(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 683 707**

(51) Int. CI.:

C12Q 1/68 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **02.05.2013 PCT/US2013/039295**

(87) Fecha y número de publicación internacional: **07.11.2013 WO13166303**

(96) Fecha de presentación y número de la solicitud europea: **02.05.2013 E 13784766 (1)**

(97) Fecha y número de publicación de la concesión europea: **11.07.2018 EP 2844772**

(54) Título: **Secuenciación de ADN**

(30) Prioridad:

02.05.2012 US 201261641715 P
15.03.2013 US 201361787437 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
27.09.2018

(73) Titular/es:

IBIS BIOSCIENCES, INC. (100.0%)
Suite 150 2251 Faraday Avenue
Carlsbad, CA 92008, US

(72) Inventor/es:

ESHOO, MARK W.

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCION

Secuenciación de ADN

CAMPO DE LA INVENCIÓN

En la presente se proporciona tecnología relacionada con la secuenciación de ácidos nucleicos y particularmente, pero no exclusivamente, con métodos, composiciones y sistemas para secuenciar un ácido nucleico usando un código de dos bases degenerado.

ANTECEDENTES

La secuenciación del ADN está impulsando la investigación y el descubrimiento genómicos. La finalización del Proyecto del Genoma Humano fue un logro monumental que involucró una cantidad increíble de esfuerzos combinados entre centros de genoma y científicos de todo el mundo. Este proyecto de una década se completó usando el método de secuenciación de Sanger para determinar el orden de las cuatro bases de nucleótidos: adenina, guanina, citosina y timina en moléculas de ADN. Este método sigue siendo la metodología de secuenciación del genoma principal en centros de secuenciación de genoma de alto rendimiento. Además, se han establecido muchas plataformas de secuenciación de "próxima generación" como alternativas prácticas al método de Sanger y se usan ampliamente. Estas incluyen enfoques de secuenciación por síntesis (SBS) como pirosecuenciación (Ronaghi et al. (1998) Science 281: 363-365), secuenciación de moléculas de ADN individuales (Braslaysky et al. (2003) Proc. Natl. Acad. Sci. USA 100: 3960-3964), y colonias de polimerasa (secuenciación de "polonias") (Mitra et al. (2003) Anal. Biochem. 320: 55-65). Aunque las tecnologías fundamentales de los varios métodos de secuenciación existentes y emergentes pueden diferir enormemente, los métodos de secuenciación convencionales comparten la característica de proporcionar una secuencia en términos de las cuatro bases de nucleótidos adenina, guanina, citosina y timina (o, en ARN, uracilo).

SUMARIO

Los métodos convencionales de secuenciación por síntesis se basan en identificar diferencialmente las cuatro bases A, C, G y T que se incorporan en un ácido nucleico durante cada evento de incorporación de base durante la síntesis. Por el contrario, la tecnología actual se basa en la secuenciación de ácidos nucleicos utilizando un código de dos bases degenerado. Por ejemplo, en lugar de determinar la secuencia de las cuatro bases en un ácido nucleico, la presente tecnología en algunas realizaciones determina el orden de las bases de purina y pirimidina en un ácido nucleico. Usando un esquema de secuenciación de acuerdo con este enfoque ejemplar, la secuencia ACGT convencionalmente derivada se obtendría en cambio determinando que la secuencia consiste de una purina en la primera posición, una pirimidina en la segunda posición, una purina en la tercera posición y una pirimidina en la cuarta posición, lo que se puede representar como RYRY. Un esquema de secuenciación de dos bases alternativo basado en la identificación de la secuencia de bases ceto y bases amino produce la secuencia de MMKK para esta misma secuencia de cuatro bases de bases ACGT. En algunas realizaciones, la información de las dos secuencias de dos bases puede fusionarse para producir una secuencia de cuatro bases convencional. De acuerdo con el ejemplo actual, la primera posición es una base amino purina, la segunda posición es una base amino pirimidina, la cuarta posición es una base ceto purina, y la cuarta posición es una base ceto pirimidina, que conduce inequívocamente a la secuencia ACGT.

Como consecuencia, las realizaciones de la tecnología requieren menos flujos de soluciones de nucleótidos y/o pasos de lavado para cada ciclo de síntesis, lo que también reduce el tiempo para adquirir una secuencia y reduce la complejidad y el coste de los aparatos usados para los tipos de esquemas de secuenciación descritos en la presente. Además, algunas realizaciones de la tecnología reducen el número de tintes fluorescentes necesarios para la secuenciación, reduciendo también de este modo el número de láseres usados para excitar marcadores (por ejemplo, fracciones fluorescentes), reduciendo o eliminando la óptica usada para dividir la señal óptica por longitud de onda, y reduciendo del número de detectores para registrar eventos de incorporación y diferenciar entre bases.

Por consiguiente, en la presente se proporcionan métodos para secuenciar un ácido nucleico objetivo, el método comprendiendo elegir un código degenerado de dos bases; y determinar una secuencia degenerada de dos bases del ácido nucleico objetivo usando el código degenerado de dos bases. Un código degenerado de dos bases puede basarse en varias clasificaciones y propiedades de las bases convencionales A, C, G y T (o U). Por ejemplo, en algunas realizaciones, el código degenerado de dos bases representa el orden de las bases de purina y las bases de pirimidina (por ejemplo, como R e Y); en algunas realizaciones, el código degenerado de dos bases representa el orden de las bases ceto y las bases amino (por ejemplo, K y M); y en algunas realizaciones, el código degenerado de dos bases representa el orden de bases fuertemente enlazadoras de hidrógeno y bases débilmente enlazadoras de hidrógeno (por ejemplo, S y W). Las realizaciones particulares proporcionan: 1) que el código degenerado de dos bases relaciona un primer elemento con una base que comprende adenina (A) o guanina (G) y un segundo elemento con una base que comprende citosina (C) o timina (T); 2) que el código degenerado de dos bases relaciona un primer elemento con una base que comprende A o C y un segundo elemento con una base que comprende G o T; y

3) que el código degenerado de dos bases relaciona un primer elemento con una base que comprende G o C y un segundo elemento con una base que comprende A o T.

Las secuencias que usan diferentes códigos degenerados de dos bases pueden usarse en combinación para derivar una secuencia de nucleótidos de cuatro bases estándar para un ácido nucleico. Por consiguiente, algunas realizaciones de la tecnología proporcionan un método que comprende fusionar una primera secuencia degenerada de dos bases y una segunda secuencia degenerada de dos bases para producir una secuencia de cuatro bases.

Las realizaciones de los métodos de acuerdo con la tecnología comprenden proporcionar un primer nucleótido y un segundo nucleótido en donde el primer nucleótido está marcado con un marcador y el segundo nucleótido está marcado con dicho marcador. Además, algunas realizaciones comprenden proporcionar un primer nucleótido, un segundo nucleótido, un tercer nucleótido y un cuarto nucleótido, en donde el primer nucleótido está marcado con un primer marcador, el segundo nucleótido está marcado con dicho primer marcador, el tercer nucleótido está marcado con un segundo marcador, y el cuarto nucleótido está marcado con dicho segundo marcador.

Adicionalmente, algunas realizaciones comprenden proporcionar un análogo de nucleótido marcado en donde el análogo de nucleótido marcado se empareja con un primer nucleótido o un segundo nucleótido. Además, las realizaciones también comprenden proporcionar un primer análogo de nucleótido marcado y un segundo análogo de nucleótido marcado en donde la base del primer análogo de nucleótido marcado se empareja con un primer nucleótido o un segundo nucleótido y la base del segundo análogo de nucleótido marcado se empareja con un tercer nucleótido o un cuarto nucleótido.

En algunas realizaciones, determinar una secuencia degenerada de dos bases del ácido nucleico objetivo usando el código degenerado de dos bases comprende medir una característica física, química y/o electrónica de una base y diferenciar entre una base de purina y una base de pirimidina, entre una base ceto y una base amino, y/o entre una base fuertemente enlazadora de hidrógeno (por ejemplo, un par de bases que consisten de tres pares de enlaces de hidrógeno) y una base débilmente enlazadora de hidrógeno (por ejemplo, un par de bases que consisten de dos pares de enlaces de hidrógeno).

En algunas realizaciones, la secuencia de dos bases del ácido nucleico objetivo se compara con una secuencia conocida, por ejemplo, para detectar un cambio en la secuencia de nucleótidos (por ejemplo, un polimorfismo de nucleótido individual, una inserción, una deleción, una variación del sitio de empalme, una transición, una transversión, una mutación de sentido erróneo, una mutación sin sentido, etc.). En algunas realizaciones, la secuencia conocida identifica todas las bases (a, t, c, g. y u) y se convierte (por ejemplo, mediante un ordenador) en un código de 2 bases.

También se proporcionan composiciones relacionadas con la secuenciación de un ácido nucleico usando un código degenerado de dos bases. Por ejemplo, algunas realizaciones proporcionan una composición que comprende un primer nucleótido y un segundo nucleótido en donde el primer nucleótido está marcado con un primer marcador y el segundo nucleótido está marcado con dicho primer marcador. En algunas realizaciones, el marcador es una fracción fluorescente. Algunas realizaciones de las composiciones proporcionan cuatro nucleótidos para la secuenciación usando un código de dos bases degenerado. En particular, las realizaciones proporcionan un tercer nucleótido y un cuarto nucleótido, en donde el tercer nucleótido está marcado con un segundo marcador y el cuarto nucleótido está marcado con dicho segundo marcador. En algunas realizaciones, el primer nucleótido es una A, el segundo nucleótido es una G, el tercer nucleótido es una C, y el cuarto nucleótido es una T. En algunas realizaciones, el primer nucleótido es una A, el segundo nucleótido es una C , el tercer nucleótido es una G, y el cuarto nucleótido es una T. Además, en algunas realizaciones, el primer nucleótido es una C, el segundo nucleótido es una G, el tercer nucleótido es una A, y el cuarto nucleótido es una T.

Las composiciones proporcionadas en la presente se refieren a la secuenciación de un ácido nucleico; como tal, la tecnología incluye realizaciones de composiciones que comprenden un ácido nucleico objetivo, un cebador de secuenciación, y una polimerasa. Tras la incorporación de un nucleótido, por ejemplo, en una reacción de secuenciación, las composiciones en algunas realizaciones comprenden un ácido nucleico que comprende el primer nucleótido y/o el segundo nucleótido.

Los métodos y composiciones de la tecnología encuentran uso en sistemas para secuenciar un ácido nucleico usando un código degenerado de dos bases. En un aspecto, la tecnología proporciona realizaciones de un sistema para secuenciar un ácido nucleico, el sistema comprendiendo un aparato de secuenciación y una funcionalidad para diferenciar un primer nucleótido y un segundo nucleótido de un tercer nucleótido y un cuarto nucleótido. En algunas realizaciones, el sistema comprende además una funcionalidad de producción para proporcionar una secuencia de nucleótidos de dos bases degenerada del ácido nucleico. Las secuencias que usan diferentes códigos de dos bases degenerados pueden fusionarse para proporcionar un código de cuatro bases para un ácido nucleico; es decir, algunas realizaciones comprenden una funcionalidad para fusionar una primera

secuencia de nucleótidos de dos bases degenerada del ácido nucleico y una segunda secuencia de nucleótidos de dos bases degenerada del ácido nucleico para proporcionar una secuencia de cuatro bases del ácido nucleico. Adicionalmente, las realizaciones de la tecnología se refieren a un sistema en el que la funcionalidad para diferenciar un primer nucleótido y un segundo nucleótido de un tercer nucleótido y un cuarto nucleótido diferencia entre una base de purina y una base de pirimidina, entre una base ceto y una base amino, y/o entre una base fuertemente enlazadora de hidrógeno y una base débilmente enlazadora de hidrógeno.

En un aspecto, el código de 2 bases se determina (por ejemplo, mediante secuenciación) sin determinar y/o conocer de otra manera el código de 4 bases.

Se proporcionan realizaciones de kits, por ejemplo, un kit para secuenciar un ácido nucleico, el kit comprendiendo un primer nucleótido, un segundo nucleótido, un tercer nucleótido y un cuarto nucleótido, en donde el primer nucleótido está marcado con un primer marcador, el segundo nucleótido está marcado con dicho primer marcador, el tercer nucleótido está marcado con un segundo marcador, y el cuarto nucleótido está marcado con dicho segundo marcador; o un primer análogo de nucleótido degenerado de dos bases y un segundo análogo de nucleótido degenerado de dos bases, en donde el primer análogo de nucleótido está marcado con un primer marcador y el segundo análogo de nucleótido está marcado con un segundo marcador. Realizaciones adicionales serán evidentes para los expertos en la técnica relevante en base a las enseñanzas contenidas en la presente.

BREVE DESCRIPCIÓN DEL DIBUJO

Estas y otras características, aspectos y ventajas de la presente tecnología se comprenderán mejor con respecto a los siguientes dibujos :
La Figura 1A muestra una secuencia de cuatro bases convencional del gen gliceraldehído 3-fosfato deshidrogenasa de *Homo sapiens*. La figura 1B muestra esta secuencia representada usando un código degenerado de dos bases de "r" e "y" que denota el orden de purinas y pirimidinas, respectivamente. La Figura 1C muestra esta secuencia representada usando un código degenerado de dos bases de "m" y "k" que denota el orden de las bases amino y ceto, respectivamente. En la Figura 1, r = A o G, y = C o T, m = A o C, y k = G o T.

DESCRIPCIÓN DETALLADA

En la presente se proporciona tecnología relacionada con la secuenciación de ácidos nucleicos y particularmente, pero no exclusivamente, con métodos, composiciones, sistemas y kits para secuenciar un ácido nucleico usando un código de dos bases degenerado.

Los encabezados de las secciones usados en la presente son con propósitos de organización solamente y no deben interpretarse como limitativos de la materia descrita de ninguna manera.

En esta descripción detallada de las varias realizaciones, con propósitos de explicación, se exponen numerosos detalles específicos para proporcionar una comprensión exhaustiva de las realizaciones divulgadas. Un experto en la técnica apreciará, sin embargo, que estas varias realizaciones pueden ponerse en práctica con o sin estos detalles específicos. En otros casos, las estructuras y los dispositivos se muestran en forma de diagramas de bloques. Además, un experto en la técnica puede apreciar fácilmente que las secuencias específicas en las que se presentan y realizan los métodos son ilustrativas y se contempla que las secuencias pueden variarse y permanezcan todavía dentro del espíritu y el alcance de las diversas realizaciones divulgadas en la presente.

