nucleótidos selectivos y el otro cebador lleva M nucleótidos selectivos. Por lo tanto, un AFLP EcoRI/MseI +1/+2 es la forma abreviada de la digestión del ADN de partida con EcoRI y MseI, ligamiento de adaptadores adecuados y amplificación con un cebador dirigido a la posición de restricción de EcoRI que lleva una base selectiva y el otro cebador dirigido al sitio de restricción de MseI que lleva 2 nucleótidos selectivos. Un cebador usado en el AFLP que lleva al menos un nucleótido selectivo en su extremo 3' también se describe como un cebador-AFLP. Los cebadores que no llevan un nucleótido selectivo en su extremo 3' y que de hecho son complementarios del adaptador y del resto del sitio de restricción se denominan a veces cebadores AFLP+0. La expresión nucleótido selectivo también se usa para nucleótidos de la secuencia diana que están situados adyacentes a la sección del adaptador y que se han identificado por el uso de cebador selectivo como consecuencia de lo cual, el nucleótido se ha vuelto conocido.

[0017] Secuenciación: El término secuenciación se refiere a determinar el orden de nucleótidos (secuencias de bases) en una muestra de ácido nucleico, p. ej. ADN o ARN. Hay muchas técnicas disponibles tales como la secuenciación de Sanger y tecnologías de secuenciación de alta productividad (también conocidas como tecnologías de secuenciación de la siguiente generación) tales como la plataforma GS FLX ofrecida por Roche Applied Science, basada en pirosecuenciación.

[0018] Endonucleasa de restricción: una endonucleasa de restricción o enzima de restricción es una enzima que reconoce una secuencia de nucleótidos específica (sitio diana) en una molécula de ADN bicatenaria, y que escindirá ambas cadenas de la molécula de ADN en o cerca de cada sitio diana, dejando un extremo romo o uno decalado.

[0019] Cortadoras frecuentes y cortadoras infrecuentes: Las enzimas de restricción típicamente tienen secuencias de reconocimiento que varían en el número de nucleótidos de 3, 4 (tal como MseI) a 6 (EcoRI) e incluso 8 (NotI). Las enzimas de restricción usadas pueden ser cortadoras frecuentes e infrecuentes. El término "frecuente" a este respecto se usa típicamente en relación con el término "infrecuente". Las endonucleasas cortadoras frecuentes (también llamadas cortadoras frecuentes) son endonucleasas de restricción que tienen una secuencia de reconocimiento relativamente corta. Las cortadoras frecuentes típicamente tienen 3-5 nucleótidos que reconocen y posteriormente cortan. Por lo tanto, una cortadora frecuente corta como media una secuencia de ADN cada 64-1024 nucleótidos. Las cortadoras infrecuentes son endonucleasas de restricción que tienen una secuencia de reconocimiento relativamente larga. Las cortadoras infrecuentes típicamente tienen 6 o más nucleótidos que reconocen y posteriormente cortan. Por lo tanto, una cortadora 6 infrecuente corta como media una secuencia de ADN cada 4096 nucleótidos, conduciendo a fragmentos más largos. Se observa otra vez que la definición de frecuente e infrecuente es en comparación de uno con otro, que significa que cuando se usa una enzima de restricción de 4 pb, tal como MseI, en combinación con una cortadora 5 tal como AvaII, AvaII se ve como la

cortadora infrecuente y MseI como la cortadora frecuente.

[0020] Fragmentos de restricción: las moléculas de ADN producidas por digestión con un 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 particular en un conjunto discreto de fragmentos de restricción. Los fragmentos de ADN que resultan de la escisión por la endonucleasa de restricción se pueden usar después en una variedad de técnicas y se pueden detectar, por ejemplo, por electroforesis en gel.

[0021] Ligamiento: la reacción enzimática catalizada por una enzima ligasa en la que dos moléculas de ADN bicatenarias se unen covalentemente entre sí se denomina ligamiento. En general, ambas cadenas de ADN son unidas covalentemente entre sí, pero también se puede prevenir el ligamiento de una de las dos cadenas mediante modificación química o enzimática de uno de los extremos de las cadenas. En este caso, la unión covalente se producirá solo en una de las dos cadenas de ADN.

[0022] Oligonucleótido sintético: las moléculas de ADN monocatenarias que tienen preferiblemente 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 tener una secuencia de nucleótidos única o deseada, aunque se pueden sintetizar familias de moléculas que tienen secuencias relacionadas y que tienen diferentes composiciones de nucleótidos en posiciones específicas dentro de la secuencia de nucleótidos. La expresión oligonucleótido sintético se usará para referirse a moléculas de ADN que tienen una secuencia de nucleótidos diseñada o deseada.

[0023] Adaptadores: moléculas cortas de ADN bicatenario 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 modo que se pueden ligar a los extremos de los fragmentos de restricción. Los adaptadores en general están compuestos de dos oligonucleótidos sintéticos que tienen secuencias de nucleótidos que son parcialmente complementarias una de otra. Cuando se mezclan los dos oligonucleótidos sintéticos en disolución en condiciones adecuadas, se asociarán entre sí formando una estructura bicatenaria. Después de la asociación, un extremo de la molécula adaptadora se diseña de modo que sea compatible con el extremo de un fragmento de restricción y se pueda ligar al mismo; el otro extremo del adaptador se pude diseñar de modo que no se pueda ligar, pero no es necesario que sea este el caso (adaptadores ligados dobles).

[0024] Fragmentos de restricción ligados a adaptador: fragmentos de restricción que se han rematado con adaptadores.

[0025] Cebadores: en general, el término cebadores se refiere a cadenas de ADN que pueden cebar la síntesis de ADN. La ADN polimerasa no puede sintetizar ADN nuevo sin cebadores: solo puede extender una cadena de ADN existente en una reacción en la que se usa la cadena complementaria como molde para dirigir el orden de los nucleótidos que se van a ensamblar. Se hará referencia a las moléculas de oligonucleótidos sintéticos que se usan en una reacción en cadena de la polimerasa (PCR) como cebadores.

[0026] Amplificación del ADN: la expresión amplificación del ADN se usará típicamente para indicar la síntesis in vitro de moléculas de ADN bicatenarias usando la PCR. Hay que indicar que existen otros procedimientos de amplificación y se pueden usar en la presente invención sin salirse de la esencia.

[0027] Ácido nucleico: un ácido nucleico de acuerdo con la presente invención puede incluir cualquier polímero u oligómero de bases púricas y pirimidínicas, preferiblemente citosina, timina y uracilo, y adenina y guanina, respectivamente (véase Albert L. Lehninger, Principles of Biochemistry, en 793-800 (Worth Pub. 1982)). La presente invención contempla cualquier desoxirribonucleótido, ribonucleótido o componente de ácido nucleico peptídico, y cualesquiera variantes de los mismos, tal como formas metiladas, hidroximetiladas o glicosiladas de estas bases, y similares. Los polímeros u oligómeros pueden ser de composición homogénea o heterogénea, y se pueden aislar de fuentes naturales o se pueden producir de forma artificial o sintética. Además, los ácidos nucleicos pueden ser ADN o ARN, o una mezcla de los mismos, y pueden existir de forma permanente o transitoria en forma monocatenaria o bicatenaria, incluyendo estados de homodúplex, heterodúplex e híbrido.

[0028] Reducción de la complejidad: la expresión reducción de la complejidad se usa para indicar un procedimiento en el que se reduce la complejidad de una muestra de ácido nucleico, tal como ADN genómico, por la generación o selección de un subconjunto de la muestra. Este subconjunto puede ser representativo de la muestra entera (es decir, complejo) y preferiblemente es un subconjunto reproducible. En este contexto, reproducible

significa que cuando la misma muestra se reduce en complejidad usando el mismo procedimiento y condiciones experimentales, se obtiene el mismo subconjunto, o al menos comparable. El procedimiento usado para reducir la complejidad puede ser cualquier procedimiento conocido en la materia para reducir la complejidad. Los ejemplos de procedimientos para reducir la complejidad incluyen, por ejemplo, AFLP® (Keygene N.V., the Netherlands; véase. p. ej. el documento EP 0534858), los procedimientos descritos por Dong (véase, p. ej., los documentos WO 03/012118, WO 00/24939), indexado (Unrau et al., véase más adelante), etc. 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 la misma muestra se reduce en complejidad de la misma forma, se obtiene el mismo subconjunto de la muestra, en contraposición a la reducción de la complejidad más aleatoria tal como la microdisección, cizalladura aleatoria, o el uso de ARNm (ADNc) que representa una parte del genoma transcrito en un tejido seleccionado y para su reproducibilidad depende de la selección del tejido, tiempo de aislamiento, etc.

[0029] Marcaje: el término marcaje se refiere a la adición de un marcador de secuencia a una muestra de ácido nucleico con el fin de poder distinguirla de una segunda o posterior muestra de ácido nucleico. El marcaje se puede realizar, p. ej., por adición de un identificador de secuencia durante la reducción de la complejidad, o por cualquier otro medio conocido en la materia tal como una etapa de ligamiento separada. Dicho identificador de secuencia puede ser, p. ej., una secuencia de bases única de longitud variable pero definida, usada unívocamente para identificar una muestra de ácido nucleico específica. Los ejemplos típicos son secuencias ZIP, conocidas en la materia como marcadores usados habitualmente para la detección única por hibridación (lannone y col. *Cytometry* 39:131-140, 2000). Usando marcadores basados en nucleótidos, se puede determinar el origen de una muestra, un clon o un producto amplificado tras procesamiento adicional. En el caso de combinar productos procesados procedentes de diferentes muestras de ácido nucleico, las muestras de ácido nucleico diferentes deben identificarse usando marcadores diferentes.

