(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 490 601**

(51) Int. CI.:

C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **15.11.2007 E 07871488 (8)**

(97) Fecha y número de publicación de la concesión europea: **07.05.2014 EP 2082063**

(54) Título: **Secuenciación multi-etiqueta y análisis ecogenómico**

(30) Prioridad:

15.11.2006 US 858948 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
04.09.2014

(73) Titular/es:

BIOSPHEREX LLC (100.0%)
15 WILTSHIRE COURT EAST SUITE 200
POTOMAC FALLS, VA 20165-5677, US

(72) Inventor/es:

GILLEVET, PATRICK M.

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Secuenciación multi-etiqueta y análisis ecogenómico

Solicitudes relacionadas

Esta solicitud es una continuación de parte y reivindica el beneficio completo de prioridad de la solicitud provisional de los Estados Unidos número 60/858.948, presentada el 15 de Noviembre de 2006 por Patrick Gillevet para Multitag Sequencing and Ecogenomics Analysis.

Campo de la Invención

La invención se refiere a métodos múltiplex para determinar cuantitativamente polinucleótidos diana de dos o más muestras.

Declaración de Derechos del Gobierno

El trabajo descrito en la presente memoria se realizó en parte con apoyo del gobierno con la subvención Núm. 1R43DK074275-01A2 otorgada por el Instituto Nacional de la Diabetes y las Enfermedades Digestivas y del Riñón de los Estados Unidos, y por lo tanto el Gobierno de los Estados Unidos puede tener ciertos derechos en la invención.

Antecedentes

Las tecnologías para la determinación de secuencias de proteínas, ARN y ADN, han sido fundamentales en el desarrollo de la biología molecular moderna. Durante los últimos quince años, la secuenciación del ADN en particular ha sido la tecnología central en una revolución en curso en el alcance y la profundidad de comprensión de la organización y la función genómica. El desarrollo continuo de la tecnología de secuenciación está, quizás, mejor simbolizado por la determinación de la secuencia completa de un genoma humano.

El proyecto de secuenciación del genoma humano sirvió para varios propósitos. Sirvió como plataforma para el desarrollo programático de la mejora de las tecnologías de secuenciación y de los esfuerzos de secuenciación del genoma. También sirvió para establecer un marco para la producción y distribución de la información de la secuenciación de proyectos de secuenciación a escala cada vez más grandes. Estos proyectos proporcionaron secuencias genómicas completas de una sucesión de organismos modelo de complementos genéticos cada vez más grandes. Estos logros, que culminaron con la realización de una secuencia del genoma humano, destacan el poder muy considerable y el rendimiento de la tecnología de secuenciación contemporánea.

Al mismo tiempo, sin embargo, ponen de relieve las limitaciones de la tecnología actual y la necesidad de considerables mejoras en la velocidad, la precisión y el coste antes de la secuenciación puede ser explotada plenamente en la investigación y la medicina. Entre las áreas que se puede ver más fácilmente que exigen avances en la tecnología de secuenciación se encuentran las aplicaciones de secuenciación clínicas que requieren información del genoma completo, las aplicaciones ambientales que involucran a múltiples organismos en mezclas, y las aplicaciones que requieren el procesamiento de muchas muestras. Estas son, por supuesto, solo unas pocas entre un gran número de áreas que se requieren o se beneficiarán en gran medida de métodos de secuenciación más capaces y menos costosos.