A menos que se defina lo contrario, todos los términos técnicos y científicos usados en la presente tienen el mismo significado que el entendido comúnmente por un experto en la técnica a la que pertenecen las varias realizaciones descritas en la presente.

Se apreciará que hay un "aproximadamente" implícito antes de las temperaturas, concentraciones, tiempos, etc. tratados en las presentes enseñanzas, de tal manera que las desviaciones insustanciales están dentro del alcance de las presentes enseñanzas. En esta solicitud, el uso del singular incluye el plural a menos que se indique específicamente lo contrario. Además, el uso de "comprender", "comprende", "comprendiendo", "contener", "contiene", "conteniendo", "incluir", "incluye" e "incluyendo" no se pretende que sea limitativo. Debe entenderse que tanto la descripción general anterior como la siguiente descripción detallada son ejemplares y explicativas solamente y no son restrictivas de las presentes enseñanzas.

Además, a menos que el contexto requiera lo contrario, los términos singulares incluirán el plural y los términos plurales incluirán el singular. Generalmente, las nomenclaturas utilizadas en conexión con, y las técnicas de, cultivo celular y tisular, biología molecular y química e hibridación de proteínas y oligonucleótidos o polinucleótidos descritas en la presente son las bien conocidas y usadas comúnmente en la técnica. A menos que se indique lo contrario, se usan técnicas estándar, por ejemplo, para la purificación y preparación de ácidos nucleicos, análisis químico, ácido nucleico recombinante, y síntesis de oligonucleótidos. Las reacciones enzimáticas y las técnicas de purificación se realizan de acuerdo con las especificaciones del fabricante o como se realiza

comúnmente en la técnica o como se describe en la presente. Las técnicas y procedimientos descritos en la presente se realizan generalmente de acuerdo con métodos convencionales bien conocidos en la técnica y como se describe en diversas referencias generales y más específicas que se citan y tratan a lo largo de la presente especificación. Ver, por ejemplo, Sambrook et al., Molecular Cloning: A Laboratory Manual (Tercera edición, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (2000)). Las nomenclaturas utilizadas en conexión con, y los procedimientos y técnicas de laboratorio descritos en la presente son los bien conocidos y usados comúnmente en la técnica.

Definiciones

Para facilitar una comprensión de la presente tecnología, se definen a continuación una serie de términos y frases. A lo largo de la descripción detallada se exponen definiciones adicionales.

A lo largo de la especificación y las reivindicaciones, los siguientes términos toman los significados explícitamente asociados en la presente, a menos que el contexto indique claramente lo contrario. La frase "en una realización" como se usa en la presente no se refiere necesariamente a la misma realización, aunque puede. Además, la frase "en otra realización" como se usa en la presente no se refiere necesariamente a una realización diferente, aunque puede. Por tanto, como se describe a continuación, pueden combinarse fácilmente varias realizaciones de la invención, sin apartarse del alcance o el espíritu de la invención.

Adicionalmente, como se usa en la presente, el término "o" es un operador "o" inclusivo y es equivalente al término "y/o" a menos que el contexto indique claramente lo contrario. El término "basado en" no es exclusivo y permite basarse en factores adicionales no descritos, a menos que el contexto indique claramente lo contrario. Adicionalmente, a lo largo de la especificación, el significado de "un" y "el" incluye referencias en plural. El significado de "en" incluye "en" y "sobre".

Un "sistema" denota un conjunto de componentes, reales o abstractos, que comprenden un todo donde cada componente interactúa o está relacionado con por lo menos otro componente dentro del todo.

Como se usa en la presente, la frase "dNTP" significa desoxinucleotidotrifosfato, donde el nucleótido comprende una base de nucleótidos, como A, T, C, G o U. Además, se pretende que el término "dNTP" haga referencia también a desoxinucleotidostrifosfatos que comprenden bases modificadas y análogos de bases que son capaces de imitar el emparejamiento de bases de A, C, G, T o U, o que son capaces de emparejar bases en un modo degenerado, por ejemplo, una base que se empareja con A o G, C o T, A o C, G o T, G o C, o A o T.

El término "monómero" como se usa en la presente significa cualquier compuesto que pueda incorporarse a una cadena molecular en crecimiento por una polimerasa dada. Tales monómeros incluyen, sin limitaciones, nucleótidos de origen natural (por ejemplo, ATP, GTP, TTP, UTP, CTP, dATP, dGTP, dTTP, dUTP, dCTP, análogos sintéticos), precursores para cada nucleótido, nucleótidos de origen no natural y sus precursores, o cualquier otra molécula que pueda incorporarse en una cadena polimérica en crecimiento por una polimerasa dada.

Como se usa en la presente, un "ácido nucleico" significará cualquier molécula de ácido nucleico, incluyendo, sin limitación, ADN, ARN e híbridos de los mismos. Las bases de ácido nucleico que forman las moléculas de ácido nucleico pueden ser las bases A, C, G, T y U, así como derivados y análogos de las mismas. Los derivados de estas bases son bien conocidos en la técnica. Debe entenderse que el término incluye, como equivalentes, análogos de o ADN o ARN elaborados a partir de análogos de nucleótidos. El término como se usa en la presente también abarca ADNc, que es ADN complementario, o copia, producido a partir de una plantilla de ARN, por ejemplo mediante la acción de la transcriptasa inversa. Es bien sabido que el ADN (ácido desoxirribonucleico) es una cadena de nucleótidos que consta de 4 tipos de nucleótidos- A (adenina), T (timina), C (citosina), y (G) guanina) y el ARN (ácido ribonucleico) es una cadena de nucleótidos que consiste de 4 tipos de nucleótidos- A, U (uracilo), G y C. También se sabe que todos estos 5 tipos de nucleótidos se unen específicamente entre sí en combinaciones denominadas emparejamiento de bases complementarias. Es decir, la adenina (A) se empareja con la timina (T) (en el caso del ARN, sin embargo, la adenina (A) se empareja con uracilo (U)) y la citosina (C) se empareja con guanina (G), de tal manera que cada uno de estos pares de bases forman una cadena doble. Como se usa en la presente, "datos de secuenciación de ácido nucleico", "información de secuenciación de ácido nucleico", "secuencia de ácido nucleico", "secuencia genómica", "secuencia genética", "secuencia de fragmento", o "lectura de secuenciación de ácido nucleico" denota cualquier información o dato que sean indicativos del orden de las bases de nucleótidos en una molécula (por ejemplo, un genoma completo, un transcriptoma completo, un exoma, oligonucleótido, polinucleótido, fragmento, etc.) de ADN o ARN usando un código de cuatro bases (por ejemplo, usando A, G, C y T o U para representar las cuatro bases adenina, guanina, citosina, y timina o uracilo) o un código degenerado de dos bases para representar las bases de purina y pirimidina; las bases ceto y amino; y/o bases fuertemente enlazadoras de hidrógeno y débilmente enlazadoras de hidrógeno.

Se usan en la presente los códigos degenerados de IUB para bases de nucleótidos. En este código, R significa cualquiera de las bases de purina A o G; Y significa cualquiera de las bases de pirimidina C o T; M significa

cualquiera de las bases amino A o C; K significa cualquiera de las bases ceto G o T; S significa cualquiera de los socios de enlace de hidrógeno más fuertes C o G; y W significa cualquiera de los socios de enlace de hidrógeno más débiles A o T.

La referencia a una base, un nucleótido o a otra molécula puede ser en singular o en plural. Es decir, una base puede referirse a una única molécula de esa base o a una pluralidad de esa base, por ejemplo, en una solución.

Como se usa en la presente, un "polinucleótido", también denominado un ácido nucleico, es una serie covalentemente enlazada de nucleótidos en donde la posición 3' de la pentosa de un nucleótido está unida por un grupo fosfodiéster a la posición 5' de la siguiente. El ADN (ácido desoxirribonucleico) y el ARN (ácido ribonucleico) son polinucleótidos que se producen biológicamente en los que los residuos de nucleótidos están enlazados en una secuencia específica mediante enlaces de fosfodiéster. Como se usa en la presente, los términos "polinucleótido" u "oligonucleótido" abarcan cualquier compuesto polimérico que tenga una cadena principal lineal de nucleótidos. Un "oligodesoxirribonucleótido" u "oligonucleótidos", también denominado un "oligómero", es generalmente un polinucleótido de una longitud más corta.

En esta divulgación, se entiende que "ADN", "oligonucleótido" o "ácido nucleico" incluye ADN y ARN, así como derivados en los que el azúcar está modificado, como en derivados de 2'-O-metilo y 2', 3'-didesoxinucleósidos, en los que la nucleobase tiene un apéndice, y estos ácidos nucleicos y sus análogos en topologías no lineales, incluidos como dendrímeros, estructuras de peine y nanoestructuras, y análogos que llevan apéndices o etiquetas (por ejemplo, fluorescentes, funcionalizadas o vinculantes, como biotina).

Como se usa en la presente, la frase "una pluralidad clonal de ácidos nucleicos" o "una población clonal de ácidos nucleicos" o "una agrupación" o "una polonia" se refiere a un conjunto de productos de ácido nucleico que son sustancial o completa o esencialmente idénticos a entre sí, y son copias complementarias de la cadena de ácido nucleico plantilla a partir de la que se sintetizan.

Como se usa en la presente, un "análogo de nucleótido de dos bases" es un análogo de nucleótido que puede formar pares de bases con dos bases de nucleótidos diferentes del conjunto A, C, G y T (o U).

Como se usa en la presente, "complementario" se refiere generalmente a formar dúplex de nucleótidos específicos para formar pares de bases de Watson-Crick canónicas, como es entendido por los expertos en la técnica. Sin embargo, complementario también incluye emparejamiento de bases de nucleótidos modificados y análogos de nucleótidos que son capaces de formar un emparejamiento de bases degenerado o universal con nucleótidos A, T, G o C y/o con ácidos nucleicos bloqueados que mejoran la estabilidad térmica de los dúplex. Un experto en la técnica reconocerá que la rigurosidad de la hibridación es un determinante en el grado de coincidencia o falta de coincidencia en el dúplex formado mediante hibridación.

Como se usa en la presente, "fracción" se refiere a una de dos o más partes en las que algo se puede dividir, como, por ejemplo, las varias partes de una cadena, una molécula o una sonda.

Una "polimerasa" es una enzima generalmente para unir nucleótidos, oligómeros de 3'-OH 5 '-trifosfato y análogos de los mismos. Además, se pretende que en esta solicitud "polimerasa" incluya ADN polimerasas de todas las familias, ARN polimerasas, y transcriptasas inversas.

El término "cebador" se refiere a un oligonucleótido, ya sea de origen natural como en un digesto de restricción purificado o producido sintéticamente, que es capaz de actuar como un punto de inicio de la síntesis cuando se coloca en condiciones en las que se induce la síntesis de un producto de extensión del cebador que es complementario a una cadena de ácido nucleico, (por ejemplo, en presencia de nucleótidos y un agente inductor como ADN polimerasa y a una temperatura y pH adecuados). El cebador es preferiblemente de cadena sencilla para una máxima eficiencia en la amplificación, pero alternativamente puede ser de cadena doble. Si es de cadena doble, primero se trata el cebador para separar sus cadenas antes de ser usado para preparar productos de extensión. Preferiblemente, el cebador es un oligodesoxirribonucleótido. El cebador debe ser lo suficientemente largo para cebar la síntesis de los productos de extensión en presencia del agente inductor. Las longitudes exactas de los cebadores dependerán de muchos factores, incluyendo la temperatura, la fuente del cebador y el uso del método.

Como se usa en la presente, "degeneración" o "degenerado" se refiere a ciertas equivalencias con respecto al código genético estándar de cuatro bases de nucleótidos A, C, G y T. En algunos contextos, un "código degenerado" es aquel en que un símbolo, carácter, color, etc. se refiere a más de una de los bases A, C, G y T (o U). Un código de dos bases degenerado es uno en el que el conjunto de símbolos que representa una secuencia de ácido nucleico tiene dos elementos y un elemento se refiere a cualquiera y/o a ambas de dos bases y el segundo elemento se refiere a cualquiera y/o a ambas de otras dos bases (es decir, no hay superposición entre el conjunto de dos bases denotado por el primer elemento y el conjunto de dos bases denotado por el segundo elemento). Ejemplos de códigos degenerados son el código de purina/pirimidina en el que R se refiere a A o G e Y se refiere a

C o T; el código de la base ceto/base amino en el que K se refiere a G o T y M se refiere a A o C; y el código fuerte/débil en el que S se refiere a C o G y W se refiere a A o T.

En algunos contextos, "degenerado" se refiere al comportamiento de emparejamiento de bases de una base de nucleótidos o análogo de bases de nucleótidos. El emparejamiento de bases degenerado se refiere a una situación en la que un nucleótido o análogo de nucleótido puede formar pares de bases con más de una socio. En algunos contextos, una "regla de emparejamiento de bases degenerada" describe o define el conjunto de socios de emparejamiento de bases con el que un nucleótido o un análogo de nucleótido forma pares de bases. Por ejemplo, una regla de emparejamiento de bases degenerada puede describir un nucleótido o análogo de nucleótido que se empareja con tanto A como G, tanto C como T, tanto A como C, tanto G como T, tanto G como C, y/o ambos de A y T.

Realizaciones de la tecnología

La tecnología se refiere de manera general a métodos, composiciones, sistemas y kits para la secuenciación de ADN usando un código degenerado de dos bases en, por ejemplo, un enfoque de secuenciación por síntesis. Aunque la divulgación de la presente se refiere a ciertas realizaciones ilustradas, debe entenderse que estas realizaciones se presentan a modo de ejemplo y no a modo de limitación.