[0030] Identificador: una secuencia corta que se puede añadir a un adaptador o un cebador o incluir en su secuencia o usar de otra forma como marcador para proporcionar un identificador único (también conocido como código de barras o índice). Dicho identificar de secuencia (marcador) puede ser una secuencia de bases única de longitud variable pero definida, típicamente de 4-16 pb, usada para identificar una muestra de ácido nucleico específica. Por ejemplo, los marcadores de 4 pb permiten 4(exp4) = 256 marcadores diferentes. Usando dicho identificador, se puede determinar el origen de una muestra de la PCR tras procesamiento adicional o se pueden relacionar fragmentos con un clon. También se pueden distinguir entre sí los clones de una mezcla usando estos identificadores basados en secuencia. Por lo tanto, los identificadores pueden ser específicos de la muestra, específicos de la mezcla, específicos del clon, específicos del amplicón, etc. En el caso de combinar productos procesados procedentes de diferentes muestras de ácido nucleico, las diferentes muestras de ácido nucleico en general se identifican usando diferentes identificadores. Los identificadores preferiblemente difieren entre sí en al menos dos pares de bases y preferiblemente no contienen dos bases consecutivas idénticas para prevenir lecturas erróneas. La función del identificador a veces se puede combinar con otras funcionalidades tales como adaptadores o cebadores y pueden estar situados en cualquier posición conveniente.

[0031] Biblioteca marcada: la expresión biblioteca marcada se refiere a una biblioteca de ácido nucleicos marcados.

[0032] Alinear y alineamiento: con los términos "alinear" y "alineamiento" se entiende la comparación de dos o más secuencias de nucleótidos basada en la presencia de tramos cortos o largos de nucleótidos idénticos o similares. Se conocen en la materia varios procedimientos para el alineamiento de secuencias de nucleótidos como se explicará mejor más adelante.

[0033] El término "cóntigo" se usa en relación con el análisis de secuencia del ADN, y se refiere a tramos contiguos ensamblados de ADN derivados de dos o más fragmentos de ADN que tienen secuencias de nucleótidos contiguas. Por lo tanto, un cóntigo es un conjunto de fragmentos de ADN que solapan, que proporciona una secuencia contigua parcial de un genoma. Un "andamiaje" se define como una serie de cóntigos que están en el orden correcto, pero no están conectados en una secuencia continua, es decir, contienen huecos. Los mapas de cóntigos también representan la estructura de regiones contiguas de un genoma especificando las relaciones de superposición entre un conjunto de clones. Por ejemplo, el término "cóntigos" abarca una serie de vectores de clonación que están ordenados de forma que tienen cada solapamiento de secuencia con la de sus vecinos. Después los clones conectados se pueden agrupar en cóntigos, sea de forma manual o preferiblemente usando programas de ordenador adecuados, tales como FPC, PHRAP, CAP3, etc.

[0034] El término "andamiaje" se usa para cóntigos generados, entre otros, por secuenciación de extremos

apareados que contienen huecos de tamaño (des)conocido. El término "superandamiaje" se usa para andamiajes que están conectados entre sí por cóntigos de BAC de WGP.

[0035] Cribado de alta productividad: el cribado de alta productividad, abreviado a menudo HTS, es un procedimiento para la experimentación científica, en especial importante para los campos de la biología y la química. Mediante una combinación de robótica moderna y otro hardware de laboratorio especializado, un investigador pueden cribar eficazmente grandes cantidades de muestras simultáneamente.

[0036] En la dirección 5' o en la dirección 3': un convenio usado para describir características de una secuencia de ADN en términos de la dirección (de 5' a 3') de la secuencia de ADN. En la dirección 3' es en la dirección del extremo 3' de la secuencia de ADN, mientras que en la dirección 5' es en la dirección del extremo 5' de la secuencia de ADN. Convencionalmente, las secuencias de ADN monocatenarias, mapas genéticos y secuencias de ARN se dibujan con transcripción (o traducción) de izquierda a derecha y así en la dirección 3' es hacia la derecha (y en la dirección 5' hacia la izquierda). La expresión en la dirección 3' o en la dirección 5' se puede usar para definir posiciones relativas entre sí de diferentes segmentos de ADN en una secuencia de ADN. Por ejemplo, en un fragmento de AFLP, el nucleótido selectivo en el fragmento está situado en la dirección 5' desde el adaptador, pero el nucleótido selectivo en el cebador está situado en la dirección 3' de la sección complementaria-adaptador del cebador.

Descripción de los dibujos

[0037]

La figura 1 es una representación visual que combina el perfilado genómico completo y la secuenciación del genoma completo usando secuencias derivadas de BAC y secuenciación aleatoria para generar cóntigos y andamiajes.
La figura 2 es una representación visual que combina el perfilado genómico completo y la secuenciación del genoma completo usando secuencias derivadas de BAC y secuenciación aleatoria para complementar cóntigos derivados de BAC y para llenar huecos entre cóntigos de BAC.
La figura 3 es una visualización de la distribución de tamaños de cóntigos obtenida en la generación de cóntigos derivados de BAC para el melón.
La figura 4 es una representación visual de la estructura de cebadores y su interacción con el adaptador y el identificador.
La figura 5 es una representación visual de la generación del andamiaje. Los bloques son cóntigos de BAC, las líneas horizontales son andamiajes de WGS y las líneas verticales son marcadores unidos.
La figura 6 es una representación visual de la generación de andamiaje ramificado. Los bloques son los cóntigos de BAC, las líneas horizontales son andamiajes de WGS y las líneas verticales son marcadores conectados. La línea horizontal de puntos muestra otro andamiaje de WGS conectado al mismo cóntigo de BAC, que da como resultado dos ramas.

Descripción detallada de la invención

[0038] Los autores de la presente invención han encontrado una nueva estrategia de secuenciación del genoma (de planta) y la han aplicado a una cosecha de plantas comerciales (melón). Esta estrategia de secuenciación del genoma se basa en dos componentes:

1) Construcción de un mapa físico basado en secuencias, preferiblemente por secuenciación de los extremos de fragmentos de clones de cromosomas artificiales (preferiblemente BAC) mezclados (Amplicon Express, Pullman, EE.UU.), preferiblemente usando el Genome Analyzer II y
2) Secuenciación del genoma completo (WGS), que preferiblemente comprende una combinación de lecturas simples, lecturas de extremos apareados de 3 kb y lecturas de extremos apareados de salto largo usando GS FLX Titanium o GA II.

[0039] La máxima potencia de ensamblado se obtiene cuando tanto el mapa físico basado en secuencias como las secuencias de WGS se generan usando la misma línea (homozigoto/consanguínea), como era el caso para la cosecha descrita en los ejemplos adjuntos.

[0040] Por lo tanto, en un primer aspecto, la invención se refiere a un procedimiento para determinar una secuencia genómica que comprende las etapas de:

- proporcionar un mapa físico de una muestra de genoma por secuenciación de los extremos de fragmentos de clones de BAC mezclados;
- proporcionar un conjunto de lecturas de secuencia de la muestra de ADN;
- generar un cóntigo del mapa físico y las lecturas de secuencia.

[0041] De esta forma se puede obtener un borrador de la secuencia genómica eficaz y de alta calidad ya que las lecturas de secuencia complementan el andamiaje proporcionado por el mapa físico obtenido por el cóntigo de los extremos de fragmentos secuenciados de los clones.

[0042] La invención en una realización se refiere a un procedimiento para determinar una secuencia genómica que comprende las etapas de:

(a) proporcionar una muestra de ADN;
(b) generar un banco de clones de cromosomas artificiales (p. ej., BAC, YAC) en el que cada clon de cromosoma artificial contiene parte de la muestra de ADN;
(c) combinar los clones de cromosomas artificiales en una o más mezclas, en las que cada clon está presente en más de una mezcla, para crear una biblioteca;
(d) proporcionar un conjunto de fragmentos para cada mezcla;
(e) ligar adaptadores a uno o ambos lados de los fragmentos,
(f) determinar la secuencia de al menos parte del adaptador y parte del fragmento;
(g) asignar las secuencias de fragmentos a los correspondientes clones;
(h) construir un cóntigo de clones generando así un mapa físico de la muestra de genoma;
(i) generar lecturas de secuencia de una muestra de ADN;
(j) alinear las lecturas de secuencia y/o cóntigos o andamiajes desde las lecturas de secuencia al cóntigo de clones construyendo así una secuencia genómica/superandamiaje.

[0043] Esta estrategia combina la potencia de la cartografía física basada en BAC con la secuenciación del genoma completo. El procedimiento de acuerdo con la invención proporciona reducciones de coste significativas comparado con las estrategias de secuenciación del genoma usadas actualmente. El procedimiento proporciona además una mayor flexibilidad para combinar información de secuencias derivada de cromosomas artificiales tales como secuencias derivadas de BAC e información de secuencias derivada de técnicas que generan directamente la información de secuencia tal como la secuenciación aleatoria del genoma completo y técnicas similares. El presente procedimiento también se puede complementar con otra información de secuencia disponible tal como la obtenida por una técnica más convencional como la didesoxisecuenciación de Sanger.

[0044] En la etapa (a) del procedimiento se proporciona una muestra de ADN. Esta se puede obtener por cualquier medio de la técnica tal como los descritos, por ejemplo, por Sambrook y col. (Sambrook y Russell (2001) "Molecular Cloning: A Laboratory Manual" (3ª edición), Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press). La muestra de ADN puede ser de cualquier especie, en particular de origen humano, vegetal o animal. Se puede usar solo una parte de un genoma, pero esto no es necesario ya que la presente invención también proporciona procedimientos para acomodar genomas de cualquier tamaño, por ejemplo, por creación de subconjuntos reproducibles a través de la reducción de la complejidad reproducible, tal como por ejemplo, amplificación selectiva basada en AFLP (documento EP534858). Por lo tanto, típicamente, el presente procedimiento usa el genoma completo.