Hasta la fecha, prácticamente todas las secuenciaciones se han realizado mediante métodos de elongación de la cadena de Sanger. Todos los métodos de Sanger requieren la separación de los productos de elongación con una resolución de una única base. Actualmente, si bien todavía se utiliza PAGE para este propósito en algunos secuenciadores comerciales, la electroforesis capilar es el método de elección para secuenciadores de ADN de alto rendimiento. Los métodos de separación tanto basados en gel como basados en capilares consumen mucho tiempo, son costosos, y tienen un rendimiento limitado. Los métodos basados en chips, tales como Affymetrix GeneChips y secuenciación Hyseq mediante métodos de hibridación, requieren chips que pueden ser producidos solo por procesos de fabricación de capital intensivo y complejos. Estas limitaciones representan obstáculos a la utilización de secuenciación para muchos propósitos, tales como los descritos anteriormente. En parte para superar las limitaciones impuestas por la necesidad de técnicas de separación poderosas en métodos de secuenciación de terminación de la cadena y los requisitos de fabricación de los métodos basados en chips, se están desarrollando actualmente numerosas tecnologías que no requieren la separación de los productos de elongación con resolución íntegra y no requieren chips.

Una tecnología avanzada de este tipo es un método basado en cuentas, amplificación en emulsión, y pirosecuenciación desarrollado por 454 Life Sciences. (Véase Marguilles, et al. (2005) Nature 437: 376, particularmente en cuanto a los métodos mencionados anteriormente. El método utiliza una serie de etapas para depositar moléculas individuales de ADN amplificado en pocillos individuales de una placa que contiene varios millones de pocillos de picotitulación. Las etapas aseguran que cada pocillo de la placa o bien no contenga ADN o bien contenga el ADN amplificado a partir de una sola molécula original. La pirosecuenciación se lleva a cabo en los pocillos mediante la elongación de un molde de cebador de la misma manera que en la secuenciación de Sanger. La

pirosecuenciación no implica terminación de la cadena y no requiere la separación de los productos de elongación. En lugar de eso la secuenciación prosigue paso a paso mediante ciclos de adición de una sola base. En cada ciclo una de las cuatro bases - A, T, G, o C - se incluye en la reacción de elongación. Las otras tres bases se omiten. Se añade una base a la cadena en crecimiento si ésta es complementaria a la siguiente posición en el molde. Se produce luz cada vez que una base se incorpora a la secuencia complementaria en crecimiento. Interrogando con cada una de A, C, G o T sucesivamente, se puede determinar la identidad de la base en cada posición. Las reacciones de secuenciación se llevan a cabo en muchos pocillos simultáneamente. Las señales se recogen de todos los pocillos a la vez utilizando un detector de formación de imágenes. Por lo tanto, se puede determinar al mismo tiempo una multitud de secuencias.

En principio, cada pocillo que contiene un ADN emitirá una señal para solo una de las cuatro bases para cada posición. En la práctica, rondas de la misma base en dos o más posiciones sucesivas conducen a la emisión de señales proporcionalmente más fuertes para la primera posición en la ronda. En consecuencia, la lectura de la secuencia de un pocillo determinado es un poco más complicada señalando a continuación simplemente, para cada posición, cuál de las cuatro bases se añade. Sin embargo, puesto que las señales son proporcionales al número de incorporaciones, las secuencias pueden ser reconstruidas con precisión a partir de la intensidad de la señal para la mayoría de las rondas.

Se ha demostrado que la tecnología lee con precisión un promedio de alrededor de 250 bases por pocillo con una precisión aceptable. Un dispositivo ofrecido por 454 Life Sciences utiliza actualmente una "placa" de pocillos de picotitulación de 6,4 cm^2 que contiene 1.600.000 pocillos de tamaño de picolitros para la secuenciación de aproximadamente 400.000 moldes diferentes. El rendimiento para una sola ronda utilizando esta placa actualmente es de aproximadamente 100 millones de bases en cuatro horas. Aunque se trata de un dispositivo de primera generación, su rendimiento es casi 100 veces mejor que los dispositivos de secuenciación de Sanger convencionales.