1. Métodos

Algunas realizaciones de la tecnología proporcionan métodos de secuenciación de ADN que usan un código de dos bases degenerado para identificar las bases en la secuencia. La tecnología abarca varias realizaciones de esquemas de secuenciación de dos bases degeneradas que identifican bases por rasgos compartidos por pares de las cuatro bases A, C, G y T (alternativamente, U). Por ejemplo, en algunas realizaciones, los métodos se basan en diferenciar bases de pirimidina (C y T) de bases de purina (A y G), produciendo una secuencia que denota pirimidinas con Y y purinas con R. En algunas realizaciones, los métodos se basan en diferenciar bases ceto (G y T) de bases amino (A y C), produciendo una secuencia que denota bases ceto con K y bases amino con M. En algunas realizaciones, los métodos se basan en diferenciar bases que forman pares de bases más fuertes (G y C) de las que forman pares de bases más débiles (A y T), produciendo una secuencia que denota bases que forman pares de bases más fuertes con S y bases que forman pares de bases más débiles con W. Debe entenderse que no se pretende que las designaciones estándar asociadas con los códigos de base degenerados R, Y, K, M, S y W limiten la tecnología a producir secuencias representadas solo por estas letras o códigos particulares. La tecnología abarca métodos que usan un código de dos bases degenerado, independientemente de la anotación usada para comunicar la secuencia.

La tecnología contempla cualquier método de secuenciación mediante el que estos pares de nucleótidos se diferencian entre sí, por ejemplo, por características físicas y/o químicas como tamaño, carga, conductividad, características de fluorescencia inherentes, masa, momento dipolar, forma, estructura, reactividad, etc., y/o interrogando a cada nucleótido en la secuencia objetivo con alguna otra molécula, como monitorizando el emparejamiento de bases de cada nucleótido con nucleótidos etiquetados (por ejemplo, marcados), nucleótidos modificados etiquetados, análogos de nucleótidos etiquetados, etc.

En algunas realizaciones, se usa un método de secuenciación basado en conjuntos y en algunas realizaciones se usa un método de secuenciación basado en una única molécula. En algunas realizaciones, se detiene una reacción de secuenciación después de la incorporación de cada nucleótido y en algunas realizaciones la síntesis se monitoriza en tiempo real sin la necesidad de interrumpir la reacción para identificar bases. En algunas realizaciones, las moléculas de un ácido nucleico se interrogan directamente sin usar una reacción de secuenciación para identificar las bases. Con respecto a los métodos y esquemas de secuenciación por síntesis que encuentran uso, por ejemplo, adaptados apropiadamente a los métodos proporcionados en la presente, Morozova y Marra proporcionan una revisión de algunas de tales tecnologías en Genomics 92: 255 (2008); exposiciones adicionales se encuentran en Mardis, Annu. Rev. Genomics Hum. Genet. (2008) 9:387-402 and in Fuller, et al. (2009) Nat. Biotechnol. 27: 1013.

En un método basado en conjuntos, de decenas de miles a decenas de millones de cadenas nominalmente idénticas se localizan en una localización dada (por ejemplo, en una perla u otra superficie o sustrato sólido) para leer en un proceso que comprende iteraciones de lavado y escaneo. En el uso convencional, este proceso implica añadir reactivos (por ejemplo, nucleótidos marcados), incorporar nucleótidos en cadenas de ADN (por ejemplo, mediante una polimerasa), detener la reacción de incorporación, eliminar o inactivar el exceso de reactivo, identificar las bases incorporadas (por ejemplo, detección óptica de emisión de fluorescencia de un marcador de nucleótido; detectar un cambio en el pH o voltaje), y, en algunas realizaciones, tratar las bases recién incorporadas para preparar plantillas de ADN para la siguiente adición de base. Estos pasos continúan hasta que el proceso secuencia el nucleótido objetivo completo o no produce resultados de secuencia satisfactorios.

En general, los métodos basados en conjuntos dependen de detener la reacción de secuenciación después

de cada incorporación de bases para mantener la población de moléculas sintetizadas en fase de tal manera que la detección (por ejemplo, obtención de imágenes) informe con precisión de la base incorporada por la síntesis en cada paso. La separación de fases se mantiene en varias realizaciones añadiendo una base cada vez (ver, por ejemplo, Margulies, M. et al. "Genome sequencing in microfabricated highdensity picolitre reactors", Nature 437: 376-380 (2005); Harris, T.D. et al. "Single-molecule DNA sequencing of a viral genome", Science 320: 106-109 (2008)) o usando nucleótidos reversiblemente bloqueados que permiten una sola incorporación de bases durante cada iteración del ciclo.

Por ejemplo, algunas realizaciones comprenden el uso de tecnologías particulares para secuenciación en paralelo de amplicones particionados (Publicación de PCT Nº: WO 2006/084132); extensión de oligonucleótidos en paralelo (ver, por ejemplo, Patente de Estados Unidos Nº 5.750.341, Patente de Estados Unidos Nº 6.306.597); secuenciación de polonias Mitra et al. (2003) Analytical Biochemistry 320: 55-65; Shendure et al. (2005) Science 309: 1728-1732; Patente de Estados Unidos Nº 6.432.360, Patente de Estados Unidos Nº 6.485.944, Patente de Estados Unidos Nº 6.511.803;); la tecnología de adición de bases individuales Solexa (ver, por ejemplo, Bennett et al. (2005), Pharmacogenomics 6: 373-382; Patente de Estados Unidos Nº 6.787.308; Patente de Estados Unidos Nº 6.833.246; la tecnología de secuenciación de firma masivamente en paralelo de Lynx (Brenner et al., (2000) Nat. Biotechnol., 18: 630-634, Patente de Estados Unidos Nº 5.695.934, Patente de Estados Unidos Nº 5.714.330). y la tecnología de colonias de PCR Adessi (Adessi et al. (2000). Nucleic Acid Res. 28: E87; WO 00/018957).

En realizaciones particulares, la extensión se bloquea momentáneamente después de cada adición de bases usando nucleótidos modificados (por ejemplo, terminadores reversibles de nucleótidos como se describe en, por ejemplo, la WO 2004/018497, Publicación de Solicitud de patente de Estados Unidos Nº 2007/0166705; Bentley, D.R. et al. "Accurate whole human genome sequencing using reversible terminator chemistry", Nature 456: 53-59 (2008); Turcatti, G. et al. "A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis", Nucleic Acids Res. 36: e25 (2008); Guo, J. et al. "Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides", Proc. Natl. Acad. Sci. USA 105: 9145-9150 (2008); Ju, J. et al. "Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators", Proc. Natl. Acad. Sci. USA 103: 19635-19640 (2006); Seo, T.S. et al. "Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides", Proc. Natl. Acad. Sci. USA 102: 5926-5931 (2005); Wu, W. et al. "Termination of DNA synthesis by N6-alkylated, not 3'-O-alkylated, photocleavable 2'-deoxyadenosine triphosphates", Nucleic Acids Res. 35: 6339-6349 (2007)) u omitiendo los componentes de la reacción como iones metálicos divalentes (ver, por ejemplo, WO2005/123957; Publicación de Solicitud de Patente de Estados Unidos Nº 20060051807).

Las realizaciones de la presente tecnología se refieren a métodos de secuenciación de conjuntos en los que se añaden 1, 2, 3 ó 4 nucleótidos en cada ronda de secuenciación. En algunas realizaciones, se usan dos marcadores para marcar dos pares de nucleótidos. Es decir, se marcan dos nucleótidos con un primer marcador (por ejemplo, una primera fracción fluorescente) y los otros dos nucleótidos se marcan con un segundo marcador (por ejemplo, una segunda fracción fluorescente). Por ejemplo, en algunas realizaciones, las purinas A y G están marcadas con un primer marcador y las pirimidinas están marcadas con un segundo marcador; en algunas realizaciones, las bases ceto G y T están marcadas con un primer marcador y las bases amino A y C están marcadas con un segundo marcador; en algunas realizaciones, las bases fuertemente enlazadoras de hidrógeno C y G están marcadas con un primer marcador y las bases débilmente enlazadoras de hidrógeno A y T están marcadas con un segundo marcador. En realizaciones en las que se añade más de un tipo de nucleótido en cada ronda, los nucleótidos se bloquean reversiblemente, por ejemplo, con un terminador reversible, para detener la síntesis después de cada incorporación de uno de los nucleótidos añadidos.

En algunas realizaciones en las que se añade un nucleótido a la vez, se usan dos análogos de nucleótidos diferencialmente marcados en los que cada base de análogo de nucleótido se empareja con dos nucleótidos de acuerdo con una regla de emparejamiento de bases degeneradas. En particular, los dos nucleótidos con los que se empareja la primera base del análogo de nucleótidos son diferentes de los dos nucleótidos con los que se empareja el segundo análogo de nucleótidos (sin sobreponerlos dos conjuntos). Por ejemplo, en algunas realizaciones, se usan dos análogos de nucleótido X e Y en una secuenciación de conjuntos mediante reacción de síntesis en donde el hidrógeno de X se une con una purina y el hidrógeno Y se une con una pirimidina. Añadir secuencialmente X e Y a la reacción (por ejemplo, añadir X, luego añadir Y, luego añadir X, luego añadir Y, etc., opcionalmente con un paso de lavado después de cada adición) genera una secuencia de purinas (por ejemplo, R = A o G) y pirimidinas (por ejemplo, Y = C o T) del ácido nucleico plantilla (objetivo. Realizaciones similares comprenden añadir dos análogos de nucleótidos que se unen por hidrógeno de acuerdo con una regla degenerada en la que un análogo de nucleótido se une a una base amino (por ejemplo, M = A o C) y el otro análogo de nucleótido se une a una base ceto (por ejemplo, K = G o T). De manera similar, las realizaciones comprenden añadir dos análogos de nucleótidos que se unen por hidrógeno de acuerdo a una regla degenerada en la que un análogo de nucleótido se une a una base de enlace de nitrógeno fuerte (por ejemplo, S = C o G) y el otro análogo de nucleótido se une a una base de enlace de hidrógeno débil (por ejemplo, W = A o T).

Si la secuencia que se está determinando es desconocida, los nucleótidos o análogos de nucleótidos añadidos se aplican habitualmente en un orden elegido que luego se repite a lo largo del análisis. Si, sin embargo, la secuencia que se está determinando es conocida y se está re-secuenciado, por ejemplo, para determinar si las hay pequeñas diferencias en la secuencia en relación a la secuencia conocida, el proceso de determinación de la secuenciación se puede hacer más rápido añadiendo los nucleótidos en cada paso en el orden apropiado, por ejemplo, elegido de acuerdo con la secuencia conocida. Las diferencias de la secuencia dada se detectan por tanto por la falta de incorporación de ciertos nucleótidos en etapas particulares de la extensión del cebador.

Adicionalmente, las realizaciones de secuenciación basadas en moléculas individuales implican métodos que comprenden diferentes tecnologías fundamentales, por ejemplo, monitorizar una molécula de polimerasa a medida que incorpora nucleótidos en una cadena de ADN sintetizada; someter a pases una molécula de ácido nucleico (o sus monómeros de nucleótidos) a través o sobre o cerca de una estructura de sonda (por ejemplo, a través de un tubo o un poro) y monitorizar las interacciones de cada base de nucleótidos con la estructura de sonda (por ejemplo, un cambio en el voltaje, un cambio en la corriente, un cambio en las propiedades ópticas); observar la síntesis de una molécula de ADN directamente usando microscopía (por ejemplo, STM, TEM); u observando directamente una molécula de un ácido nucleico e identificando las bases individuales mediante observación directa.

Las realizaciones de los métodos proporcionados en la presente comprenden la secuenciación de moléculas individuales basada en un código degenerado de dos bases. Por ejemplo, en algunas realizaciones, se observa directamente una molécula de ADN y se discierne la secuencia de las bases de purina y pirimidina (o, alternativamente, las bases ceto y amino o fuertemente y débilmente enlazadoras de hidrógeno) en base a características físicas como la forma, tamaño y/o masa de cada base. Como otro ejemplo, en algunas realizaciones una molécula de un ácido nucleico se enhebra a través de un nanoporo y se discierne la secuencia de bases ceto y amino (o, alternativamente, las bases de purina y pirimidina o las bases fuertemente y débilmente enlazadoras de hidrógeno) por los diferentes cambios en la corriente y/o potencial a través del nanoporo inducidos por bases ceto y amino.

Algunas realizaciones de secuenciación de moléculas individuales en las que se monitoriza la síntesis (por ejemplo, mediante observación directa, detectando cambios en la fluorescencia, etc.) usan dos marcadores para marcar pares de nucleótidos como se ha descrito anteriormente para los métodos de conjuntos. En particular, estas realizaciones comprenden usar un primer par de nucleótidos marcados con un primer marcador (por ejemplo, una primera fracción fluorescente) y un segundo par de nucleótidos marcados con un segundo marcador (por ejemplo, una segunda fracción fluorescente) y/o un par de análogos de nucleótidos marcados en los que cada base de análogo de nucleótido se empareja con dos nucleótidos de acuerdo con una regla de emparejamiento de bases degeneradas (por ejemplo, como se ha tratado anteriormente para las realizaciones de conjuntos).

Durante cada ciclo, la detección de una señal de salida apropiada para la base añadida en el paso anterior indica una incorporación exitosa de esa base y por tanto identifica la base incorporada en ese paso. La detección puede ser por modos convencionales. Por ejemplo, si el marcador es una fracción fluorescente, entonces la detección de una base incorporada puede llevarse a cabo usando un microscopio de escaneo confocal para escanear la colección de agrupaciones (por ejemplo, unidas a una superficie) con un láser para obtener imágenes de las fracciones fluorescentes unidas directamente a las bases incorporadas. Alternativamente, puede usarse un detector 2D sensible, como un detector de carga acoplada (CCD), para visualizar las señales generadas. Sin embargo, hay disponibles otras técnicas como la microscopía óptica de barrido de campo cercano (SNOM) y pueden usarse cuando se obtienen imágenes de matrices densas. Por ejemplo, usando SNOM, pueden distinguirse polinucleótidos individuales cuando están separados por una distancia de menos de 100 nm, por ejemplo de 10 nm a 10 fm. Para una descripción de la microscopía óptica de barrido de campo cercano, ver Moyer et al., Laser Focus World (1993) 29:10. Se conocen aparatos adecuados usados para obtener imágenes de matrices de polinucleótidos y la configuración técnica es evidente para el experto en la técnica. La detección se usa preferiblemente en combinación con un sistema de análisis para determinar el número y la naturaleza de las bases de nucleótidos incorporadas para cada paso de la síntesis. Este análisis, que puede llevarse a cabo inmediatamente después de cada paso de síntesis, o más tarde utilizando datos registrados, permite determinar la secuencia de la plantilla de ácido nucleico.