[0045] En la etapa (b) se genera un banco de clones artificiales. La biblioteca puede ser una biblioteca de cromosomas artificiales bacterianos (BAC) o basada en levaduras (YAC). También son posibles otras bibliotecas tales como basadas en fósmidos, cósmidos, PAC, TAC o MAC. Se prefiere una biblioteca de BAC. La biblioteca preferiblemente es una genoteca de alta calidad y preferiblemente es una genoteca de tamaño de inserto grande. Esto significa que el BAC individual contiene un inserto relativamente grande del ADN genómico que se investiga (típicamente > 125 kpb). El tamaño del inserto grande preferido depende de la especie. A lo largo de esta solicitud, se puede hacer referencia a los BAC como ejemplos de cromosomas artificiales. Sin embargo, hay que indicar que la presente invención no está limitada a los mismos, y que se pueden usar otros cromosomas artificiales sin salirse del espíritu de la invención. Preferiblemente las bibliotecas contienen al menos 5 equivalentes genómicos, más preferiblemente al menos 7, los más preferiblemente al menos 8. Se prefiere en particular, al menos 10. Cuanto mayor es el número de equivalentes genómicos en la biblioteca, más fiables serán los cóntigos y el mapa físico resultantes.

[0046] Los clones individuales en la biblioteca se mezclan para formar mezclas que contienen una multitud de cromosomas artificiales o clones. La mezcla puede ser la simple combinación de una serie de clones individuales en una muestra (por ejemplo, 100 clones en 10 mezclas, que contiene cada una 10 clones), pero también se pueden usar estrategias de mezcla más elaboradas. La distribución de los clones en las mezclas preferiblemente es tal que cada clon está presente en al menos una o dos o más mezclas, generando así una biblioteca. Preferiblemente, las mezclas contienen de 10 a 10000 clones por mezcla, preferiblemente de 100 a 1000, más preferiblemente de 250 a 750. Se observa que el número de clones por mezcla puede variar ampliamente, y esta variación está relacionada, por ejemplo, con el tamaño del genoma que se investiga. Típicamente, el tamaño máximo de una mezcla o submezcla está controlada por la capacidad de identificar unívocamente un clon en una mezcla mediante un conjunto de identificadores. Un intervalo típico para un equivalente genómico en una mezcla es del orden de 0,2-0,3, y esto puede variar de nuevo por genomas. Las mezclas se generan basándose en estrategias de mezclado bien conocidas en la materia. El experto en la materia puede seleccionar la estrategia de mezclado óptima basándose en factores tales como el tamaño genómico, etc. La estrategia de mezclado resultante dependerá de las circunstancias, y son ejemplos de la misma el mezclado en placa, mezclado N-dimensional tal como mezclado 2D, mezclado 3D, mezclado 6D o mezclado complejo. Para facilitar la manipulación de números grandes de mezclas, las mezclas a su vez se pueden combinar en supermezclas (es decir, las supermezclas son mezclas de mezclas de clones) o dividirlas en submezclas. Otros ejemplos de estrategias de mezclado y su deconvolución (es decir, la identificación correcta del clon individual en una biblioteca por detección de la presencia de un indicador asociado conocido (es decir, marcador o identificador) del clon en una o más mezclas o submezclas) se describen por ejemplo en el documento US 6975943 o en Klein y col. en *Genome Research*, (2000), 10, 798-807. La estrategia de mezcla preferiblemente es tal que cada clon en la biblioteca está distribuido en las mezclas de modo que se hace para cada clon una combinación única de mezclas. El resultado de esto es que una combinación determinada de (sub)mezclas identifica unívocamente un clon.

[0047] En la etapa (d) del procedimiento, las mezclas se fragmentan y se produce un conjunto de fragmentos para cada mezcla. La fragmentación puede ser aleatoria, es decir, por cizalladura o nebulización para crear un conjunto de fragmentos. En una realización preferida, las mezclas se digieren con endonucleasas de restricción para dar fragmentos de restricción. Cada mezcla, preferiblemente por separado, se somete a una digestión por endonucleasas. Cada mezcla se trata con la(s) misma(s) (combinación de) endonucleasa(s). En principio se puede usar cualquier endonucleasa de restricción. Las endonucleasas de restricción pueden ser cortadoras frecuentes (cortadoras 4 ó 5, tales como MseI o AvaII) o cortadoras infrecuentes (cortadoras 6 y más, tales como EcoRI, HindIII, PacI). Típicamente, las endonucleasas de restricción se seleccionan de modo que se obtengan fragmentos de restricción que, como media, estén presentes en una cantidad o tengan una determinada distribución de longitudes de que sea adecuada para la resolución de perfilado requerida y/o etapas posteriores. En algunas realizaciones, se pueden usar dos o más endonucleasas de restricción y en determinadas realizaciones, se pueden usar combinaciones de cortadoras infrecuentes y frecuentes. Para genomas grandes, se pueden usar ventajosamente, por ejemplo, tres o más endonucleasas de restricción. En algunas realizaciones, las enzimas de restricción se pueden usar para proporcionar extremos romos. Los correspondientes adaptadores (véase a continuación) entonces también pueden ser de extremos romos.

[0048] Se ligan adaptadores a uno o ambos extremos de los fragmentos en la etapa (e) para proporcionar fragmentos ligados a adaptador. Típicamente, los adaptadores son oligonucleótidos sintéticos definidos en otra parte en el presente documento. Los adaptadores usados en la presente invención, preferiblemente contienen una sección identificadora, en esencia como se define en otra parte en el presente documento. En algunas realizaciones, el adaptador contiene un identificador específico de mezcla, es decir, para cada mezcla se usa un adaptador que contiene un identificador único que indica inequívocamente la mezcla a partir de la cual se origina el fragmento. En algunas realizaciones, el adaptador contiene una sección identificadora degenerada, que se usa en combinación con un cebador que contiene un identificador específico de mezcla. El adaptador puede contener además sitios de unión del cebador en los que más tarde se puede iniciar la amplificación. Estos sitios de unión del cebador también se pueden ligar en una etapa posterior. Se prefiere que la sección identificadora (degenerada o no) esté situada entre el fragmento y el sitio de unión del cebador de modo que la amplificación desde el sitio de unión del cebador usando cebadores complementarios al sitio de unión del cebador amplifique al menos el identificador.

[0049] En algunas realizaciones, los fragmentos ligados a adaptador se pueden combinar en grupos más grandes, en particular cuando los adaptadores contienen un identificador específico de mezcla. Esta combinación en grupos más grandes puede ayudar a reducir el número de amplificaciones paralelas de cada conjunto de restricción ligado a adaptador obtenido de una mezcla.

[0050] Los fragmentos ligados a adaptador se pueden amplificar usando un conjunto de cebadores de los que al

menos un cebador amplifica el identificador específico de mezcla en la posición del identificador específico de mezcla o degenerado en el adaptador. El cebador puede contener (parte de) el identificador, pero el cebador también puede ser complementario de una sección del adaptador que está situada fuera del identificador, es decir en la dirección 3' del adaptador. La amplificación entonces también amplifica el identificador (véase también la fig. 4). En una realización, el cebador puede contener un identificador en una posición situada 5' desde la parte que es complementaria del adaptador, de modo que la amplificación introduce el identificador en el amplicón resultante.

[0051] Esta realización también permite la agrupación de fragmentos ligados a adaptador antes de la amplificación como se ha señalado antes. En una realización alternativa, cada mezcla de fragmentos ligados a adaptador, en las que el adaptador contenía una sección identificadora degenerada, se amplifica por separado usando un conjunto de cebadores de los que al menos un cebador contiene una sección específica de mezcla que sirve como un identificador, identificando así unívocamente la mezcla. En otra realización, el cebador es complementario de al menos parte del adaptador y proporciona un identificador en el fragmento ligado a adaptador amplificado, por ejemplo, conteniendo en el extremo 5' de la parte que es complementaria del adaptador, una secuencia identificadora. La amplificación con este cebador añade un identificador al amplicón.

[0052] De cualquier manera, el resultado es un conjunto de fragmentos ligados a adaptador amplificados, también representados como amplicones, que están conectados a la mezcla de la que proceden por la presencia en el amplicón del identificador específico de mezcla. En algunas realizaciones, se pueden crear subconjuntos de amplicones mediante amplificación selectiva, por ejemplo, usando cebadores que llevan nucleótidos selectivos en su extremo 3', esencialmente como se describe en otra parte en el presente documento.

[0053] Estos amplicones se pueden combinar en determinadas realizaciones, en un conjunto de amplicones combinados o la llamada biblioteca de secuencias.

[0054] En la etapa (f) del procedimiento, los fragmentos (o, cuando están amplificados, los amplicones) se someten a secuenciación, preferiblemente secuenciación de alta productividad como se describe en el presente documento a continuación. Durante la secuenciación, se determina al menos parte de la secuencia de nucleótidos del fragmento ligado a adaptador. Preferiblemente, se determina al menos la secuencia de parte del adaptador y parte de la secuencia del fragmento. Preferiblemente, la parte secuenciada permite la correlación de la secuencia con el clon de BAC. Preferiblemente, se determina la secuencia del identificador específico de mezcla y parte del fragmento (es decir, obtenida de la muestra de genoma). Preferiblemente, se determina una secuencia de al menos 10 nucleótidos del fragmento. En algunas realizaciones, se determinan al menos 11, 12, 13, 14 ó 15 nucleótidos del fragmento. El número de nucleótidos que se va a determinar como mínimo, otra vez dependerá del genoma así como de la plataforma de secuenciación. Por ejemplo, en plantas están presentes secuencias más repetitivas, por lo tanto se van a determinar secuencias más largas (25-75 nucleótidos) para un cóntigo de calidad comparable. Por ejemplo, los cálculos por ordenador de la secuencia genómica conocida de Arabidopsis han mostrado que, cuando se incluye un sitio de restricción de 6 pb en la etapa de secuenciación, es necesario determinar aproximadamente 20 pb por fragmento con el fin de asegurar que la mayoría de las secuencias son únicas en el genoma. Se puede determinar la secuencia del fragmento entero, pero esto no es una necesidad absoluta para la construcción de cóntigo de un clon de BAC.