Se están desarrollando muchos otros métodos para la secuenciación de ultra-alto rendimiento por otras instituciones y empresas. La secuenciación mediante métodos de síntesis que dependen de la amplificación de la diana están siendo desarrollados y/o comercializados por George Church de la Universidad de Harvard, por Solexa, y por otros. Se han desarrollado y/o están siendo comercializados métodos de secuenciación mediante ligación por Applied Biosystems y Solexa, entre otros. Los métodos de matrices y secuenciación por hibridación están disponibles comercialmente y/o están siendo desarrollados por Affymetrix, Hyseq, Biotrove, Nimblegen, Illumina, y otros. Los métodos de secuenciación de moléculas individuales están siendo buscados por Helicos basándose en la secuenciación mediante síntesis y U.S. Geomics (entre otros) basándose en la poración.

Estos métodos representan una mejora considerable en el rendimiento respecto a los métodos anteriores, en algunos aspectos. Y prometen una mejora considerable en la economía también. Sin embargo, en la actualidad son muy costosos de implementar y utilizar, se limitan a lecturas relativamente cortas y, aunque en paralelo a gran escala, tienen limitaciones que deben superarse para alcanzar su pleno potencial.

Una desventaja concreta de estos métodos, por ejemplo, es que las muestras deben ser procesadas en serie, reduciendo el rendimiento y aumentando de coste. Esto es particularmente un gran inconveniente cuando se están procesando grandes cantidades de muestras, como puede ser el caso en los estudios clínicos y la toma de muestras ambientales, por nombrar solo dos aplicaciones. La incorporación de secuencias de indexación por ligación a las bibliotecas por perdigonazo "shotgun" aleatorias se ha descrito en los números de patente de los Estados Unidos: 7264929, 7244559, y 7211390, pero los métodos de ligación directa descritos allí distorsionan la distribución de los componentes dentro de las muestras (como se ilustra en la Figura 4 en la presente memoria) y por lo tanto son inapropiados para la enumeración de componentes dentro de cada muestra.

El documento WO02/061143 se refiere a un análisis comparativo de ácidos nucleicos utilizando etiquetado poblacional. El documento WO2005/068656 se refiere a la caracterización de ácidos nucleicos, en particular a un método de secuenciación y distinción entre secuencias de ácidos nucleicos procedentes de diferentes fuentes en una matriz. El documento WO 2005/042759 se refiere a métodos para el análisis de la expresión génica utilizando tecnología basada en micromatrices.

Por lo tanto, existe una necesidad de mejorar el rendimiento de la muestra, reducir los costes de secuenciación de polinucleótidos de muchas muestras al mismo tiempo, y enumerar con precisión los componentes de muestras analizadas mediante técnicas paralelizadas y múltiplex de alto rendimiento.

Compendio

Es por tanto un objeto de la presente invención proporcionar métodos de secuenciación con mejora de rendimiento de la muestra. Los siguientes párrafos describen algunas realizaciones ilustrativas de la invención que ilustran algunos de sus aspectos y características. No son exhaustivas en la ilustración de sus muchos aspectos y realizaciones, y por lo tanto no son en modo alguno limitantes de la invención. Muchos otros aspectos, características y realizaciones de la invención se describen en la presente memoria. Muchos otros aspectos y realizaciones serán fácilmente evidentes para los expertos en la técnica después de leer la solicitud y prestarle la

debida atención a la luz de la técnica anterior y el conocimiento en el campo. La materia sujeto para la que se solicita protección es la definida en las reivindicaciones.