Ejemplos de tecnologías de secuenciación para las que la presente tecnología es apropiada y/o para las que se adapta la presente tecnología se tratan a continuación. En algunas realizaciones, se usan métodos de pirosecuenciación. En la pirosecuenciación (Voelkerding et al., Clinical Chem., 55: 641-658, 2009; MacLean et al., Nature Rev. Microbiol., 7: 287-296; Patente de Estados Unidos Nº 6.210.891; Patente de Estados Unidos Nº 6.258.568; el ADN plantilla se fragmenta, se repara al final, se liga a adaptadores y se amplifica clonalmente in situ capturando moléculas plantilla individuales con perlas portadoras que llevan oligonucleótidos complementarios a los adaptadores. Cada perla que lleva un único tipo de plantilla se compartimentaliza en una microvesícula de agua en aceite, y la plantilla se amplifica clonalmente usando una técnica referida como PCR en emulsión. La emulsión se interrumpe después de la amplificación y las perlas se depositan en pocillos individuales de una placa picotituladora que funciona como una celda de flujo durante las reacciones de secuenciación. La introducción iterativa ordenada de cada uno de los cuatro reactivos dNTP tiene lugar en la celda de flujo en presencia de enzimas de secuenciación y

un indicador luminiscente como luciferasa. En el caso de que se añada un dNTP apropiado al extremo 3' del cebador de secuenciación, la producción resultante de ATP provoca una explosión de luminiscencia dentro del pocillo, que se graba usando una cámara CCD. Es posible lograr longitudes de lectura mayores o iguales a 400 bases, y se pueden lograr 10^6 lecturas de secuencia, dando como resultado hasta 500 millones de pares de bases (Mb) de secuencia.

En la plataforma Solexa/Illumina (Voelkerding et al., Clinical Chem., 55: 641-658, 2009; MacLean et al., Nature Rev. Microbiol., 7: 287-296;; Patente de Estados Unidos Nº 6.833.246; Patente de los Estados Unidos Nº 7.115.400; Patente de Estados Unidos Nº 6.969.488; los datos de secuenciación se producen en forma de lecturas de longitud más corta. En este método, el ADN fragmentado de cadena sencilla se repara en el extremo para generar extremos romos 5'-fosforilados, seguido de la adición mediada por Klenow de una única base A al extremo 3' de los fragmentos. La adición de A facilita la adición de oligonucleótidos adaptadores de saliente T, que son usados posteriormente para capturar las moléculas adaptadoras de plantillas en la superficie de una célula de flujo que está tachonada con anclajes de oligonucleótidos. El anclaje se usa como un cebador de PCR, pero debido a la longitud de la plantilla y su proximidad a otros oligonucleótidos de anclaje cercanos, la extensión mediante PCR da como resultado el "arqueamiento" de la molécula para hibridar con un oligonucleótido de anclaje adyacente para formar una estructura puente en la superficie de la célula de flujo. Estos giros de ADN se desnaturalizan y escinden. Las cadenas directas se secuencian luego con terminadores de colorante reversibles. La secuencia de nucleótidos incorporados se determina mediante detección de fluorescencia posterior a la incorporación, eliminándose cada flúor y bloque antes del siguiente ciclo de adición de dNTP. La longitud de lectura de la secuencia varía de 36 nucleótidos a más de 50 nucleótidos, con l producción total superando 1 billón de pares de nucleótidos por ejecución del análisis.

La secuenciación de moléculas de ácido nucleico usando tecnología SOLiD (Voelkerding et al., Clinical Chem., 55: 641-658, 2009; MacLean et al., Nature Rev. Microbiol., 7:287-296;; Patente de Estados Unidos Nº 5.912.148; Patente de Estados Unidos Nº 6.130.073; también implica la fragmentación de la plantilla, el ligamiento a adaptadores de oligonucleótidos, la unión a perlas y la amplificación clonal por PCR en emulsión. Después de esto, las perlas que llevan la plantilla se inmovilizan sobre una superficie derivada de una célula de flujo de vidrio, y se aparea un cebador complementario al oligonucleótido adaptador. Sin embargo, en lugar de utilizar este cebador para la extensión 3', se usa en cambio para proporcionar un grupo fosfato 5' para la ligación a sondas de interrogación que contienen dos bases específicas de sondas seguidas por 6 bases degeneradas y uno de cuatro marcadores fluorescentes. En el sistema SOLiD, las sondas de interrogación tienen 16 combinaciones posibles de las dos bases en el extremo 3' de cada sonda, y uno de los cuatro flúores en el extremo 5'. El color del flúor, y por tanto la identidad de cada sonda, corresponde a esquemas de codificación de espacio especificados por color. Las múltiples rondas (habitualmente 7) de apareamiento de sondas, ligamiento, y detección de flúor van seguidas de desnaturalización, y luego una segunda ronda de secuenciación usando un cebador que está desplazado por una base con respecto al cebador inicial. De esta manera, la secuencia plantilla puede reconstruirse computacionalmente, y las bases plantilla se interrogan dos veces, lo que da como resultado una mayor precisión. La longitud de lectura de secuencia promedia 35 nucleótidos, y la producción total excedo los 4 billones de bases por ejecución de secuencia.

En ciertas realizaciones, se emplea la secuenciación de nanoporos (ver, por ejemplo, stier et al. (2006), J. Am. Chem. Soc. 128: 1705-10. La teoría detrás de la secuenciación de nanoporos tiene que ver con lo que ocurre cuando un nanoporo se sumerge en un fluido conductor y se aplica un potencial (voltaje) a través de él. Bajo estas condiciones puede observarse una ligera corriente eléctrica debida a la conducción de iones a través del nanoporo, y la cantidad de corriente es extremadamente sensible al tamaño del nanoporo. A medida que cada base de un ácido nucleico pasa a través del nanoporo, esto provoca un cambio en la magnitud de la corriente a través del nanoporo que es distinto para cada una de las cuatro bases, permitiendo de este modo que se determine la secuencia de la molécula de ADN.

En ciertas realizaciones, se emplea HeliScope de Helicos Biosciences (Voelkerding et al., Clinical Chem., 55: 641-658, 2009; MacLean et al., Nature Rev. Microbiol., 7:287-296; Patente de Estados Unidos Nº 7.169.560; Patente de Estados Unidos Nº 7.282.337; Patente de Estados Unidos Nº 7.482.120; Patente de Estados Unidos Nº 7.501.245; Patente de Estados Unidos Nº 6.818.395; Patente de Estados Unidos Nº 6.911.345; Patente de Estados Unidos Nº 7.501.245. El ADN plantilla está fragmentado y poliadenilado en el extremo 3', con la adenosina final llevando un marcador fluorescente. Los fragmentos de plantilla poliadenilados desnaturalizados se ligan a poli(dT)oligonucleótidos en la superficie de una célula de flujo. Las localizaciones físicas iniciales de las moléculas de plantilla capturadas se graban con una cámara CCD, y luego el marcador se escinde y se lava. La secuenciación se logra mediante la adición de polimerasa y la adición en serie de reactivos dNTP marcados fluorescentemente. Los eventos de incorporación dan como resultado la señal de flúor correspondiente a dNTP, y la señal se captura por una cámara CCD antes de cada ronda de adición de dNTP. La longitud de lectura de secuencia varía de 25-50 nucleótidos, con una producción total que excede 1 billón de pares de nucleótidos por ejecución de análisis.

La tecnología Ion Torrent es un método de secuenciación de ADN basado en la detección de iones de hidrógeno que se liberan durante la polimerización de ADN (ver, por ejemplo, Science 327(5970): 1190 (2010); Publicaciones de Solicitud de Patente de Estados Unidos Nº 20090026082, 20090127589, 20100301398, 20100197507, 20100188073 y 20100137143). Un micropocillo contiene una cadena de ADN plantilla a ser secuenciada. Por debajo de la capa de micropocillos hay un sensor de iones ISFET hipersensible. Todas las capas

están contenidas dentro de un chip semiconductor CMOS, similar al usado en la industria electrónica. Cuando se incorpora un dNTP en la cadena complementaria en crecimiento, se libera un ion de hidrógeno que activa un sensor de iones hipersensible. Si hay repeticiones de homopolímeros en la secuencia de plantilla, se incorporarán múltiples moléculas de dNTP en un solo ciclo. Esto lleva a un número correspondiente de hidrógenos liberados y una señal electrónica proporcionalmente más alta. Esta tecnología difiere de otras tecnologías de secuenciación en que no se usan nucleótidos u ópticas modificados. La precisión por base del secuenciador Ion Torrent es de ~ 99,6% para 50 lecturas de base, con ~100 Mbp generados por ejecución. La longitud de lectura es de 100 pares de bases. La precisión para las repeticiones de homopolímeros de 5 repeticiones de longitud es de ~98%. Los beneficios de la secuenciación de semiconductores iónicos son una velocidad de secuenciación rápida y bajos costos iniciales y operativos.

Otro enfoque de secuenciación de ácidos nucleicos ejemplar que puede adaptarse para su uso con la presente invención fue desarrollado por Stratos Genomics, Inc. e implica el uso de Xpandomeros. Este proceso de secuenciación incluye típicamente proporcionar una cadena hija producida por una síntesis dirigida a plantilla. La cadena hija generalmente incluye una pluralidad de subunidades acopladas en una secuencia correspondiente a una secuencia de nucleótidos contigua de toda o una porción de un ácido nucleico objetivo en el que las subunidades individuales comprenden una cadena, por lo menos una sonda o residuo de nucleobases, y por lo menos un enlace selectivamente escindible. El enlace(s) selectivamente escindible se escinde para producir un Xpandomero de una longitud más larga que la pluralidad de las subunidades de la cadena hija El Xpandomero incluye típicamente las cadenas y los elementos informadores para analizar la información genética en una secuencia correspondiente a la secuencia de nucleótidos contiguos de toda o una parte del ácido nucleico objetivo. Los elementos informadores del Xpandomero son luego detectados. Detalles adicionales relacionados con los enfoques basados en Xpandomeros se describen en, por ejemplo, la Publicación de Patente de Estados Unidos Nº 20090035777, titulada "HIGH THROUGHPUT NUCLEIC ACID SEQUENCING BY EXPANSION", presentada el 19 de junio de 2008. Otros métodos de secuenciación de moléculas individuales emergentes incluyen la secuenciación en tiempo real mediante síntesis usando una plataforma VisiGen ((Voelkerding et al., Clinical Chem., 55: 641-58, 2009, la Patente de Estados Unidos Nº 7.329.492, la Solicitud de Patente de Estados Unidos Nº de Serie 11/671956; Patente de Estados Unidos Aplicación Nº de Serie 11/781166; en las que la plantilla de ADN inmovilizado, cebado se somete a extensión de cadenas usando una polimerasa modificada fluorescentemente y moléculas aceptoras fluorescentes, dando como resultado una transferencia de energía de resonancia de fluorescencia (FRET) detectable tras la adición de nucleótidos.

Otro sistema de secuenciación de moléculas individuales en tiempo real desarrollado por Pacific Biosciences (Voelkerding et al., Clinical Chem., 55: 641-658, 2009; MacLean et al., Nature Rev. Microbiol., 7: 287-296; Patente de Estados Unidos Nº 7.170.050; Patente de Estados Unidos Nº 7.302.146; Patente de Estados Unidos Nº 7.313.308;. Patente de Estados Unidos Nº 7.476.503; utiliza pocillos de reacción de 50-100 nm de diámetro y que abarca un volumen de reacción de aproximadamente 20 zeptoliters (10^{-21} l). Las reacciones de secuenciación se realizan usando plantilla inmovilizada, ADN polimerasa phi29 modificada, y altas concentraciones locales de dNTP marcado fluorescentemente. Las concentraciones locales altas y las condiciones de reacción continuas permiten que se capturen eventos de incorporación en tiempo real mediante detección de señal de flúor usando excitación láser, una guía de onda óptica y una cámara CCD.

Con esta tecnología de secuenciación de ADN de moléculas individuales en tiempo real (SMRT), la secuenciación de ADN se realiza en chips SMRT, cada uno conteniendo miles de guías de onda de modo cero (ZMW). Una ZMW es un agujero, de decenas de nanómetros de diámetro, fabricado en una película de metal de 100 nm depositada sobre un sustrato de dióxido de silicio. Cada ZMW se convierte en una cámara de visualización nanofotónica que proporciona un volumen de detección de solo 20 zeptolitros (10^{-21} l). En este volumen, la actividad de una molécula individual puede detectarse entre un fondo de miles de nucleótidos marcados. La ZMW proporciona una ventana para observar la ADN polimerasa ya que realiza la secuencia mediante síntesis. Dentro de cada cámara, una única molécula de ADN polimerasa está unida a la superficie inferior de tal manera que reside permanentemente dentro del volumen de detección. Los nucleótidos fosfoenlazados, cada tipo marcado con un fluoróforo de diferente color, se introducen luego en la solución de reacción a concentraciones altas que promueven la velocidad, precisión y procesividad de las enzimas. Debido al pequeño tamaño de la ZMW, incluso a estas concentraciones altas biológicamente relevantes, el volumen de detección está ocupado por nucleótidos solo una pequeña fracción del tiempo. Además, las visitas al volumen de detección son rápidas, duran solo unos pocos microsegundos, debido a la muy pequeña distancia que la difusión tiene que llevar los nucleótidos. El resultado es un fondo muy bajo.