[0055] En la etapa de secuenciación, para proporcionar el máximo cubrimiento de todos los fragmentos y mayor precisión, la biblioteca de secuencias se puede secuenciar con un nivel de redundancia medio (también conocido como tasa de sobremuestreo) de al menos 5. Esto significa que, como media, se determina la secuencia de al menos 5 amplicones obtenidos de la amplificación de un fragmento ligado a adaptador específico. En otras palabras: cada fragmento es secuenciado (estadísticamente) al menos 5 veces. Se prefiere una mayor redundancia ya que mejora la fracción de fragmentos que son muestreados en cada mezcla y la precisión de esas secuencias, así preferiblemente el nivel de redundancia es al menos 7, más preferiblemente al menos 10. Se usan niveles medios de redundancia de secuenciación mayores para compensar un fenómeno conocido como "variación de muestreo", es decir la fluctuación estadística aleatoria en los subconjuntos de muestreo de una "población" grande. Además, un nivel medio de redundancia de secuenciación más alto alivia las posibles diferencias en la abundancia de fragmentos amplificados que resulta de diferencias en sus tasas de amplificación causadas por la variación de longitud entre fragmentos y diferencias en la composición de la secuencia.

[0056] En la siguiente etapa (g), los fragmentos ligados a adaptador (parcialmente) secuenciados o amplicones se correlacionan o asignan al correspondiente clon, típicamente por ordenador mediante procedimientos computerizados. Se seleccionan los fragmentos ligados a adaptador o amplicones que contienen secciones de nucleótidos idénticas en la parte derivada del fragmento. Posteriormente, se identifican los diferentes identificadores

específicos de mezcla que están presentes en esos fragmentos ligados a adaptador o amplicones. La combinación de los diferentes identificadores específicos de mezcla y por lo tanto la secuencia del fragmento, se puede asignar unívocamente a un clon específico (un procedimiento conocido como "deconvolución"). Por ejemplo, en el caso de una estrategia de mezcla 3D (X,Y,Z), cada mezcla en la biblioteca es dirigida unívocamente por una combinación de 3 identificadores específicos de mezcla. Cada clon aparece más de una vez en la biblioteca, de modo que por cada aparición de un clon en la biblioteca, se puede hacer una combinación de 3 identificadores específicos de mezcla en combinación con la misma sección derivada del fragmento. En otras palabras, una sección derivada de fragmento procedente de un clon estará marcada con 3 identificadores diferentes. Las secciones derivadas de fragmento únicas, cuando se observan en combinación con los 3 identificadores se pueden asignar a un solo clon de BAC. Esto se puede repetir para cada uno de los fragmentos ligados a adaptador o amplicones que contienen otras secciones únicas de nucleótidos en la parte derivada de fragmento. Este procedimiento de deconvolución se puede hacer más fácil manteniendo el equivalente genómico por mezcla relativamente bajo (<0,3, preferiblemente 0,2), reduciendo así la posibilidad de que el mismo fragmento esté presente dos veces en la misma mezcla derivada de diferentes clones.

[0057] Una muestra de ADN se convierte en una biblioteca de BAC. La biblioteca de BAC se puede mezclar en un conjunto de mezclas (M) (p. ej., 3 mezclas, que contiene cada una aproximadamente 0,3 EG). Cada mezcla se divide en (X+Y+Z) submezclas (típicamente una pila de microplacas de valoración o mezclas en filas y/o columnas).

[0058] Los fragmentos ligados a adaptador secuenciados o amplicones que ahora están conectados a un clon particular en la biblioteca, se pueden usar en la construcción de un cóntigo basado en el emparejamiento de las secuencias de las secciones derivadas de fragmento. Los cóntigos de cada clon después se pueden alinear para generar un mapa físico. En una realización, los fragmentos derivados del mismo clon se pueden ordenar para construir un cóntigo a partir del clon. Basándose en la aparición de las secuencias de fragmentos en dos o más clones (marcadores de WPG), los clones se pueden conectar entre sí en la etapa (h) de la invención, formando así un cóntigo de clones y por lo tanto un mapa físico de la muestra de genoma.

[0059] La secuenciación de alta productividad usada en la presente invención, es un procedimiento para la experimentación científica específicamente importante en los campos de la biología y la química. Mediante una combinación de procedimientos robóticos modernos y otro hardware de laboratorio especializado, el investigador puede cribar de forma eficaz grandes cantidades de muestras simultáneamente.

[0060] Se prefiere que la secuenciación se realice usando procedimientos de secuenciación de alta productividad, tal como los procedimientos descritos en los documentos WO 03/004690, WO03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, y WO 2005/003375, de Seo y coll. (2004) *Proc. Natl. Acad. Sci. USA* 101:5488-93, y tecnologías de Helicos, Illumina), US Genomics, etc.

Roche Applied Science

[0061] En algunas realizaciones, se prefiere realizar la secuenciación usando el aparato y/o procedimiento descrito en los documentos WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007 y WO 2005/003375. Actualmente, la tecnología descrita permite secuenciar 400.000 lecturas de secuencia en una sola serie de GS FLX Titanium, y es 100 veces más rápida y barata que la tecnología de la competencia. La tecnología de secuenciación contiene esencialmente 5 etapas: 1) fragmentación del ADN y ligamiento de adaptadores específicos para crear una biblioteca de ADN monocatenario (ADNmc); 2) asociación del ADNmc a perlas, emulsión de las perlas en microrreactores de agua en aceite y realización de la PCR en emulsión para amplificar las moléculas de ADNmc individuales; 3) selección de/enriquecimiento de las perlas que contienen moléculas de ADNmc amplificadas en su superficie, 4) deposición de las perlas que llevan ADN en una placa PicoTiter™; y 5) secuenciación simultánea en más de 1 millón de pocillos de una placa PicoTiter™ por generación de una señal de luz de pirofosfato. El procedimiento se explicará con más detalle a continuación.

[0062] En una realización preferida, la secuenciación comprende las etapas de:

a. asociar fragmentos adaptados a perlas, asociándose cada perla con un solo fragmento adaptado;
b. emulsionar y amplificar los fragmentos asociados en las perlas en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una sola perla;
c. cargar las perlas en pocillos, comprendiendo cada pocillo una sola perla; y generar una señal de pirofosfato.

[0063] En la primera etapa (a), se ligan adaptadores de secuenciación a fragmentos en la biblioteca de

combinación. Dicho adaptador de secuenciación incluye al menos una región para la asociación con un oligonucleótido complementario unido a una perla, una región de cebador de secuenciación y una región de cebador de PCR. Así se obtienen los fragmentos adaptados.

[0064] En la primera etapa, los fragmentos adaptados se asocian a las perlas, asociándose cada perla con un solo fragmento adaptado. A la mezcla de fragmentos adaptados se añaden las perlas en exceso para asegurar la asociación de un solo fragmento adaptado por perla para la mayoría de las perlas (distribución de Poisson). En la presente invención, los adaptadores que se ligan a los fragmentos de restricción obtenidos de los clones pueden comprender una sección que es capaz de asociarse a una perla.

[0065] En la siguiente etapa, las perlas se emulsionan en microrreactores de agua en aceite, comprendiendo cada microrreactor de agua en aceite una sola perla. Los reaccionantes de la PCR están presentes en los microrreactores de agua en aceite, permitiendo que tenga lugar una reacción de PCR en los microrreactores. Posteriormente, los microrreactores se rompen, y se enriquecen las perlas que comprenden ADN (perlas positivas para ADN), es decir, se separan de las perlas que no contienen fragmentos amplificados.

[0066] En una etapa siguiente, las perlas enriquecidas se cargan en pocillos, comprendiendo cada pocillo una sola perla. Los pocillos preferiblemente son parte de una placa PicoTiter™, que permite la secuenciación simultánea de un gran número de fragmentos.

[0067] Después de añadir perlas que llevan enzima, se determina la secuencia de los fragmentos usando pirosecuenciación. En etapas sucesivas, la placa PicoTiter™ y las perlas así como las perlas con enzima en las mismas, se someten a diferentes desoxirribonucleótidos en presencia de reaccionantes de secuenciación convencionales, y tras la incorporación de un desoxirribonucleótido se genera una señal de luz que se registra. La incorporación del nucleótido correcto generará una señal de pirosecuenciación que se puede detectar.

[0068] La propia pirosecuenciación se conoce en la técnica y se describe, entre otros en www.biotagebio.com; www.pyrosequenc-ing.com/section technology. La tecnología se aplica además, 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 ahora Roche Diagnostics), y Margulies y col., *Nature* 2005, 437, 376-380.

[0069] En la presente invención, las perlas están equipadas preferiblemente con secuencias de cebadores o partes de las mismas, que son capaces de ser extendidas por polimerización para dar amplicones unidos a perlas. En otras realizaciones, los cebadores usados en la amplificación están equipados con secuencias, por ejemplo en su extremo 5', que permiten la unión de los amplicones a las perlas con el fin de permitir la posterior polimerización por emulsión seguida de secuenciación. Alternativamente, los amplicones se pueden ligar con adaptadores de secuenciación antes de ligamiento a las perlas o la superficie. Los amplicones secuenciados pondrán de manifiesto la identidad del identificador y por consiguiente la combinación de identificadores pone de manifiesto la identidad del clon.