En particular, la presente invención proporciona un método múltiplex para determinar cuantitativamente polinucleótidos diana de dos o más muestras, que comprende: anclar una primera secuencia etiqueta específica de la muestra a uno o más polinucleótidos de una primera muestra; unir una segunda secuencia etiqueta específica de la muestra diferente de dicha primera secuencia etiqueta a uno o más polinucleótidos de una segunda muestra; mezclar las muestras etiquetadas juntas; y secuenciar dichos polinucleótidos que comprenden dicha primera y dicha segunda etiquetas; en donde las etiquetas específicas de la muestra están incluidas en los cebadores para la amplificación por PCR y las etiquetas específicas de la muestra se anclan a los polinucleótidos por medio de amplificación mediante PCR, y en donde los cebadores para la amplificación mediante PCR comprenden, en orden 5' a 3', un radical para la inmovilización y/o una secuencia para la amplificación mediante PCR, la etiqueta de secuencia específica de la muestra y una secuencia sonda específica de una secuencia diana localizada 3' con respecto a una región genética variable; el método comprende adicionalmente las etapas de: a partir de las secuencias etiqueta incluidas en las secuencias de polinucleótidos determinadas de este modo identificar la muestra en la que se produjeron las secuencias de polinucleótidos; a partir de las secuencias de la región genética variable incluida en las secuencias de polinucleótidos determinadas de este modo identificar las variantes concretas de dicho elemento genético variable; a partir de esta información determinar el número de veces que una o más variantes determinadas aparecen en cada muestra; y a partir del número de cada variante en los polinucleótidos determinados de este modo, cuantificar dichos polinucleótidos en dichas muestras.

Las realizaciones proporcionan métodos múltiplex para la determinación cuantitativa de polinucleótidos en dos o más muestras, que comprenden:

hibridar un primer cebador con los polinucleótidos en una primera muestra, comprendiendo dicho primer cebador una primera secuencia etiqueta y una primera secuencia sonda específica para una primera secuencia diana, en donde dicha primera secuencia diana se encuentra en posición 3' con respecto a una región genética variable;
elongar moldes de cebadores formados de esta manera para formar una primera población de polinucleótidos etiquetados que comprende: dicho primer cebador que incluye dicha primera secuencia etiqueta; y secuencias de dicha región genética variable;
hibridar un segundo cebador con los polinucleótidos en una segunda muestra, comprendiendo dicho segundo cebador una segunda secuencia etiqueta y una segunda secuencia sonda específica para una segunda secuencia diana, en donde dicha segunda secuencia diana se encuentra en posición 3' con respecto a la misma región genética variable como dicha primera secuencia diana, en donde adicionalmente dicha segunda secuencia sonda puede ser la misma que o diferente de dicha primera secuencia sonda;
elongar los moldes de los cebadores formados de esta manera para formar una segunda población de polinucleótidos etiquetados que comprende: dicho segundo cebador que incluye dicha segunda secuencia etiqueta; y secuencias de dicha región genética variable;
mezclar entre sí dichas primera y segunda poblaciones;
determinar las secuencias de polinucleótidos que comprenden secuencias etiqueta y las secuencias del elemento genético variable en dicha mezcla;
a partir de las secuencias etiqueta incluidas en las secuencias de polinucleótidos determinadas de este modo identificar la muestra en la que se produjeron las secuencias de polinucleótidos;
a partir de las secuencias de la región genética variable incluida en las secuencias de polinucleótidos determinadas de este modo identificar las variantes concretas de dicho elemento genético variable;
a partir de esta información determinar el número de veces que una o más variantes dadas aparecen en cada muestra, y
a partir del número de cada variante en los polinucleótidos determinados de este modo, cuantificar dichos polinucleótidos en dichas muestras;
en donde dichas secuencias se determinan sin transferencia Southern y/o sin separar por tamaño los productos de extensión del cebador y/o sin electroforesis.

Las realizaciones proporcionan métodos de acuerdo con cualquiera de lo anterior o lo siguiente en donde secuencias de polinucleótidos dadas en una muestra se cuantifican mediante un método que comprende normalizar el número de apariciones determinado por la secuencia dada. En realizaciones el número de apariciones se normaliza dividiendo el número de apariciones determinado por la secuencia de polinucleótidos dada por el número total de apariciones de secuencias de polinucleótidos en la muestra. En realizaciones la secuencia de polinucleótidos dada es la de una variante dada de una región genética variable y, en realizaciones, la cantidad de la variante dada en la muestra se normaliza dividiendo el número de apariciones de esa variante por el número total de apariciones de todas las variantes de la región genética variable en la muestra.