Los procesos, composiciones y sistemas para la secuenciación que pueden adaptarse para su uso con la invención se describen, por ejemplo, en las Patentes de Estados Unidos Nº 7.405.281, titulada "Fluorescent nucleotide analogs and uses therefor", expedida el 29 de julio de 2008 a Xu et al.; 7.315.019, titulada ""Arrays of optical confinements and uses thereof", expedida el 1 de enero de 2008 a Turner et al., 7.313.308, titulada "Optical analysis of molecules", expedida el 25 de diciembre de 2007 a Turner et al.; 7.302.146, titulada "Apparatus and method for analysis of molecules", expedida el 27 de noviembre de 2007 a Turner et al.; y 7.170.050, titulada "Apparatus and methods for optical analysis of molecules", expedida el 30 de enero de 2007 a Turner et al.; y las

Publicaciones de Patente de Estados Unidos Nº 20080212960, titulada "Methods and systems for simultaneous realtime monitoring of optical signals from multiple sources", presentada el 26 de octubre de 2007 por Lundquist et al.; 20080206764, titulada "Flowcell system for single molecule detection", presentada el 26 de octubre de 2007 por Williams et al.; 20080199932, titulada "Active surface coupled polymerases", presentada el 26 de octubre de 2007 por Hanzel et al.; 20080199874, titulada "CONTROLLABLE STRAND SCISSION OF MINI CIRCLE DNA", presentada el 11 de febrero de 2008 por Otto et al.; 20080176769, titulada "Articles having localized molecules disposed thereon and methods of producing same", presentada el 26 de octubre de 2007 por Rank et al.; 20080176316, titulada "Mitigation of photodamage in analytical reactions", presentada el 31 de octubre de 2007 por Eid et al.; 20080176241, titulada "Mitigation of photodamage in analytical reactions", presentada el 31 de octubre de 2007 por Eid et al.; 20080165346, titulada "Methods and systems for simultaneous real-time monitoring of optical signals from multiple sources", presentada el 26 de octubre de 2007 por Lundquist et al.; 20080160531, titulada "Uniform surfaces for hybrid material substrates and methods for making and using same", presentada el 31 de octubre de 2007 por Korlach; 20080157005, titulada "Methods and systems for simultaneous realtime monitoring of optical signals from multiple sources", presentada el 26 de octubre de 2007 por Lundquist et al.; 20080153100, titulada "Articles having localized molecules disposed thereon and methods of producing same", presentada el 31 de octubre de 2007 por Rank et al.; 20080153095, titulada "CHARGE SWITCH NUCLEOTIDES", presentada el 26 de octubre de 2007 por Williams et al.; 20080152281, titulada "Substrates, systems and methods for analyzing materials", presentada el 31 de octubre de 2007 por Lundquist et al.; 20080152280, titulada "Substrates, systems and methods for analyzing materials", presentada el 31 de Octubre de 2007 por Lundquist et al.; 20080145278, titulada "Uniform surfaces for hybrid material substrates and methods for making and using same", presentada el 31 de octubre de 2007 por Korlach; 20080128627, titulada "SUBSTRATES, SYSTEMS AND METHODS FOR ANALYZING MATERIALS", presentada el 31 de agosto de 2007 por Lundquist et al.; 20080108082, titulada "Polymerase enzymes and reagents for enhanced nucleic acid sequencing", presentada el 22 de octubre de 2007 por Rank et al.; 20080095488, titulada "SUBSTRATES FOR PERFORMING ANALYTICAL REACTIONS", presentada el 11 de junio de 2007 por Foquet et al.; 20080080059, titulada "MODULAR OPTICAL COMPONENTS AND SYSTEMS INCORPORATING SAME", presentada el 27 de septiembre de 2007 por Dixon et al.; 20080050747, titulada "Articles having localized molecules disposed thereon and methods of producing and using same", presentada el 14 de agosto de 2007 por Korlach et al.; 20080032301, titulada "Articles having localized molecules disposed thereon and methods of producing same", presentada el 29 de marzo de 2007 por Rank et al.; 20080030628, titulada "Methods and systems for simultaneous real-time monitoring of optical signals from multiple sources", presentada el 9 de febrero de 2007 por Lundquist et al.; 20080009007, titulada "CONTROLLED INITIATION OF PRIMER EXTENSION", presentada el 15 de junio, 2007 por Lyle et al.; 20070238679, titulada "Articles having localized molecules disposed thereon and methods of producing same", presentada el 30 de marzo de 2006 por Rank et al.; 20070231804, titulada "Methods, systems and compositions for monitoring enzyme activity and applications thereof", presentada el 31 de marzo de 2006 por Korlach et al., 20070206187, titulada "Methods and systems for simultaneous real-time monitoring of optical signals from multiple sources", presentada el 9 de febrero de 2007 por Lundquist et al., 20070196846, titulada "Polymerases for nucleotide analogue incorporation", presentada el 21 de diciembre de 2006 por Hanzel et al., 20070188750, titulada "Methods and systems for simultaneous real-time monitoring of optical signals from multiple sources", presentada el 7 de julio de 2006 por Lundquist et al., 20070161017, titulada "MITIGATION OF PHOTODAMAGE IN ANALYTICAL REACTIONS", presentada el 1 de diciembre de 2006 por Eid et al.; 20070141598, titulada "Nucleotide Compositions and Uses Thereof" presentada el 3 de noviembre de 2006 por Turner et al.; 20070134128, titulada "Uniform surfaces for hybrid material substrate and methods for making and using same", presentada el 27 de noviembre de 2006 por Korlach; 20070128133, titulada "Mitigation of photodamage in analytical reactions", presentada el 2 de diciembre de 2005 por Eid et al.; 20070077564, titulada "Reactive surfaces, substrates and methods of producing same", presentada el 30 de septiembre de 2005 por Roitman et al.; 20070072196, titulada "Fluorescent nucleotide analogs and uses therefore", presentada el 29 de septiembre de 2005 por Xu et al; y 20070036511, titulada "Methods and systems for monitoring multiple optical signals from a single source", presentada el 11 de agosto de 2005 por Lundquist et al. y Korlach et al. (2008) "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures" PNAS 105(4): 1176-81.

2. Composiciones

La tecnología proporciona composiciones que comprenden una o más bases, por ejemplo, para su uso en la secuenciación de ácidos nucleicos usando un código de dos bases degenerado. En algunas realizaciones, las bases se marcan por pares, por ejemplo, dos de las cuatro bases de nucleótidos se marcan con un primer marcador y las (otras) dos restantes de las cuatro bases de nucleótidos se marcan con un segundo marcador de tal manera que se diferencia una base del primer par de una base en el segundo par, pero las bases dentro de cada par no se diferencian entre sí con respecto al marcador unido a ellas.

La tecnología abarca varias realizaciones de composiciones que comprenden bases marcadas y/o análogos de bases. Por ejemplo, en algunas realizaciones, las composiciones comprenden una o más bases que están marcadas para diferenciar bases de pirimidina (C y T) de bases de purina (A y G), por ejemplo, para su uso en un método para producir una secuencia que denota la secuencia de pirimidinas (por ejemplo, "Y") y purinas (por ejemplo, con" R "). En algunas realizaciones, las bases se marcan para diferenciar bases ceto (G y T) de bases

amino (A y C), por ejemplo, para su uso en un método para producir una secuencia que denota la secuencia de bases ceto (por ejemplo, "K") y bases amino (por ejemplo, "M"). En algunas realizaciones, las bases se marcan para diferenciar bases que forman pares de bases más fuertes (G y C) de las que forman pares de bases más débiles (A y T), por ejemplo, para su uso en un método para producir una secuencia que denota la secuencia de bases que formar pares de bases más fuertes (por ejemplo, "S") y bases que forman pares de bases más débiles (por ejemplo, "W").

En algunas realizaciones, las composiciones comprenden uno o más análogos de bases que forman pares de bases con las cuatro bases de nucleótidos de acuerdo con una regla de emparejamiento de bases degenerada; es decir, cada base de análogo de nucleótido se empareja con dos nucleótidos de acuerdo con una regla de emparejamiento de bases degenerada. En particular, los dos nucleótidos con los que se empareja el primer par de bases del análogos de nucleótidos son diferentes (sin superposición) de los dos nucleótidos con los que se empareja el segundos par de bases del análogo de nucleótidos. Por ejemplo, en algunas realizaciones, las composiciones comprenden uno o dos análogos de nucleótidos X e Y en donde X es un par de bases con una purina e Y un par de bases con una pirimidina. Realizaciones similares comprenden composiciones de uno o dos análogos de nucleótidos que emparejan bases de acuerdo con una regla degenerada en la que un análogo de nucleótido se empareja con una base amino (por ejemplo, M = A o C) y el otro análogo de nucleótido se empareja con una base ceto (por ejemplo, K = G o T). De manera similar, las realizaciones comprenden composiciones de uno o dos pares de bases de análogos de nucleótidos de acuerdo con una regla degenerada en la que un análogo de nucleótido se empareja con una base de enlace de hidrógeno fuerte (por ejemplo, S = C o G) y el otro análogo de nucleótido se empareja con una base de enlace de hidrógeno débil (por ejemplo, W = A o T).

Se describen análogos de pares de bases que actúan como un nucleótido de purina y pirimidina degenerado (por ejemplo, ese par de bases de acuerdo con una regla de emparejamiento de bases degenerada en la que el nucleótido reconoce enlaces, y pares de bases con, o ambos A y G o ambos C y T), por ejemplo, en Abraham, et al., "Nucleobase analogs for degenerate hybridization devised through conformational pairing analysis" (2007), Biotechniques 43: 617. Ver también Linet al., "Synthesis of oligodeoxyribonucleotides containing degenerate bases and their use as primers in the polymerase chain reaction" (1992), Nucleic Acids Res. 19: 5149. Ejemplos adicionales son 8-hidroxiguanina, 2-hidroxiadenina, 6-O-metilguanina y xantina, cuya base se empareja con C y A (por ejemplo, M); T y A (W); T y C (por ejemplo, Y); y T y C (por ejemplo, Y), respectivamente, y actúan por tanto como bases que se pueden denotar como K, S, R, y R, respectivamente. Las bases no estándar se pueden incorporar mediante polimerasas, por ejemplo, como se describe en la Solicitud de Patente Internacional WO 2009/154733.

De acuerdo con algunas realizaciones de la tecnología, las bases se marcan con una fracción que da como resultado la producción de una señal detectable tras la incorporación de la base en la cadena de ADN que se está sintetizando. En algunas realizaciones, la fracción produce una señal (por ejemplo, fluorescencia) antes de la incorporación y/o después de la incorporación. En algunas realizaciones, la fracción está enlazada de una manera que es apropiado para eliminar la fracción después de la incorporación o después de la obtención de imágenes. La fracción de marcación es, en algunas realizaciones, un colorante orgánico fluorescente derivado para la unión a la base directamente o a través de un conector. En la bibliografía está disponible guía práctica que proporciona una lista de moléculas fluorescentes y cromogénicas y sus propiedades ópticas relevantes (ver, por ejemplo, Berlman, Handbook of Fluorescence Spectra of Aromatic Molecules, 2ª edición (Academic Press, New York, 1971); Griffiths, Colour and Constitution of Organic Molecules (Academic Press, New York, 1976); Bishop, Ed., Indicators (Pergamon Press, Oxford, 1972); Haugland, Handbook of Fluorescent Probes and Research Chemicals (Molecular Probes, Eugene, 1992); Pringsheim, Fluorescence and Phosphorescence (Interscience Publishers, New York, 1949); y similares.

Además, hay una guía en la bibliografía para derivar moléculas fluorescentes para la unión covalente a través de grupos reactivos comunes que se pueden añadir a un nucleótido, como se ejemplifica en Haugland (supra); Ullman et al., Patente de Estados Unidos Nº 3.996.345; Khanna et al, Patente de Estados Unidos Nº 4.351.760. Hay muchas fracciones de enlace y metodologías para unir marcadores fluorescentes o fracciones neutralizantes a nucleótidos, como se ejemplifica por las siguientes referencias: Eckstein, editor, Oligonucleotides and Analogues: A Practical Approach (IRL Press, Oxford, 1991); Zuckerman et al. (1987), Nucleic Acids Research 15: 5305-5321; Sharma et al. (1991), Nucleic Acids Research 19: 3019; Giusti et al., PCR Methods and Applications 2: 223-227 (1993); Fung et al., Patente de Estados Unidos Nº 4.757.141; Stabinsky, Patente de Estados Unidos Nº 4.739.044; Agrawal et al. (1990), Tetrahedron Letters 31: 1543-1546; Sproat et al. (1987), Nucleic Acids Research 15: 4837; Nelson et al. (1989), Nucleic Acids Research 17: 7187-7194; y similares. Una variedad de metodologías de secuenciación basadas en fluorescencia de ADN son conocidas en la técnica (ver, por ejemplo, Birren et al., Genome Analysis: Analyzing DNA, (Cold Spring Harbor, NY).

Las realizaciones de la tecnología comprenden composiciones que comprenden una plantilla de ácido nucleico objetivo. En algunas realizaciones, la composición comprende un cebador, por ejemplo, en algunas realizaciones que está unido a la plantilla de ácido nucleico objetivo.

El ácido nucleico objetivo no es crítico y puede provenir de una variedad de fuentes estándar. Puede ser ARNm, ARN ribosomal, ADN genómico o ADNc. Cuando el objetivo es de una fuente biológica, se conocen procedimientos para extraer ácido nucleico y opcionalmente amplificarlo a una concentración conveniente para el genotipado o el trabajo de secuencia. El ácido nucleico puede obtenerse de cualquier célula viva de una persona, animal, o planta (y en muchos casos de células muertas u otra materia de origen biológico). Los humanos, los microbios patogénicos, y los virus son fuentes particularmente interesantes. También se conocen métodos de amplificación de ácidos nucleicos. Preferiblemente, la amplificación se lleva a cabo mediante reacción en cadena de la polimerasa (PCR) (Patentes de Estados Unidos Nº 4.683.202, 4.683.195 y 4.889.818; Gyllenstein et al (1988) Proc. Natl. Acad. Sci. USA 85: 7652-7656; Ochman et al. (1988) Genetics 120: 621 - 623; Loh et al. (1989) Science 243: 217-220; Innis et al (1990) PCR Protocols (Academic Press, San Diego, CA). Pueden usarse otros métodos de amplificación conocidos en la técnica, incluyendo, pero sin limitación, reacción en cadena de la ligasa (ver, por ejemplo, EP 320308), el uso de Q-beta replicasa, o los métodos enumerados en Kricka et al., 1995, Molecular Probing, Blotting and Sequencing (Academic Press, Nueva York), especialmente el Capítulo 1 y la Tabla IX.