Tecnologías Illumina

[0070] Uno de los procedimientos para la secuenciación de alta productividad está disponible en tecnologías Illumina (www.illumina.com) y se describe, entre otros, en los documentos WO0006770, WO0027521, WO0058507, W00123610, WO0157248, WO0157249, W002061127, W003016565, WO03048387, W02004018497, W02004018493, WO2004050915, WO2004076692, W02005021786, W02005047301, WO2005065814, W02005068656, WO2005068089, WO2005078130. En esencia, el procedimiento empieza con los fragmentos de ADN ligados a adaptador, en este caso particular fragmentos de restricción ligados a adaptador de las mezclas de cromosomas artificiales, como se describen en otra parte en el presente documento. El ADN ligado a adaptador se une aleatoriamente a un césped denso de cebadores que están unidos a una superficie sólida, típicamente en una celda de flujo. El otro extremo del fragmento ligado a adaptador hibrida con un cebador complementario en la superficie. Los cebadores se extienden en presencia de nucleótidos y polimerasas en una amplificación llamada de puente de fase sólida para proporcionar fragmentos bicatenarios. Esta amplificación de puente en fase sólida puede ser una amplificación selectiva. La desnaturalización y repetición de la amplificación de puente en fase sólida da como resultado agrupaciones densas de fragmentos amplificados distribuidos por la superficie. La secuenciación se inicia por adición de cuatro nucleótidos terminadores reversibles marcados de forma diferente, cebadores y polimerasa a la celda de flujo. Después del primer ciclo de extensión del cebador, se detectan los marcadores, se registra la identidad de la primera base incorporada y se eliminan de la base incorporada el extremo 3' bloqueado y el fluoróforo. Después se determina la identidad de la segunda base de la misma forma y así continua la

secuenciación.

[0071] En la presente invención, los fragmentos ligados a adaptador o los amplicones se unen a la superficie por la secuencia de unión del cebador o la secuencia del cebador. La secuencia se determina como se ha señalado, incluyendo la secuencia identificadora y (parte de) el fragmento. La tecnología actualmente disponible permite la secuenciación de longitudes de lecturas de un máximo de 125 bases. Para el propósito del perfil del genoma completo, puede ser suficiente una longitud de lectura de secuencia de 36 bases, pero esto depende del tamaño del genoma y la composición de la secuencia (véase más adelante). Mediante un diseño económico de adaptadores y los cebadores unidos a la superficie, la etapa de secuenciación lee por el identificador de la muestra el resto de la secuencia de reconocimiento de la endonucleasa de restricción, cualquier base selectiva opcional y la secuencia interna del fragmento de restricción. Por ejemplo, en el caso de lecturas de secuencia de 36 bases, cuando se usa una muestra de identificador de 6 bases, cuando el resto de la cortadora infrecuente EcoRI (GAATTC) son 6 bases, cuando se usan 2 bases selectivas, entonces la longitud de la secuencia interna del fragmento de restricción será 36 menos 14 = 22 bases, que se puede usar para identificar unívocamente el fragmento de restricción en la muestra. Obsérvese que la secuencia del sitio de la enzima de restricción y las bases selectivas (opcionales) también están presentes en el genoma, pero puesto que estas secuencias son comunes para todos los fragmentos de restricción, no contribuyen a la capacidad de asignar las lecturas de secuencia a clones únicos en las bibliotecas.

[0072] En la etapa (i) del procedimiento, las lecturas de secuencia se generan a partir de una muestra de ADN. Esta puede ser la misma muestra que la usada en la generación del banco de clones, pero también puede ser otra muestra de la misma especie. El uso de un origen diferente para la muestra para generar lecturas de secuencia, permite usar bancos de clones que ya existen, aunque a expensas de la calidad de la secuencia genómica así obtenida (la generación de cóntigos puede ser más difícil) o la secuencia genómica resultante es de menor calidad o contiene más huecos. A partir de las lecturas de secuencia se pueden generar andamiajes o cóntigos por alineación de las mismas, como en la etapa (j) y anclándolas al cóntigo de clones para construir un superandamiaje o una secuencia genómica.

[0073] En una realización de la presente invención, también se pueden usar fragmentos generados aleatoriamente a partir de los BAC (u otros cromosomas artificiales) o las mezclas de BAC y determinar (parte de) la secuencia de los mismos, usando las tecnologías de secuenciación descritas en el presente documento. La calidad del ensamblado del cóntigo mejoraría entonces incluso más ya que no solo conectan los extremos de los fragmentos de restricción de BAC, sino que se puede generar un cóntigo de (una parte de) los BAC. Preferiblemente, combinado con las lecturas de secuencia y los cóntigos obtenidos a partir de la muestra de ADN, esto aumenta más la calidad.

[0074] Por lo tanto, en una realización preferida, se genera una secuencia genómica borrador a partir de una combinación de cóntigos derivados de BAC, es decir, de secuencias de los extremos de BAC y/o secuencias de fragmentos de restricción de BAC y/o clones secuenciados aleatoriamente con (cóntigos generados a partir de) lecturas de secuencia de la muestra de ADN que se pueden haber obtenido por fragmentación (enzima de restricción).

[0075] En paralelo o posteriormente a la generación del cóntigo de clones y/o el mapa físico, se pueden generar lecturas de secuencia a partir de una muestra usando un procedimiento más directo, indicado también como "secuenciación aleatoria" o "secuenciación aleatoria del genoma completo" (WGS). En esta etapa, los datos de secuencia se generan a partir de una muestra de ADN y/o a partir de uno o más clones de cromosomas artificiales de la muestra de ADN. La muestra puede ser la muestra que pone en marcha el banco de clones, pero también puede ser otra muestra de la misma especie o variedad y por lo tanto contiene una pequeña cantidad de polimorfismos comparado con la muestra del banco de clones. Los datos de secuencias típicamente se generan por fragmentación de la muestra de ADN, por ejemplo por cizalladura, nebulización o digestión con enzimas de restricción. Los fragmentos pueden estar o no ligados a adaptador. El adaptador puede contener marcadores para identificar el origen de los fragmentos o la muestra usando los llamados identificadores. Los fragmentos ligados a adaptador se pueden amplificar de forma selectiva o no selectiva, por ejemplo, usando tecnología basada en AFLP usando cebadores complementarios del adaptador que se pueden extender en el extremo 3' con uno o más nucleótidos selectivos, esencialmente como se describe en otra parte en el presente documento. De cualquier forma se generan las lecturas de secuencia usando preferiblemente tecnologías de secuenciación de alta productividad tales como las tecnologías de secuenciación basadas en pirosecuenciación como se describe en otra parte en el presente documento.

[0076] Las lecturas de secuencia después se ensamblan en cóntigos y/o se anclan al cóntigo generado a partir de la biblioteca de BAC.

[0077] En una realización preferida, se usa más de una tecnología de secuenciación para generar las lecturas de secuencia de la muestra de ADN. Como se señala en los dibujos y el texto, las diferentes tecnologías proporcionan lecturas de diferentes longitudes que preferiblemente ayudan al anclaje y construcción de un cóntigo extendido.

[0078] El uso de las lecturas de secuencia "directas" no solo completa el cóntigo de BAC, sino que también es capaz de llenar los huecos dejados por el cóntigo generado por BAC. De hecho, esta es una de las principales ventajas de la presente invención. En estrategias previas, se contempla el uso de datos de secuencias adicionales (sea generados de nuevo o tomados de fuentes conocidas) solo en vista de la posibilidad de anclar los datos de secuencia al cóntigo para llenar los datos de secuencia para el cóntigo de BAC, no en el contexto de conectar los diferentes cóntigos de clones entre sí para generar un cóntigo (andamiaje) que cubra una parte mayor del genoma. La presente invención ahora también proporciona en alguna realización, la posibilidad de extender el cóntigo de BAC y llenar los huecos dejados entre los BAC por una parte y el o los cóntigos generados con lecturas de secuencia por otra parte, conduciendo así a una calidad mejorada del genoma borrador resultante, como se ilustra en las figuras.

[0079] Opcionalmente, los datos de las lecturas de secuencia también se pueden complementar mediante datos de secuencias obtenidos por técnicas de didesoxisecuenciación de Sanger ya que esto puede ayudar más al ensamblado de cóntigos de alta calidad. También se pueden complementar datos mediante las técnicas de secuenciación llamadas de siguiente-siguiente generación, tales como las de Pacific Biosciences que pueden suministrar resultados de secuencias de hasta múltiples kb de longitud.

[0080] Al obtener la lectura de secuencia, en realizaciones preferidas, la muestra de ADN también se puede someter a tecnologías de reducción de la complejidad más reproducibles tales como por ejemplo AFLP (documento EP534858) y/o estrategias basadas en AFLP para la secuenciación de genomas complejos, por ejemplo tal como se describe en el documento WO2006/13773 en el que se usan dos combinaciones de enzimas de restricción diferentes en la tecnología de AFLP para generar un cóntigo de datos de secuencias.

[0081] Por lo tanto, la presente invención determina una secuencia genómica borrador basada en una ruta de dos vías. La primera ruta es la generación de un cóntigo de clones de cromosomas artificiales (BAC) usando el perfilado genómico completo (WPG). Usando preferiblemente un subconjunto aleatorio pero reproducible de fragmentos de restricción de mezclas de BAC, se pueden generar cóntigos en una cantidad relativamente pequeña de datos de cobertura, conduciendo a un cóntigo de clones de BAC que se pueden describir como "finos" o de "baja densidad". Fino en cuanto a que hay un espacio relativamente grande entre los fragmentos de restricción secuenciados que permite que los cóntigos de BAC se ensamblen con una cantidad relativamente económica de secuenciación y potencia computacional. Por consiguiente, la fracción del genoma completo que se secuencia en el procedimiento de WPG es relativamente baja (ya que el objetivo del WPG son los cóntigos de clones y no la secuenciación del genoma completo).

[0082] La segunda vía es generar/recoger datos de secuencias, preferiblemente de la misma muestra (ADN íntegro) usando máquinas y procedimientos de secuencias de alta productividad tales como los conocidos de Roche Applied Science (que producen lecturas de hasta 1 kb) e Illumina (GS FLX) que producen lecturas de 36-125 nt, y de otros suministradores (tales como Helicos, Intelligent Biosystems, Danaher Motion-Dover, Pacific Biosciences etc.). Los datos de secuencia se pueden anclar directamente en el cóntigo de BAC, pero se pueden usar primero para generar cóntigos a partir de los datos de secuenciación. En una etapa posterior, estos cóntigos basados en secuencias se pueden anclar al cóntigo de BAC de la primera ruta. Además, los datos de secuencia y los cóntigos a partir de las lecturas de secuencia se pueden usar para conectar los cóntigos de BAC existentes entre sí, es decir, cerrar huecos entre y dentro de los andamiajes. La ventaja de combinar la tecnología es que los datos de secuencia se obtienen de la misma muestra usando diferentes procedimientos que se pueden complementar entre sí, como se ilustra en las figuras adjuntas. Es particularmente ventajoso combinar el WPG con dos o más tecnologías de secuenciación (de alta productividad) diferentes. Una de las ventajas particulares de las estrategias descritas en el presente documento, y a diferencia de las estrategias de la técnica anterior que se basan muy a menudo en procedimientos de fuerza bruta tales como en el documento WO 03/027311, es que se usan conjuntos de datos relativamente pequeños que después se combinan.

[0083] La secuenciación del ADN obtenido a partir de una muestra (ADN íntegro) se puede basar en representaciones de complejidad reducida del ADN íntegro, p. ej., usando ADN digerido con endonucleasas de restricción que produce fragmentos de restricción que se pueden marcar ("código de barras") para indicar su origen cuando sea necesario. Estos fragmentos de restricción después se pueden someter a secuenciación usando, preferiblemente, las tecnologías de secuenciación de alta productividad, tales como las mencionadas en otra parte

en el presente documento. También se pueden considerar otras formas de reducción de la complejidad, incluyendo, pero sin limitar, la fragmentación aleatoria (por nebulización, ultrasonidos, cizalladura u otros medios mecánicos) seguido de selección de tamaños de los fragmentos en un intervalo de tamaños particular, selección de Cot (basado en cinéticas de hibridación diferentes de secuencias únicas frente a repetidas) u otros procedimientos de reducción de la complejidad. En principio, el uso de fragmentos de restricción, por ejemplo, los obtenidos por restricción del ADN íntegro con cortadoras infrecuentes tales como EcoRI, típicamente de una longitud de 2-3 kb (en genomas ricos en AT), y la determinación de las secuencias de nucleótidos de los extremos de los fragmentos de restricción (típicamente 30-400 pb por extremo, dependiendo de la tecnología de secuenciación usada), puede ser suficiente para crear un cóntigo y anclar estos fragmentos al cóntigo de WGP (el mapa físico). Estará claro que se pueden usar otras enconucleasas de restricción (p. ej., EcoRI/MseI). Se prefiere crear cóntigos a partir de los datos de secuencias (fragmentos relativamente cortos) obtenidos del ADN íntegro y posteriormente anclar estos cóntigos (relativamente más largos) al cóntigo de BAC disponible en lugar de anclar inmediatamente las lecturas de secuenciación al cóntigo de BAC. Otra vez, la ventaja reside en el uso de subconjuntos relativamente más pequeños de datos que permiten mayor eficacia en el "procesamientos de los datos" y por lo tanto requisitos menos duros en términos de potencia computacional. Dicho procedimiento también puede permitir ventajosamente realizar los cálculos o partes o elementos del mismo en un ordenador de sobremesa o portátil, en lugar de servidores y ordenadores centrales potentes. Otra ventaja de este procedimiento de dos rutas está en el uso del ADN íntegro como la segunda fuente de información de secuencias (frente al uso de la biblioteca de BAC como la primera fuente). Una biblioteca de BAC siempre carece de la cobertura íntegra y completa de un genoma. Usando el ADN íntegro como una fuente adicional de ADN, es posible y ventajoso lograr o al menos estar cerca de la cobertura casi completa del genoma que se investiga.

[0084] Los ejemplos de dicha secuenciación basada en fragmentos de restricción se describe, por ejemplo, en el documento WO2006/13773 que describe el uso de AFLP como una tecnología de reducción de la complejidad en combinación con la secuenciación de alta productividad para crear también secuencias genómicas borrador de alta calidad. Por lo tanto, en esta realización, se generan cóntigos de BAC como se ha indicado antes en el presente documento, y se combinan con los cóntigos derivados de ejecutar el procedimiento descrito en el documento WO2006/137734.

[0085] En una realización alternativa, la secuenciación del ADN basada en el ADN íntegro se puede basar en "marcadores de secuencia aleatorios". En combinación con los sistemas de alta productividad conocidos de, entre otros, Illumina, la información de secuencia generada también se puede anclar en el cóntigo de BAC obtenido del WGP. La realización deriva del hecho de que el BAC es el último "extremo apareado". La ventaja de esta tecnología reside en el hecho de que la "secuenciación profunda" (es decir, secuenciación de varios equivalentes genómicos (EG) con el fin de obtener una mayor calidad de los datos) no es esencial para obtener un ensamblado de genoma de alta calidad, ya que la ordenación principal del genoma ya la ha proporcionado el cóntigo de BAC (y los datos de secuencia se usan principalmente para llenar los huecos en el cóntigo de BAC). Por lo tanto, la presente metodología permite el uso de menos datos de secuencias (por la "secuenciación menos profunda", es decir secuenciación de solo uno o unos pocos EG) sin afectar a la calidad del ensamblado. Esto da como resultado un procedimiento más económico, puesto que la "secuenciación menos profunda" es inherentemente más barata que la secuenciación profunda. En el caso de que para algunas áreas sean necesarios datos de secuencias de alta calidad, se puede realizar la secuenciación profunda en regiones seleccionadas, seleccionando algunos clones de BAC o cóntigos de BAC.

[0086] Por lo tanto, un aspecto de la invención se refiere al uso del procedimiento descrito en el presente documento para la secuenciación selectiva de parte de un genoma o región genómica seleccionada, preferiblemente con niveles de cobertura variables.

[0087] Comparado con el documento WO03/027311, el presente procedimiento difiere en que los subconjuntos se hacen basándose en fragmentos de restricción o en fragmentes de restricción en combinación con cizalladura aleatoria y no en solo cizalladura aleatoria. Además, y al contrario del documento WO03/027311, la secuenciación de los fragmentos de restricción se basa en una cobertura muy baja. Basándose en esta baja cobertura, se genera un cóntigo de BAC que es muy "fino", es decir, contiene una cantidad relativamente pequeña de datos. Este cóntigo "fino" después se complementa con datos obtenidos de las lecturas de secuencia. Este es un procedimiento más eficaz para la generación del mapa físico y hace un uso más eficaz de la potencia computacional (limitada) para los proyectos de esta escala.

[0088] Los resultados de generar el genoma borrador se pueden proporcionar como un producto separado que comprende, opcionalmente en un formato digital:

- los datos de secuencia asociados con la biblioteca de BAC y los cóntigos de BAC asociados;
- los datos de secuencias asociados con la secuenciación del ADN íntegro y los cóntigos asociados;
- software para presentar los cóntigos de BAC, los cóntigos de ADN, los cóntigos combinados y la secuencia genómica borrador, desde un nivel de secuencia genómica borrador general al nivel de los nucleótidos y la superposición entre dos fragmentos
- software para generar cóntigos a partir de los datos de secuencia separados
- una aplicación para presentar marcadores moleculares en diferentes cóntigos y mapas
- software para visualizar la calidad de datos y huecos en la secuencia.

[0089] El producto se puede proporcionar en un ordenador portátil equipado con memoria flash o un disco duro, un soporte de datos de solo lectura tal como un CD-ROM o DVD o similar. Alternativamente, el producto se puede proporcionar en forma de un servidor basado en web de modo que el producto se proporciona en un formato digital, en un servidor preferiblemente seguro.