La tecnología proporcionada en la presente se refiere al uso de una polimerasa en una reacción de secuenciación. En general, las polimerasas que encuentran uso en la tecnología toleran marcadores en varias posiciones, por ejemplo, en la nucleobase, en el gamma-fosfato, en el 3' hidroxilo, etc. Por ejemplo, las polimerasas que encuentran uso en la tecnología incluyen, pero no están limitadas a, ADN polimerasas dependientes de ADN, ARN polimerasas dependientes de ADN, ADN polimerasas dependientes de ARN, ARN polimerasas dependientes de ARN, ADN polimerasa T7, ADN polimerasa T3, ADN polimerasa T4, ARN polimerasa T7, ARN polimerasa T3, ARN polimerasa SP6, ADN polimerasa 1, fragmento de Klenow, ADN polimerasa de *Thermophilus aquaticus*, ADN polimerasa *Tth*, ADN polimerasa Vent ((New England Biolabs) ADN polimerasa , Deep Vent (New England Biolabs), Fragmento grande de ADN polimerasa *Bst*, fragmento de Stoeffel, ADN polimerasa 9º N, ADN polimerasa *Pfu*, ADN polimerasa *Tfl*, polimerasa RepliPHI Phi29, ADN polimerasa *Tli*, ADN polimerasa beta eucariota, telomerasa, polimerasa Therminator (New England Biolabs), ADN polimerasa KOD HiFi. (Novagen), ADN polimerasa KOD1, Q-beta replicasa, transferasa terminal, transcriptasa inversa AMV, transcriptasa inversa M-MLV, transcriptasa inversa Phi6, transcriptasa inversa VIH-1, nuevas polimerasas descubiertas por bioprospección y polimerasas enumeradas en la Publicación de Solicitud de Patente de Estados Unidos Nº 2007/0048748 y en las Patentes de Estados Unidos Nº 6.329.178; 6.602.695; y 6.395.524. Estas polimerasas incluyen isoformas mutantes de tipo salvaje y variantes genéticamente diseñadas. En algunas realizaciones se usa una polimerasa defectuosa de exonucleasas. En algunas realizaciones (por ejemplo, una tecnología de terminación reversible), se usa una polimerasa que tiene una actividad de exonucleasas para algunos o todos los pasos.

Los cebadores (para síntesis por ADN polimerasa) o promotores (para síntesis por ARN polimerasa) típicamente se hacen sintéticamente usando tecnología de síntesis de ácidos nucleicos convencional, por ejemplo, usando un sintetizador de ADN automatizado y químicas estándar, tales como química de fosforamiditas, por ejemplo, como se divulga en las siguientes referencias: Beaucage and Iyer, Tetrahedron 48: 2223-211 (1992); Patente de Estados Unidos Nº 4.980.460; Patente de Estados Unidos Nº 4.725.677; Patentes de Estados Unidos Nº 4.415.732; 4.458.066; y 4.973.679; y similares. También pueden emplearse químicas alternativas, por ejemplo, que dan como resultado grupos estructurales no naturales, como fosforotioato, fosforamidato, y similares, siempre que los oligonucleótidos resultantes sean compatibles con la polimerasa. Se pueden pedir comercialmente de una variedad de compañías que se especializan en oligonucleótidos personalizados como Operon, IDT, Dharmacon, etc.

Los cebadores en combinación con polimerasas se usan para secuenciar ADN objetivo. La longitud del cebador se selecciona para proporcionar hibridación con ADN plantilla complementario. Los cebadores son generalmente de por lo menos 10 nt de longitud, habitualmente de por lo menos entre 15 y 30 nt de longitud. Los cebadores están diseñados para hibridar con sitios internos conocidos en el ADN objetivo sujeto. Alternativamente, los cebadores pueden unirse a adaptadores de oligonucleótidos sintéticos unidos a los extremos del ADN objetivo mediante una ligasa. De manera similar, cuando se usan promotores, pueden ser internos al ADN objetivo o ligados como adaptadores a los extremos.

La mezcla de la reacción para la secuenciación comprende un medio tampón acuoso que está optimizado para la polimerasa particular elegida. En general, el tampón incluye típicamente una fuente de iones monovalentes, una fuente de cationes divalentes, y un agente de tamponamiento. Puede emplearse cualquier fuente conveniente de iones monovalentes, como cloruro de potasio, acetato de potasio, acetato de potasio, glutamato de potasio, cloruro de amonio, sulfato de amonio y similares.

El catión divalente puede ser magnesio, manganeso, zinc y similares, donde el catión será típicamente magnesio. Puede emplearse cualquier fuente conveniente de catión magnesio, incluyendo $MgCl_2$, acetato de magnesio y similares. La cantidad de ion de Mg presente en el tampón puede variar de 0,5 a 20 mM, pero preferiblemente variará de aproximadamente 1 a 12 mM, más preferiblemente de 2 a 10 mM,, e idealmente será de aproximadamente 5 mM.

Los agentes de tamponamiento representativos o sales que pueden estar presentes en las composiciones de acuerdo con la tecnología descrita (por ejemplo, en una composición que comprende un nucleótido marcado o en

una reacción de SBS) incluyen Tris, Tricina, HEPES, MOPS y similares, donde la cantidad de agente de tamponamiento variará típicamente de aproximadamente 5 a 150 mM, habitualmente de aproximadamente 10 a 100 mM, y más habitualmente de aproximadamente 20 a 50 mM, donde en ciertas realizaciones preferidas, el agente de tamponamiento estará presente en una cantidad suficiente para proporcionar un pH que varía de aproximadamente 6.0 a 9.5. Otros agentes que pueden estar presentes en el medio tampón incluyen agentes quelantes, como EDTA, EGTA y similares.

En algunas realizaciones, el marcador (por ejemplo, fracción fluorescente) se une a la base de nucleótidos y, en algunas realizaciones el marcador se une a la cadena de fosfato (por ejemplo, en métodos como la secuenciación SMRT de Pacific Biosciences).

3. Análisis de datos

Algunas realizaciones comprenden un sistema informático sobre el que pueden implementarse las realizaciones de las presentes enseñanzas. En varias realizaciones, un sistema informático incluye un bus u otro mecanismo de comunicación para comunicar información y un procesador acoplado con el bus para procesar información. En varias realizaciones, el sistema informático incluye una memoria, que puede ser una memoria de acceso aleatorio (RAM), u otro dispositivo de almacenamiento dinámico, acoplado al bus para identificar bases (por ejemplo, haciendo "llamadas de base"), e instrucciones para ser ejecutadas por el procesador. La memoria también puede usarse para almacenar variables temporales u otra información intermedia durante la ejecución de las instrucciones a ser ejecutadas por el procesador. En varias realizaciones, el sistema informático puede incluir además una memoria de solo lectura (ROM) u otro dispositivo de almacenamiento estático acoplado al bus para almacenar información estática e instrucciones para el procesador. Puede proporcionarse un dispositivo de almacenamiento, como un disco magnético o disco óptico, y acoplarlo al bus para almacenar información e instrucciones.

En varias realizaciones, el sistema informático se acopla a través del bus a una pantalla, como un tubo de rayos catódicos (CRT) o una pantalla de cristal líquido (LCD), para mostrar información a un usuario del ordenador. Un dispositivo de entrada, incluyendo teclas alfanuméricas y otras, pueden acoplarse al bus para comunicar información y selecciones de comandos al procesador. Otro tipo de dispositivo de entrada del usuario es un control de cursor, como un ratón, una rueda de desplazamiento, o teclas de dirección del cursor para comunicar información de dirección y selecciones de comandos al procesador y para controlar el movimiento del cursor en la pantalla. Este dispositivo de entrada tiene típicamente dos grados de libertad en dos ejes, un primer eje (por ejemplo, x) y un segundo eje (por ejemplo, y), que le permite al dispositivo especificar posiciones en un plano.

Un sistema informático puede realizar realizaciones de la presente tecnología. De acuerdo con ciertas implementaciones de las presentes enseñanzas, pueden proporcionarse resultados por el sistema informático en respuesta al procesador que ejecuta una o más secuencias de una o más instrucciones contenidas en la memoria. Tales instrucciones se pueden leer en la memoria desde otro medio legible por ordenador, como un dispositivo de almacenamiento. La ejecución de las secuencias de instrucciones contenidas en la memoria puede hacer que el procesador realice los métodos descritos en la presente. Alternativamente, puede usarse una circuitería de cable físico en lugar de o en combinación con instrucciones de software para implementar las presentes enseñanzas. Por tanto, las implementaciones de las presentes enseñanzas no están limitadas a ninguna combinación específica de circuitos de hardware y software.

El término "medio legible por ordenador" como se usa en la presente se refiere a cualquier medio que participe en la provisión de instrucciones al procesador para su ejecución. Tal medio puede tomar muchas formas, incluyendo pero no limitadas a, medios no volátiles, medios volátiles y medios de transmisión. Los ejemplos de medios no volátiles pueden incluir, pero no están limitados a, discos ópticos o magnéticos. Los ejemplos de medios volátiles pueden incluir, pero no están limitados a, memoria dinámica y flash. Los ejemplos de medios de transmisión pueden incluir, pero no están limitados a, cables coaxiales, cable de cobre y fibras ópticas, incluyendo los cables que componen el bus.

Las formas comunes de medios legibles por ordenador incluyen, por ejemplo, un disquete, un disco flexible, disco duro, cinta magnética o cualquier otro medio magnético, un CD-ROM, cualquier otro medio óptico, tarjetas perforadas, cinta de papel, cualquier otro medio físico con patrones de orificios, una RAM, PROM, y EPROM, un FLASH-EPROM, cualquier otro chip o cartucho de memoria, o cualquier otro medio tangible del que pueda leer un ordenador.

Varias formas de medios legibles por ordenador pueden estar implicadas en llevar una o más secuencias de una o más instrucciones al procesador para su ejecución. Por ejemplo, las instrucciones pueden llevarse inicialmente en el disco magnético de un ordenador remoto. El ordenador remoto puede cargar las instrucciones en su memoria dinámica y enviar las instrucciones a través de una conexión de red (por ejemplo, una LAN, una WAN, Internet, una línea telefónica). Un sistema informático local puede recibir los datos y transmitirlos al bus. El bus puede llevar los datos a la memoria, de la que el procesador recupera y ejecuta las instrucciones. Las instrucciones

recibidas por la memoria pueden almacenarse opcionalmente en un dispositivo de almacenamiento ya sea antes o después de la ejecución por parte del procesador.

De acuerdo con varias realizaciones, las instrucciones configuradas para ser ejecutadas por un procesador para realizar un método se almacenan en un medio legible por ordenador. El medio legible por ordenador puede ser un dispositivo que almacena información digital. Por ejemplo, un medio legible por ordenador incluye una memoria de solo lectura de disco compacto (CD-ROM) como se conoce en la técnica para almacenar software. Al medio legible por ordenador se accede mediante un procesador adecuado para ejecutar instrucciones configuradas para ejecutarse.

De acuerdo con dicho sistema informático, algunas realizaciones de la tecnología proporcionada en la presente comprenden además funcionalidades para recoger, almacenar y/o analizar datos (por ejemplo, datos de secuencias de nucleótidos). Por ejemplo, algunas realizaciones contemplan un sistema que comprende un procesador, una memoria y/o una base de datos para, por ejemplo, almacenar y ejecutar instrucciones, analizar datos de imágenes de una reacción de secuenciación, realizar cálculos usando los datos, transformar los datos y almacenarlos los datos. En algunas realizaciones, un algoritmo de llamada base asigna una secuencia de bases a los datos y asocia puntuaciones de calidad a llamadas base en base a un modelo estadístico. En algunas realizaciones, el sistema está configurado para ensamblar una secuencia a partir de múltiples sub-secuencias, en algunos casos teniendo en cuenta la superposición y calculando una secuencia consenso. En algunas realizaciones, una secuencia se alinea con una secuencia de referencia o con un supercóntigo.

En algunas realizaciones, se analizan dos o más secuencias degeneradas del mismo ácido nucleico en combinación para proporcionar una secuencia "fusionada" en la anotación convencional de cuatro bases. Por ejemplo, una primera secuencia degenerada de dos bases de RYRY y una segunda secuencia degenerada de dos bases MMKK para la misma secuencia indica que la primera posición es una base de amino purina, la segunda posición es una base de amino pirimidina, la cuarta posición es una base de ceto purina, y la cuarta posición es una base de ceto pirimidina, dando como resultado por tanto la ACGT de secuencia de cuatro bases convencional para el ácido nucleico.

Muchos diagnósticos implican determinar la presencia de, o una secuencia de nucleótidos de, uno o más ácidos nucleicos. Por tanto, en algunas realizaciones, una ecuación que comprende variables que representan la presencia o las propiedades de secuencia de múltiples ácidos nucleicos produce un valor que encuentra uso al hacer un diagnóstico o evaluar la presencia o las cualidades de un ácido nucleico. Como tal, en algunas realizaciones este valor se presenta mediante un dispositivo, por ejemplo, mediante un indicador relacionado con el resultado (por ejemplo, un LED, un icono en una LCD, un sonido o similar). En algunas realizaciones, un dispositivo almacena el valor, transmite el valor o usa el valor para cálculos adicionales.

Además, en algunas realizaciones, un procesador está configurado para controlar las reacciones de secuenciación y recoger los datos (por ejemplo, imágenes). En algunas realizaciones, el procesador se usa para iniciar y/o finalizar cada ronda de secuenciación y recogida de datos relacionados con una reacción de secuenciación. Algunas realizaciones comprenden un procesador configurado para analizar los datos y discernir la secuencia del ácido nucleico objetivo y/o de su complemento.

En algunas realizaciones, el procesador usa un dispositivo que comprende una interfaz de usuario (por ejemplo, un teclado, botones, diales, conmutadores y similares) para recibir la entrada del usuario para dirigir una medición. En algunas realizaciones, el dispositivo comprende además una salida de datos para transmitir (por ejemplo, mediante una conexión por cable o inalámbrica) datos a un destino externo, por ejemplo, un ordenador, una pantalla, una red y/o un medio de almacenamiento externo.

En algunas realizaciones, la tecnología encuentra uso en el ensayo de la presencia de uno o más ácidos nucleicos y/o proporcionar la secuencia de uno o más ácidos nucleicos. Por consiguiente, la tecnología proporcionada en la presente encuentra uso en los campos médico, clínico y de medicina de emergencia. En algunas realizaciones, se usa un dispositivo para analizar muestras biológicas. En tal ensayo, la muestra biológica comprende un ácido nucleico y la secuenciación del ácido nucleico es indicativa de un estado o una propiedad de la muestra y, en algunas realizaciones, del sujeto del que se tomó la muestra. Algunas muestras relevantes incluyen, pero no están limitadas a, sangre total, linfa, plasma, suero, saliva, orina, heces, sudoración, moco, lágrimas, líquido cefalorraquídeo, secreción nasal, secreción cervical o vaginal, semen, líquido pleural, líquido amniótico, líquido peritoneal, líquido del oído medio, líquido articular, aspirado gástrico, homogeneizado tisular, homogeneizado celular o similares.

La secuencia de señales de salida proporciona la secuencia del ADN sintetizado y, por las reglas de complementariedad de bases, también proporciona por tanto la secuencia de la cadena plantilla.

Aparatos

Un aspecto adicional de la invención proporciona un aparato para llevar a cabo los métodos o para preparar las composiciones de la tecnología. Tal aparato podría comprender, por ejemplo, una pluralidad de plantillas y cebadores de ácidos nucleicos unidos, preferiblemente covalentemente, a un soporte sólido, junto con una polimerasa de ácido nucleico, una pluralidad de nucleótidos o análogos de nucleótidos como los descritos anteriormente, y una funcionalidad para controlar la temperatura y/o las adiciones de nucleótidos. Preferiblemente, el aparato también comprende una funcionalidad de detección para detectar y distinguir señales de agrupaciones de ácidos nucleicos individuales. Dicha funcionalidad de detección podría comprender un dispositivo acoplado a carga conectado operativamente a un dispositivo de amplificación como un microscopio. Preferiblemente, cualquier aparato de la invención se proporciona en una forma automatizada, por ejemplo, bajo el control de un programa de pasos y decisiones, por ejemplo, como se implementa en un software informático.

Algunas realizaciones de dicho aparato incluyen una unidad de administración y control de fluidos; una unidad de procesamiento de muestra; una unidad de detección de señales; y una unidad de adquisición, análisis y control de datos. Varias realizaciones del aparato pueden proporcionar una secuenciación automatizada que puede usarse para recopilar información de secuencia de una pluralidad de secuencias en paralelo, por ejemplo, sustancialmente de forma simultánea.

En varias realizaciones, la unidad de administración y control de fluidos incluye un sistema de administración de reactivos. El sistema de administración de reactivos puede incluir un depósito de reactivos para el almacenamiento de varios reactivos (por ejemplo, composiciones de nucleótidos o análogos de nucleótidos de acuerdo con la tecnología). Los reactivos pueden incluir cebadores basados en ARN, cebadores de ADN directos/inversos, mezclas de oligonucleótidos para la secuenciación por ligamiento, mezclas de nucleótidos para secuenciación por síntesis, tampones, reactivos de lavado, reactivo de bloqueo, reactivos de agotamiento y similares. Además, el sistema de administración de reactivos puede incluir un sistema de pipeteo o un sistema de flujo continuo que conecta la unidad de procesamiento de muestra con el depósito de reactivos.

En varias realizaciones, la unidad de procesamiento de muestras puede incluir una cámara de muestra, como una celda de flujo, un sustrato, una micromatriz, una bandeja de múltiples pocillos o similar. La unidad de procesamiento de muestras puede incluir múltiples carriles, múltiples canales, múltiples pocillos u otros modos de procesamiento de conjuntos de muestras múltiples de manera sustancialmente simultánea. Adicionalmente, la unidad de procesamiento de muestras puede incluir múltiples cámaras de muestras para permitir el procesamiento de múltiples ejecuciones simultáneamente. En realizaciones particulares, el sistema puede realizar la detección de señales en una cámara de muestras a la vez que procesa sustancialmente de manera simultánea otra cámara de muestras. Adicionalmente, la unidad de procesamiento de muestras puede incluir un sistema de automatización para mover o manipular la cámara de muestras.

En varias realizaciones, la unidad de detección de señales puede incluir un sensor de obtención de imágenes o detección. La unidad de detección de señales puede incluir un sistema de excitación para provocar que una sonda, como un colorante fluorescente, emita una señal. El sistema de excitación puede incluir una fuente de iluminación, como una lámpara de arco, un láser, un diodo emisor de luz (LED) o similar. En realizaciones particulares, la unidad de detección de señales puede incluir óptica para la transmisión de luz desde una fuente de iluminación a la muestra o desde la muestra al sensor de obtención de imágenes o detección. Alternativamente, la unidad de detección de señales puede no incluir una fuente de iluminación, como, por ejemplo, cuando se produce una señal espontáneamente como resultado de una reacción de secuenciación. Por ejemplo, una señal puede producirse por la interacción de una fracción liberada, como un ion liberado que interactúa con una capa sensible a iones, o un pirofosfato que reacciona con una enzima u otro catalizador para producir una señal quimioluminiscente.

En varias realizaciones, una unidad de análisis y control de adquisición de datos puede monitorizar varios parámetros del sistema. Los parámetros del sistema pueden incluir la temperatura de varias partes del instrumento, como una unidad de procesamiento de muestras o depósitos de reactivos, volúmenes de varios reactivos, el estado de varios subcomponentes del sistema, como un manipulador, un motor a pasos, una bomba o similares, o cualquier combinación de los mismos.

Un experto en la técnica apreciará que pueden usarse varias realizaciones de dicho instrumento para poner en práctica una variedad de métodos de secuenciación que incluyen métodos basados en ligamiento, secuenciación por síntesis, métodos de moléculas individuales y otras técnicas de secuenciación. La secuenciación por ligamiento puede incluir técnicas de ligamiento único, o técnicas de ligamiento de cambio donde se realizan múltiples ligamientos en secuencia en un único primario. La secuenciación por síntesis puede incluir la incorporación de nucleótidos marcados con colorante, terminación de cadena o similares. Las técnicas de moléculas individuales pueden incluir secuenciación escalonada, donde las reacciones de secuenciación se pausan para determinar la identidad del nucleótido incorporado.

En varias realizaciones, el instrumento de secuenciación puede determinar la secuencia de un ácido nucleico, como un polinucleótido o un oligonucleótido. El ácido nucleico puede incluir ADN o ARN, y puede ser de cadena sencilla, como ADNmc y ARN, o de cadena doble, como ADNcd o un par de ARN/ADNc. En varias

realizaciones, el ácido nucleico puede incluir o derivarse de una biblioteca de fragmentos, una biblioteca de parejas acopladas, un fragmento ChIP o similar. En realizaciones particulares, el instrumento de secuenciación puede obtener la información de secuencia de un grupo de moléculas de ácido nucleico sustancialmente idénticas.

En varias realizaciones, el instrumento de secuenciación puede producir datos de lectura de la secuenciación de ácidos nucleicos en una variedad de diferentes tipos/formatos de archivos de datos de salida, incluyendo, pero no limitados a: *.fasta, *.csfasta, *seq.txt, *qseq.txt, *.fastq, *.sff, *prb.txt, *.sms, *srs y/o *.qv.

Algunas realizaciones comprenden un sistema para reconstruir una secuencia de ácido nucleico, por ejemplo, una secuencia de base generada de dos bases o una secuencia de cuatro bases "fusionada", de acuerdo con las varias realizaciones proporcionadas en la presente. El sistema puede incluir un secuenciador de ácidos nucleicos, un almacenamiento de datos de secuencias de la muestra, un almacenamiento de datos de secuencias de referencia y un dispositivo/servidor/nodo de computación analítica. En varias realizaciones, el dispositivo/servidor/nodo de computación analítica puede ser una estación de trabajo, un ordenador central, un ordenador personal, un dispositivo móvil, etc.

El secuenciador de ácidos nucleicos puede configurarse para analizar (por ejemplo, interrogar) un fragmento de ácido nucleico (por ejemplo, fragmento individual, fragmento de pareja acoplada, fragmento de extremo emparejado, etc.) utilizando todas las variedades apropiadas de técnicas, plataformas o tecnologías para obtener información de la secuencia de ácido nucleico, por ejemplo, usando una secuenciación de conjuntos por síntesis. En varias realizaciones, el secuenciador de ácidos nucleicos puede estar en comunicación con el almacenamiento de datos de secuencia de muestra directamente a través de un cable de datos (por ejemplo, un cable en serie, una conexión de cable directa, etc.) o enlace de bus o, alternativamente, a través de una conexión de red (por ejemplo, Internet, LAN, WAN, VPN, etc.). En varias realizaciones, la conexión de red puede ser una conexión física "cableada". Por ejemplo, el secuenciador de ácidos nucleicos puede estar conectado de forma comunicativa (a través de Categoría 5 (CAT5), fibra óptica o cableado equivalente) a un servidor de datos que puede conectarse de forma comunicativa (a través de CAT5, fibra óptica o cableado equivalente) a través de Internet y al almacenamiento de datos de secuencias de muestra. En varias realizaciones, la conexión de red puede ser una conexión de red inalámbrica (por ejemplo, Wi-Fi, WLAN, etc.), por ejemplo, utilizando un formato de transmisión 802.11b/g o equivalente. En la práctica, la conexión de red utilizada depende de los requisitos particulares del sistema. En varias realizaciones, el almacenamiento de datos de secuencias de muestra puede ser una parte integrada del secuenciador de ácidos nucleicos.

En varias realizaciones, el almacenamiento de datos de secuencias de muestra puede ser cualquier dispositivo, sistema o implementación de almacenamiento de bases de datos (por ejemplo, partición de almacenamiento de datos, etc.) que está configurado para organizar y almacenar datos de lecturas de secuencias de ácidos nucleicos generados por el secuenciador de ácidos nucleicos de tal manera que los datos pueden buscarse y recuperarse manualmente (por ejemplo, por un administrador de base de datos/operador de cliente) o automáticamente a través de un programa informático/aplicación/script de software. En varias realizaciones, el almacenamiento de datos de referencia puede ser cualquier dispositivo de base de datos, sistema de almacenamiento, o implementación (por ejemplo, partición de almacenamiento de datos, etc.) que esté configurado para organizar y almacenar secuencias de referencia (por ejemplo, genoma completo/parcial, exoma completo/parcial, etc.) de tal manera que los datos puedan buscarse y recuperarse manualmente (por ejemplo, por un administrador de base de datos/operador de cliente) o automáticamente a través de un programa informático/aplicación/script de software. En varias realizaciones, los datos de lecturas de secuencias de ácidos nucleicos de muestra pueden almacenarse en el almacenamiento de datos de secuencias de muestra y/o el almacenamiento de datos de referencia en una variedad de tipos/formatos de archivos de datos diferentes, incluyendo, pero no limitados a: *.fasta, *.csfasta, *seq.txt, *qseq.txt, *.fastq, *.sff, *prb.txt, *.sms, *srs y/o *.qv.

En varias realizaciones, el almacenamiento de datos de secuencias de muestra y el almacenamiento de datos de referencia son dispositivos/sistemas autónomos independientes o implementados en diferentes dispositivos. En varias realizaciones, el almacenamiento de datos de secuencias de muestra y el almacenamiento de datos de referencia se implementan en el mismo dispositivo/sistema. En varias realizaciones, el almacenamiento de datos de secuencias de muestra y/o el almacenamiento de datos de referencia pueden implementarse en el dispositivo/servidor/nodo de computación analítica.

El dispositivo/servidor/nodo de computación analítica puede estar en comunicación con el almacenamiento de datos de secuencias de muestra y el almacenamiento de datos de referencia directamente a través de un cable de datos (por ejemplo, cable serial, conexión de cable directo, etc.) o enlace de bus o, alternativamente, a través de una conexión de red (por ejemplo, Internet, LAN, WAN, VPN, etc.). En varias realizaciones, el dispositivo/servidor/nodo de computación analítica puede alojar un motor de mapeo de referencia, un módulo de mapeo de novo y/o un motor de análisis terciario. En varias realizaciones, el motor de mapeo de referencia puede configurarse para obtener lecturas de secuencias de ácidos nucleicos de muestra del almacenamiento de datos de muestra y mapearlas contra una o más secuencias de referencia obtenidas del almacenamiento de datos de referencia para ensamblar las lecturas en una secuencia que es similar pero no necesariamente idéntica a la

secuencia de referencia usando todas las variedades de técnicas y métodos de mapeo/alineación de referencia. La secuencia reensamblada puede luego analizarse adicionalmente mediante uno o más motores de análisis terciarios opcionales para identificar diferencias en la composición genética (genotipo), expresión génica o estado epigenético de los individuos que pueden dar como resultado grandes diferencias en las características físicas (fenotipo). Por ejemplo, en varias realizaciones, el motor de análisis terciario puede configurarse para identificar varias variantes genómicas (en la secuencia ensamblada) debidas a mutaciones, recombinación/cruce, o deriva genética. Los ejemplos de tipos de variantes genómicas incluyen, pero no están limitados a: polimorfismos de nucleótido único (SNP), variaciones en el número de copias (CNV), inserciones/deleciones (Indels), inversiones, etc.

El módulo de mapeo de novo opcional puede configurarse para ensamblar lecturas de secuencias de ácidos nucleicos de muestra a partir del almacenamiento de datos de muestras en secuencias nuevas y previamente desconocidas.

Debe entenderse, sin embargo, que los varios motores y módulos alojados en el dispositivo/servidor/nodo de computación analítica pueden combinarse o contraerse en un único motor o módulo, dependiendo de los requisitos de la aplicación particular o de la arquitectura del sistema. Además, en varias realizaciones, el dispositivo/servidor/nodo de computación analítica puede alojar motores o módulos adicionales según lo necesite la aplicación particular o la arquitectura del sistema.

En varias realizaciones, los motores de mapeo y/o de análisis terciario están configurados para procesar las lecturas de secuencias de ácidos nucleicos y/o de referencia en el espacio de proporción de señal. En varias realizaciones, los datos de lecturas de secuenciación de ácidos nucleicos de muestra y secuencias referenciadas pueden suministrarse al dispositivo/servidor/nodo de computación analítica en una variedad de tipos/formatos de archivos de datos de entrada diferentes, incluyendo, pero no limitados a: *.fasta, *.csfasta, *seq.txt, *qseq.txt, *.fastq, *.sff, *prb.txt, *.sms, *srs y/o *.qv.

Usos

La tecnología proporciona el uso de los métodos de la tecnología, o las composiciones de la tecnología, para secuenciar y/o re-secuenciar moléculas de ácido nucleico para la monitorización de la expresión génica, elaborar perfiles de diversidad genética, diagnóstico, selección, secuenciación del genoma completo, descubrimiento y puntuación de polimorfismos del genoma completo, o cualquier otra aplicación que implique el análisis de ácidos nucleicos cuando la información de secuencia o de secuencia parcial sea relevante.