[0090] Por lo tanto, un producto de ejemplo puede contener uno o más de los siguientes componentes:

a) Un mapa físico ensamblado (perfilado genómico completo; WGP).
El mapa se puede ensamblar usando software de construcción de cóntigos tal como software de cóntigos de huella dactilar (FPC) adaptados para usar con secuencias en lugar de movilidades de banda. El cóntigo se puede construir basándose en secuencias de nucleótidos derivadas de los clones mezclados, tal como clones de BAC que se han asignado a clones individuales por deconvolución basada en secuencias identificadoras;
b) Ensamblados, incluyendo cóntigos, supercóntigos y/o andamiajes de secuenciación genómica completa (WGS)
Los ensamblados se pueden generar usando paquetes de software de ensamblado de genoma tales como Newbler (454 Life Sciences/Roche Applied Sciences y Short Oligonucleotide Analysis Package (SOAP) de novo (http://soap.genom-ics.org.cn), basado en la secuenciación de siguiente generación (es decir, pirosecuenciación de alta productividad) y/o datos de secuenciación de Sanger;
c) Una secuencia genómica borrador.
La secuencia genómica borrador se puede basar en la integración del WGP (el mapa y los datos de (a)) y WGS (los datos de (b)). La secuencia genómica borrador se puede proporcionar en diferentes formatos incluyendo archivos Fasta y delimitados por tabuladores;
d) Software de visualización.
Software de visualización tal como FPC para ver ensamblados de WGP y WGS, secuencias y clon asociado así como sus combinaciones;
e) Datos de secuencia.
Los datos de secuencia reales que se han usado en la generación del mapa físico o la secuenciación genómica completa. Esto puede ayudar a mejorar más los datos, a la verificación de los datos, permitir la generación de un mapa físico mejorado, por ejemplo, basado en la obtención de datos adicionales.
f) Un dispositivo de almacenamiento o soporte de datos.
El dispositivo o soporte puede ser un disco duro o disco flash que comprende uno o más de los datos y software descritos en (a) a (f);
g) Un ordenador tal como un ordenador portátil u ordenador portátil pequeño que comprende uno o más de los componentes (a) a (f) o parte de los mismos.

Ejemplos

Arabidopsis thaliana ecotipo Columbia

[0091] Se usó una biblioteca de BAC que contenía 6144 BAC (aproximadamente 5 equivalentes genómicos).

[0092] Se llevó a cabo un serie de Illumina clásico en mezclas fragmentadas con enzimas de restricción (EcoRI y MseI), dando como resultado aproximadamente 65.000 lecturas de secuencia deconvolucionables distintas desde el lado de EcoRI. El ensamblado de las lecturas (FPC, Soderlund, C., S. Humphrey, A. Dunhum, y L. French (2000). Contigs built with fingerprints, markers and FPC V4.7. *Genome Research* 10:1772-1787) en 4599 BAC (74,8%) dio como resultado 234 cóntigos con 2-125 BAC por cóntigo. La validación de la secuencia genómica publicada por el análisis por BLAST de las lecturas de secuencia, mostró que aproximadamente 52.000 lecturas dieron 100% de aciertos, cubriendo 99% del genoma con un hueco máximo de 125 kpb. Había 50.000 aciertos únicos; como media 2.355 pb entre marcadores y estaban representados 80% de todos los sitios EcoRI.

Melón

[0093] El melón tiene un tamaño de genoma calculado de 450 Mpb.

[0094] Se analizaron 47.616 BAC derivados de bibliotecas de EcoRI y HindIII, dando un total de aproximadamente 13 equivalentes genómicos. El 50% de todas las lecturas se podían deconvolucionar a BAC (40.063 BAC; 85%) y se marcaron unívocamente. Disponibles para la construcción de cóntigos: 9.417.2459 lecturas de Illumina GA II de 36 bases; obtenidas en 5 series de secuenciación GA II. 196.256 lecturas de secuencia únicas se conectaron con 40.063 clones de BAC con una media de 33 lecturas ancladas. Estas lecturas se ensamblaron en 670 cóntigos y 8.213 BAC de una unidad. Como media 15 BAC por cóntigo (> 1,8 Mpb) y >90% de cobertura genómica calculada. Véase la figura 3 para la distribución de tamaños de cóntigos.

Melón:

[0095] Los andamiajes de WGS del melón se integraron con los cóntigos de BAC de WGP del melón. El tamaño calculado del genoma del melón es 450 Mpb.

Entrada:

[0096]

Tipo de datos	nº andamiajes/cóntigos	tamaño de andamiaje N50 (kb)	Tamaño del andamiaje mayor (Mpb)	Cobertura total (Mpb)
WGP	1,88	546	3,1	375
WGS*	21.126 (> 1000 bp)	422	3,07	375
* comprende:				

[0097] Se produjeron los siguientes datos de secuencia en la plataforma GS FLX Titanium para el ADN nuclear de la línea de melón:

1) 17 series aleatorias, que comprendían un total de 16.171.153 lecturas
2) 5 series de extremos apareados de 3 Kb que comprendían un total de 4.844.561 lecturas
3) 3,5 series de extremos apareados de salto largo (~20 Kb), que comprendían un total de 3.448.598 lecturas.
4) 1 serie de extremo aleatorio-EcoRI, que comprendía un total de 789.048 lecturas.

[0098] El número total de series realizadas fue 26,5 y el número total de lecturas generadas es de 25.253.360. Estas lecturas representan un total de 8.691.334.029 bases (8,69 Gpb) del genoma nuclear del melón (es decir, excluyendo secuencias de cloroplastos y mitocondriales y secuencias conectoras de las bibliotecas de extremos apareados). Con un tamaño de genoma calculado de 450 Mbp, esto representa 19,43 veces la cobertura del genoma del melón: (rotura ~12,44 X aleatoria; ~3,72 X 3 kb PE; ~2,65 X salto largo; y ~0,61 X *Eco*RI-extremo aleatorio).

[0099] El procedimiento conecta los andamiajes de WGS con los cóntigos de WGP. Como etapa del procedimiento, se determina si un andamiaje de WGS solapa/empareja con un solo cóntigo de WGP o con múltiples cóntigos de WGP (basado en la presencia de secuencias con marcador de WGP en el andamiaje de WGS). Los criterios para la conexión de andamiajes de WGS con cóntigos de WGP son el número de marcadores de WGP que tienen un 100% de emparejamiento de secuencia. Todos los emparejamientos que se hacen se anotan para saber si se basan en al menos 1, 2 o más de 2 secuencias con marcador de WGP que se emparejan. Para los andamiajes de WGS que cubren un cóntigo de WGP entero, se han distinguido cuatro situaciones diferentes, que reflejan 4 niveles de confianza diferentes para conectar estos WGS y cóntigos de WGP. Se conectaron 5630 andamiajes de WGS que cubrían 77 Mpb con 838 cóntigos de BAC únicos. Se conectaron 470 andamiajes de WGS que cubrían 231 Mpb con 903 cóntigos de BAC múltiples. Estos dos conjuntos de datos se superponen porque el número total de cóntigos de BAC disponibles era 1088, que es inferior a 838+903. Los BAC de una unidad (no colocados en cóntigos de BAC) no se incluyeron en el análisis.

[0100] Un cóntigo de BAC aleatorio se considera una "semilla" para construir un superandamiaje basado en la presencia de secuencias con marcadores de WGP compartidas en cóntigos de BAC y sus andamiajes de WGS conectados. Véase la figura 5 más adelante. En esencia, los cóntigos de WGP y andamiajes de WGS se conectaron si al menos una secuencia con marcador de WGP era compartida entre ellos y no se identificaban marcadores en

conflicto. La semilla crecerá hasta que no se puedan hacer más conexiones o en el caso de que se produzca un punto de ramificación, p. ej., cuando andamiajes de WGS de solapamiento múltiple se conectan con el mismo cóntigo de BAc (figura 6).

[0101] Siguiendo el procedimiento descrito antes, se generaron 329 superandamiajes que comprendían 289 Mpb de la secuencia genómica del melón.

REIVINDICACIONES

1. Procedimiento para determinar la secuencia genómica que comprende las etapas de:

- proporcionar un mapa físico de una muestra de genoma por secuenciación de los extremos de fragmentos de fragmentos de clones de cromosomas artificiales mezclados;
- proporcionar un conjunto de lecturas de secuencia de la muestra de genoma;
- generar un cóntigo del mapa físico y las lecturas de secuencia para construir una secuencia genómica.

2. Procedimiento para determinar una secuencia genómica que comprende las etapas de:

(a) proporcionar una muestra de ADN;
(b) generar un banco de clones de cromosomas artificiales (p. ej., BAC, YAC) en el que cada clon de cromosoma artificial contiene parte de la muestra de ADN;
(c) combinar los clones de cromosomas artificiales en una pluralidad de mezclas, en las que cada clon está presente en más de una mezcla;
(d) proporcionar un conjunto de fragmentos para cada mezcla;
(e) ligar adaptadores a uno o ambos lados de los fragmentos,
(f) determinar la secuencia de al menos parte del adaptador y parte del fragmento;
(g) asignar las secuencias de los fragmentos a los correspondientes clones;
(h) construir un cóntigo de clones generando así un mapa físico de la muestra de genoma;
(i) generar lecturas de secuencia de una muestra de ADN;
(j) alinear las lecturas de secuencia y/o cóntigos o andamiajes desde las lecturas de secuencia al cóntigo de clones para construir así una secuencia genómica/superandamiaje.

3. Procedimiento de acuerdo con la reivindicación 2, en el que al menos un adaptador contiene un identificador específico de mezcla o una sección de identificador degenerado, respectivamente, para proporcionar fragmentos ligados a adaptador que contienen identificador.

4. Procedimiento de acuerdo con las reivindicaciones 2 ó 3, en el que los fragmentos ligados a un adaptador se amplifican usando

- un cebador que amplifica al menos el identificador y parte del fragmento; o
- un cebador que contiene una sección que es complementaria de la sección degenerada en el adaptador e introduce un identificador en el fragmento amplificado; o
- un cebador que es complementario de al menos parte del adaptador y proporciona un identificador en el fragmento ligado a adaptador amplificado.

5. Procedimiento de acuerdo con las reivindicaciones 2-4, en el que los fragmentos para una mezcla se generan por fragmentación aleatoria de las mezclas y/o fragmentación con enzimas de restricción de las mezclas.

6. Procedimiento de acuerdo con las reivindicaciones 2-5, en el que las lecturas de secuencia se obtienen de la muestra de ADN fragmentada y/o de uno o más clones de cromosomas artificiales de la muestra de ADN.