Kits

Un aspecto adicional de la invención proporciona un kit para su uso en secuenciación, re-secuenciación, monitorización de la expresión génica, elaboración de perfiles de diversidad genética, diagnóstico, selección, secuenciación del genoma completo, descubrimiento y puntuación de polimorfismos del genoma completo, o cualquier otra aplicación que implique la secuenciación de ácidos nucleicos. En algunas realizaciones, los kits comprenden por lo menos un nucleótido o análogo de nucleótido marcado, marcado de acuerdo con la tecnología descrita.

Ejemplos

Ejemplo 1 - Secuenciación de ADN de próxima generación con nucleótidos degenerados

Durante el desarrollo de las realizaciones de la tecnología, se realizaron experimentos en los que se determinó una secuencia degenerada usando una mezcla de nucleótidos degenerada y la secuencia degenerada se usó para identificar una secuencia objetivo de ADN.

Materiales y métodos

Se construyó una biblioteca de ADN a partir de los productos de una reacción de amplificación del genoma completo de transcriptasa inversa. La plantilla era el virus de ARN, MS2, que es un bacteriófago bien descrito. Esta biblioteca ha sido secuenciada anteriormente usando secuenciación convencional de 4 bases (por ejemplo, tecnología de secuenciación Ion Torrent). El producto de la amplificación del genoma completo se secuenció con la tecnología de secuenciación de ADN de Próxima Generación Ion Torrent con y sin bases degeneradas. Para este experimento, las bases de adenosina (A) y citosina (C) se mantuvieron separadas mientras que las bases de guanina (G) y timina (T) se mezclaron entre sí (por ejemplo, G/T degenerada) y la mezcla se usó en lugar de tanto G como T. Aparte de las bases mixtas G/T, se utilizaron condiciones estándar para la secuenciación con la plataforma Ion Torrent.

Resultados

Las reacciones de secuenciación de ADN tuvieron éxito en generar 64 megabases (64 millones de bases) de datos de secuencia. La secuencia de datos comprende 573.000 lecturas totales que tienen una longitud de lectura media de 116 pb. Se usaron dos lecturas experimentales elegidas aleatoriamente adquiridas usando realizaciones de la tecnología (Figura 2, "lectura 2.1" y Figura 3, "lectura 3.1") para demostrar la capacidad de mapear las secuencias adquiridas por el experimento con el genoma MS2 conocido (número de acceso NC_001417.2).

Como se muestra para cada lectura a continuación en la Figura 2 y en la Figura 3, la "secuencia observada" es la lectura de secuencia generada por el secuenciador Ion Torrent bajo condiciones estándar. Como se muestra para cada lectura a continuación en la Figura 2 y la Figura 3, la "secuencia degenerada" se generó bajo las condiciones experimentales en las que las G y las T se mezclaron entre sí y se usaron en lugar de tanto G como T. La secuencia degenerada se muestra usando el código de única letra degenerado de K, que denota una posición en la que se encuentra G o T. El software del secuenciador de Ion Torrent llamó a las K degeneradas como G o T ya que el software y el sistema de Ion Torrent no están diseñados para usar bases mixtas; como tal, los resultados se convirtieron manualmente para usar la K de código degenerado. La "homología" indicada denota el emparejamiento de la "secuencia degenerada" con el genoma MS2. Para la lectura 2.1, la alineación de la lectura con el genoma MS2 identificó más de 180 bases, lo que se corresponde con una precisión del 94%. Para la lectura 3.1, la alineación de la lectura con el genoma MS2 identificado más de 193 bases, que se corresponde con una precisión del 90%.

El análisis de los datos indicó que los errores se debían al secuenciador usado y no específicamente al resultado de los nucleótidos degenerados (mixtos).

En resumen, los datos recogidos en este experimento demostraron que el uso de nucleótidos degenerados con una tecnología de secuenciación de próxima generación identifica correctamente un objetivo.

Varias modificaciones y variaciones de las composiciones, métodos, sistemas y usos de la tecnología descritos serán evidentes para los expertos en la técnica sin apartarse del alcance y el espíritu de la tecnología tal como se describe. Aunque la tecnología se ha descrito en relación con realizaciones ejemplares específicas, debe entenderse que la invención tal como se reivindica no debería estar indebidamente limitada a tales realizaciones específicas. De hecho, varias modificaciones de los modos descritos para llevar a cabo la invención que son obvias para los expertos en campos relacionados se pretende que esté dentro del alcance de las reivindicaciones siguientes.

REIVINDICACIONES

1. Un método para identificar un ácido nucleico en una muestra, el método comprendiendo:

(a) determinar una secuencia degenerada de dos bases del ácido nucleico objetivo en la muestra, usando un método de secuenciación que determina una secuencia degenerada de dos bases del ácido nucleico objetivo sin determinar una secuencia de cuatro bases del ácido nucleico objetivo, en donde la secuencia degenerada de dos bases del ácido nucleico objetivo se determina sin determinar o conocer de otra manera la secuencia de cuatro bases del ácido nucleico objetivo; y
(b) comparar la secuencia degenerada de dos bases del ácido nucleico objetivo en la muestra determinada en el paso (a) con una secuencia de referencia conocida para identificar el ácido nucleico objetivo, en donde:

(i) el código degenerado de dos bases consiste de un primer elemento que representa una base de purina y un segundo elemento que representa una base de pirimidina;
(ii) el código degenerado de dos bases consiste de un primer elemento que representa una base ceto y un segundo elemento que representa una base amino; o
(iii) el código degenerado de dos bases consiste de un primer elemento que representa una base fuertemente enlazadora de hidrógeno y un segundo elemento que representa una base débilmente enlazadora de hidrógeno.

2. El método de la reivindicación 1 en donde el código degenerado de dos bases:

(i) consiste de un primer elemento que representa una base que comprende adenina (A) o guanina (G) y un segundo elemento que representa una base que comprende citosina (C) o timina (T);
(ii) consiste de un primer elemento que representa una base que comprende A o C y un segundo elemento que representa una base que comprende G o T; o
(iii) consiste de un primer elemento que representa una base que comprende G o C y un segundo elemento que representa una base que comprende A o T.

3. El método de la reivindicación 2 en donde:

(i) las purinas A y G están marcadas con un primer marcador y las pirimidinas C y T están marcadas con un segundo marcador;
(ii) las bases ceto G y T están marcadas con un primer marcador y las bases amino A y C están marcadas con un segundo marcador; o
(iii) las bases fuertemente enlazadoras de hidrógeno C y G están marcadas con un primer marcador y las bases débilmente enlazadoras de hidrógeno A y T están marcadas con un segundo marcador.

4. El método de la reivindicación 1 que comprende además proporcionar un primer nucleótido y un segundo nucleótido en donde el primer nucleótido está marcado con un marcador y el segundo nucleótido está marcado con dicho marcador.

5. El método de la reivindicación 1 que comprende además proporcionar un primer nucleótido, un segundo nucleótido, un tercer nucleótido y un cuarto nucleótido, en donde el primer nucleótido está marcado con un primer marcador, el segundo nucleótido está marcado con dicho primer marcador, el tercer nucleótido está marcado con un segundo marcador, y el cuarto nucleótido está marcado con dicha segundo marcador.

6. El método de la reivindicación 5 en donde el primer marcador es una primera fracción fluorescente y en donde el segundo marcador es una segunda fracción fluorescente.

7. El método de la reivindicación 1 que comprende además proporcionar un análogo de nucleótido marcado en donde la base del análogo de nucleótido marcado se empareja con un primer nucleótido o un segundo nucleótido de acuerdo con una regla de emparejamiento de bases degenerada.

8. El método de la reivindicación 1 que comprende además proporcionar un primer análogo de nucleótido marcado y un segundo análogo de nucleótido marcado en donde la base del primer análogo de nucleótido marcado se empareja con un primer nucleótido o un segundo nucleótido y la base del segundo análogo de nucleótido marcado se empareja con un tercer nucleótido o un cuarto nucleótido.

9. El método de la reivindicación 8 en donde el primer análogo de nucleótido marcado está marcado con una primera fracción fluorescente y en donde el segundo análogo de nucleótido marcado está marcado con una segunda fracción fluorescente.

10. El método de cualquier reivindicación anterior en donde la determinación comprende medir una característica física, química y/o electrónica de una base y diferenciar entre una base de purina y una base de pirimidina, entre una

base ceto y una base amino, y/o entre una base fuertemente enlazadora de hidrógeno y una base débilmente enlazadora de hidrógeno.

11. El método de la reivindicación 6 o la reivindicación 9 en donde la determinación comprende la detección de la emisión fluorescente de un marcador de nucleótido.

12. El método de cualquier reivindicación anterior en donde se usa el método de secuenciación basado en conjuntos.

13. El método de cualquier reivindicación anterior en donde se usa el método de secuenciación basado en moléculas individuales.

14. El método de cualquier reivindicación anterior en donde el ácido nucleico objetivo en la muestra es un ARNm, ARN ribosómico, ADN genómico, o ADNc.

A

aactgagcggcagcagcgagacctggcctactgtgtgtcacagctgcccctcacagagcgaggcctccgtaagat
gcttgacaattttgactgttttggagacaaactgtcagatgagtccatcttcagtgctttttgtcagttgtagg
caagctgcgacgtggggccaagcctgagggcaag

B

rryyrrryrryrryrryrrrryyyrryyyryyryryryyryrryyryyyyyyryrrrryrrrryyyyyryrrrry
ryyyrryrryyyyrryyryyyyyrrrrryrrryyryyrrryrrryyyryyyyyyrryryyyyyyyryyrryyryrrr
yrrryyryrryryrrrryyrrryyyrrrrryrrr

C

mmmkkmkmkkmmkmmkmkmkmmmkkkmmkmmkkkkkkkmmmmkmkkmmmmkmmmmkmkmkmkkmmkmmkkmmkmk
kmkkkmmmmmkkkkkmmkkkkkkkkmkmmmmmmmkkkmmkmkkmkkmmmkmkkmmkkkmkkkkkkkkmmkkkkkmkk
mmmkmkkmkmmkkkkkkmmmmkmmkkmkkkmmmk

Figura 1

observed sequence

CTTCTATCGGATTTTTAAGGCAAGGCAAGGGCTCCGAAAATATTTAACCCGGAGGCCCGGCAGCAATCGCATCAACTACCGGC
GAGCTACGAATAGACGCCGGCCAGGCAAATCATTATTACGGGACCCAGGGCTAAGATCAACAAATAGGCAAGGCGGGAGGATT
GATTCGCCGCTCGA

secuencia degenerada

CKKCKAKCKKAKKKKAAKKCAAKKCAAKKKCKCCKAAAAKAKKKAACCCKKAKKCCCKKCAKCAAKCKCAKCAACKACCKKC
KAKCKACKAAKAKACKCCKKCCAKKCAAAKCAKKAKKACKKKACCCAKKKCKAAKAKCAACAAAKAKKCAAKKCKKKAKKAKK
KAKKCKCCKCKCKA

Homología identificada sobre 180 bases (94% de precisión)

secuencia degenerada

CKKCKAKCKKAKKKKAAKKCAAKKCAAKKKCKCCKAAAAKAKKKAA-CCCKKAKKCCCKKCAKCAAKCK
||||||||||||||||||||||||||||||||||||||||||||||| |||| ||||||| |||||||||
CTGCGAGCTTATTGTTAAGGCAATGCAAGGTCTCCTAAAAGATGGAAACCCG-ATTCCCT-CAGCAATCG

Secuencia MS2

secuencia degenerada (continuación)

-AKCAACKACCKKCKAKCKACKAAKAKACKCCKKCCAKKCAAAKCAKKAKKACKKKACCCAKKKCKAAKA
 ||||| ||||| ||||||||||||||||||||||||||||||||||||| ||||||||||||||||
CAGCAAACTCCGGC-ATCTACTAATAGACGCCGGCCATTCAAA-CATGAGGA--TTACCCATGTCGAAGA

Secuencia MS2 (continuación)

secuencia degenerada (continuación)

KCAACAAAKA-KK-CAAKKCKKKAKK-AKKKAKKCKCCKCKCKA
 ||||||||| || ||| |||||||| |||||| |||||||||
-CAACAAAGAAGTTCAACTCTTTATGTATTGATCTTCCTCGCGA

Secuencia MS2 (continuación)

Figura 2

secuencia observada

AGATTTAATTACGAACGCCAGGCTTCGACATTAAGCTCGACACCACCAACATTCGGGGGGGGCCTACGGAGGGCAGCCTCGAC
GGGATTTTTAGGGGCTAGGGGCGGGCTCATATCGCGGACGAAGCTACATTTACGGGGAATTCCGGGGAACGCGGAGGGTTATT
CGATCGCTAGCGAGCCCTTATCGAATT

secuencia degenerada

AKAKKKAAKKACKAACKCCAKKCKKCKACAKKAAKCKCKACACCACCAACAKKCKKKKKKKKCCKACKKAKKKCAKCCKCKAC
KKKAKKKKKAKKKKCKAKKKKCKKKCKCAKAKCKCKKACKAAKCKACAKKKACKKKKAAKKCCKKKKAACKCKKAKKKKAKK
CKAKCKCKAKCKAKCCCKKAKCKAAKK

Homología identificado sobre 193 bases (90% de precisión)

secuencia degenerada (continuación)

AKAKKKAAKKACKAACKCCAKKCKKCKACAKKAAKCKCKACACCACCAACAKKCKKKKKKKKCCKACKK-
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| || |||||
ATATTTAAGTACGAACGCCATGCGGCTACAGGAAGCTCTACACCACCAACAGTCTGGGTTG-CC-ACTTT

Secuencia MS2

secuencia degenerada

AKKKCAKCCKCKACKKK-AKKKKK-AKKKKCKAKKKKCKKKCKCAKAKCKCKKACKAA-KCKACAKKK-A
||| || ||||||||||||||||||||||||||||||| |||||||||||||||||| |||||||||| |
AGG-CA-CCTCGACTTTGATGGTGTATTTGCGATTCT---GCGCAGAGCTCTGACGAACGCTACAGGTTA

Secuencia MS2

secuencia degenerada

CKKKK-AAKKCCKKKKAACKCKKAKKKKKAKKCKAKC--KCKAKCKAKCCC-KKAKCKAAKK
||||| ||| |||||||||||| |||||||| |||| | ||||| ||| ||||||||||
CTTTGTAAG-CCTGTGAACGCG-AGTTAGAGCTGATCCATTCAGCGACCCCGTTAGCGAAGT

Secuencia MS2

Figura 3