7. Procedimiento de acuerdo con las reivindicaciones 2-6, en el que las lecturas de secuencia se obtienen de la muestra de ADN fragmentada aleatoriamente y/o de uno o más clones de cromosomas artificiales de la muestra de ADN.

8. Procedimiento de acuerdo con las reivindicaciones 2-6, en el que las lecturas de secuencia se obtienen de fragmentos de restricción que se han obtenido por fragmentación con enzimas de restricción de la muestra de ADN y/o de uno o más clones de cromosomas artificiales de la muestra de ADN.

9. Procedimiento de acuerdo con la reivindicación 8, en el que los fragmentos de restricción son fragmentos de restricción ligados a adaptador.

10. Procedimiento de acuerdo con la reivindicación 9, en el que los fragmentos de restricción ligados a adaptador se amplifican de forma selectiva o no selectiva.

11. Procedimiento de acuerdo con cualquiera de las reivindicaciones anteriores, en el que la secuenciación se lleva a cabo mediante secuenciación de alta productividad.

12. Procedimiento de acuerdo con la reivindicación 11, en el que la secuenciación de alta productividad se realiza sobre un soporte sólido.

13. Procedimiento de acuerdo con la reivindicación 11 ó 12, en el que la secuenciación de alta productividad se basa en secuenciación por síntesis.

14. Procedimiento de acuerdo con la reivindicación 11 ó 12, en el que la secuenciación se basa en pirosecuenciación.

FIG 1

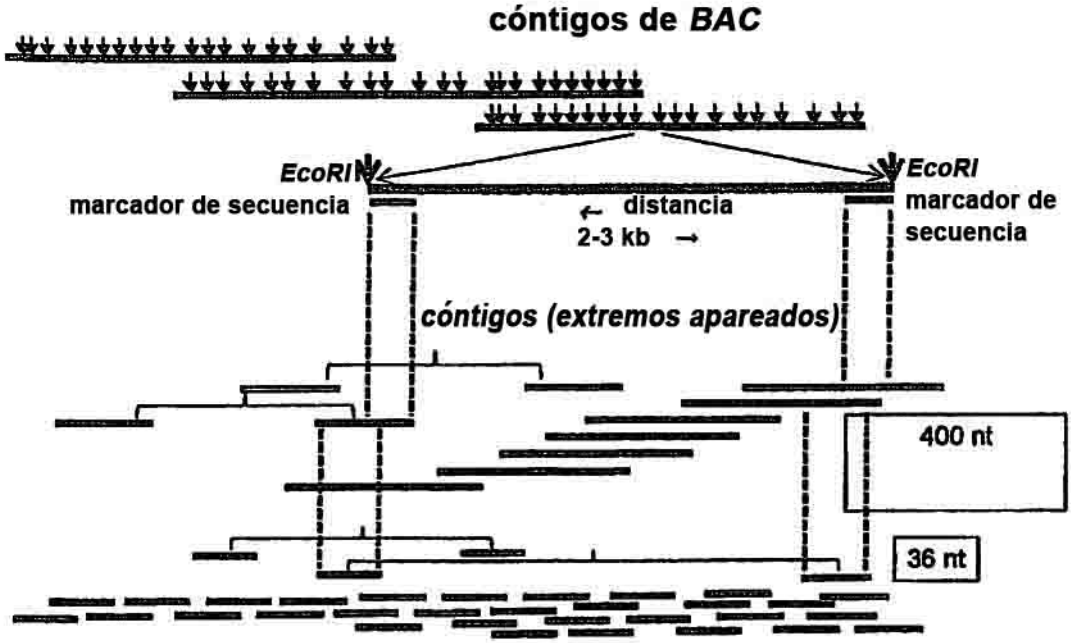
cóntigos de BAC
EcoRI
marcador de secuencia
distancia
2-3 kb
EcoRI
marcador de secuencia
cóntigos (extremos apareados)
400 nt
36 nt

FIG 2

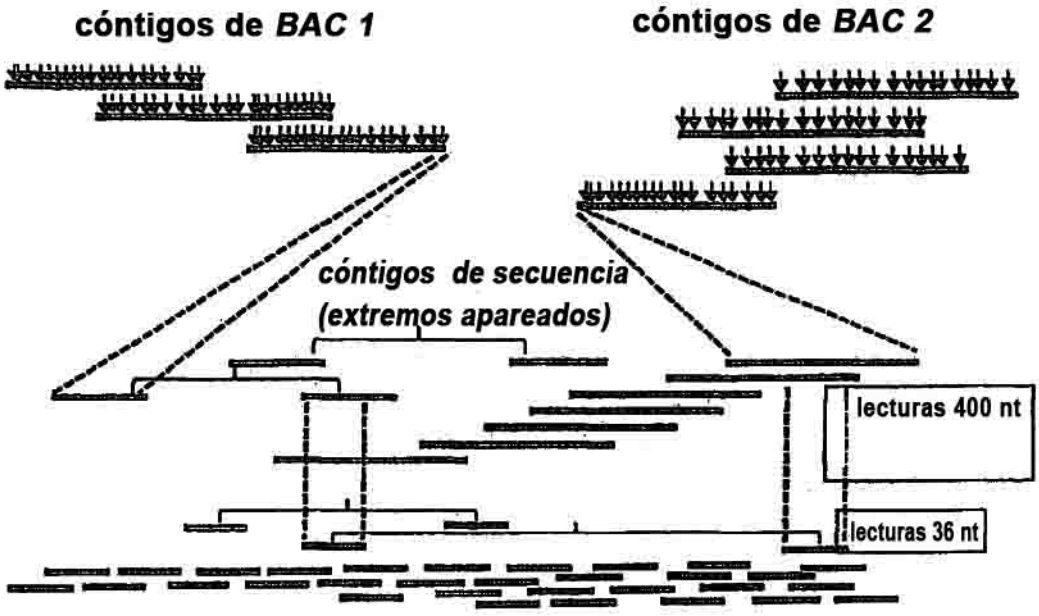
cóntigos de *BAC 1*
cóntigos de *BAC 2*
cóntigos de secuencia
(extremos apareados)
lecturas 400 nt
lecturas 36 nt

FIG 3

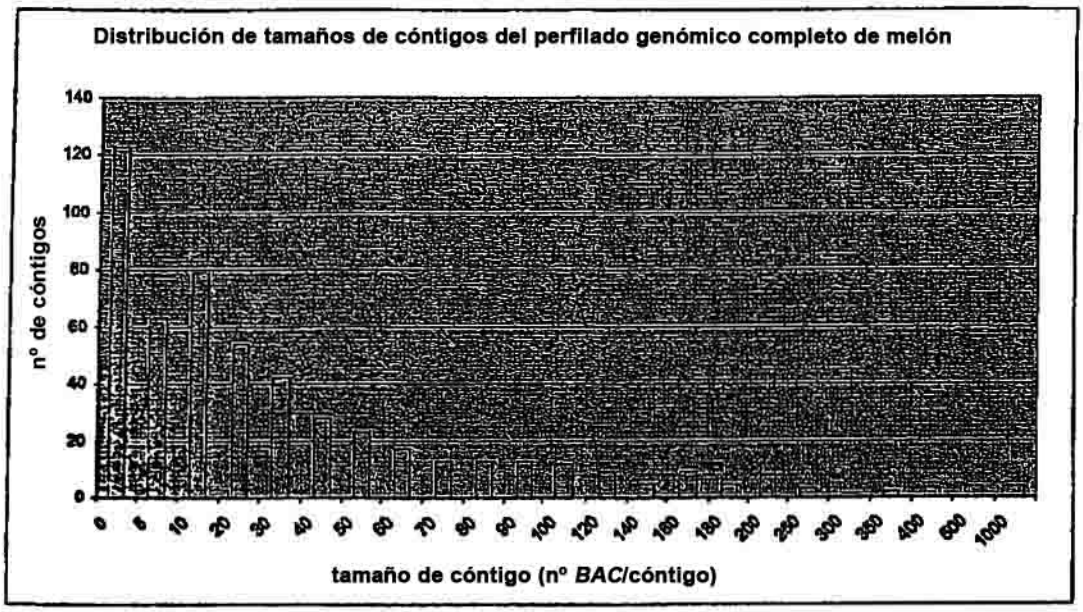
Distribución de tamaños de cóntigos del perfilado genómico completo de melón
nº de cóntigos
0
20
40
60
80
100
120
140
0
5
10
20
30
40
50
60
70
80
90
100
120
140
160
180
200
250
300
350
400
600
1000
tamaño de cóntigo (nº BAC/cóntigo)

FIG 4

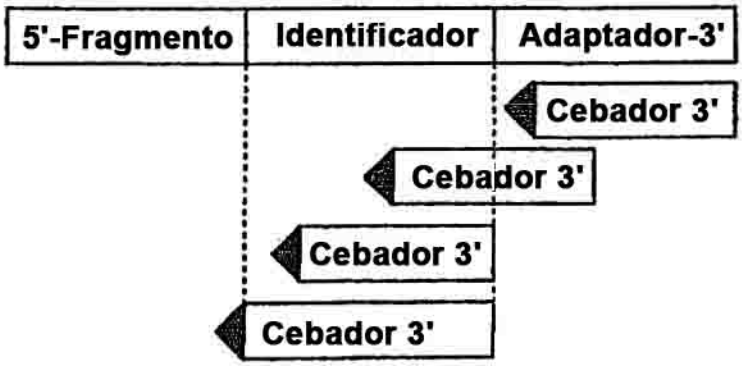
5'-Fragmento
Identificador
Adaptador-3'
Cebador 3'
Cebador 3'
Cebador 3'
Cebador 3'

FIG 5

FIG 6