Las realizaciones proporcionan un método múltiplex para determinar secuencias de polinucleótidos en dos o más muestras, que comprende: anclar una primera secuencia etiqueta a uno o más polinucleótidos de una primera muestra; anclar una segunda secuencia etiqueta diferente de dicha primera secuencia etiqueta a uno o más polinucleótidos de una segunda muestra; mezclar entre sí los polinucleótidos etiquetados de dichas primera y

segunda muestras; determinar las secuencias de dichos polinucleótidos que comprenden dicha primera y dicha segunda etiquetas; e identificar dichas primera y segunda etiquetas en dichas secuencias; identificando de este modo las secuencias de dichos polinucleótidos de dichas primera muestra y segunda muestra, en donde dichas secuencias se determinan sin transferencia Southern y/o sin separar por tamaño los productos de extensión del cebador y/o sin electroforesis.

Las realizaciones proporcionan un método múltiplex para determinar secuencias de polinucleótidos en dos o más muestras que comprende:

anclar una primera secuencia etiqueta, t_1, a polinucleótidos $P_{1\text{-}1}$ a $P_{1\text{-}N1}$ en una primera muestra, para proporcionar de este modo una primera pluralidad de polinucleótidos etiquetados con dicha primera etiqueta, $t_1P_{1\text{-}1}$ a $t_1P_{1\text{-}N1}$;
anclar una segunda secuencia etiqueta, t_2, a polinucleótidos $P_{2\text{-}1}$ a $P_{2\text{-}N2}$ en una segunda muestra, para proporcionar de este modo una segunda pluralidad de polinucleótidos etiquetados con dicha segunda etiqueta, $t_2P_{2\text{-}1}$ a $t_2P_{2\text{-}N2}$;
mezclar entre sí dichos polinucleótidos etiquetados con dicha primera y dicha segunda etiquetas;
determinar las secuencias de polinucleótidos que comprenden dichas etiquetas en dicha mezcla;
identificar dichas primera y segunda etiquetas en dichas secuencias y;
mediante dicha primera etiqueta identificar las secuencias de polinucleótidos de dicha primera muestra y mediante dicha segunda etiqueta identificar las secuencias de polinucleótidos de dicha segunda muestra;
en donde dichas secuencias se determinan sin transferencia Southern y/o sin separar por tamaño los productos de extensión del cebador y/o sin electroforesis.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde el número de dichos polinucleótidos en dicha primera muestra, n_1, es cualquiera de 2, 5, 10, 25, 50, 100, 150, 200, 250, 500, 1.000, 1.500, 2.000, 2.500, 5.000, 7.500, 10.000, 12.500, 15.000, 17.500, 20.000, 25.000, 30.000, 35.000, 40.000, 50.000, 75.000, 100.000, 150.000, 200.000, 250.000, 500.000, 1.000.000 o más, y el número de dichos polinucleótidos en dicha segunda muestra, n_2, es cualquiera de 2, 5, 10, 25, 50, 100, 150, 200, 250, 500, 1.000, 1.500, 2.000, 2.500, 5.000, 7.500, 10.000, 12.500, 15.000, 17.500, 20.000, 25.000, 30.000, 35.000, 40.000, 50.000, 75.000, 100.000, 150.000, 200.000, 250.000, 500.000, 1.000.000 o más.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde el número de dichas muestras y de dichas diferentes etiquetas para las mismas es de 5, 10, 15, 20, 25, 50, 75, 100, 150, 200, 250, 500, 1.000, 2.500, 5.000, 10.000 o más.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las etiquetas son secuencias de nucleótidos que tienen 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36 nucleótidos de longitud o cualquier combinación de las mismas.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las etiquetas se incorporan a dichos polinucleótidos mediante una etapa de amplificación.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde dichas etiquetas se incluyen en cebadores para la amplificación y se incorporan a dichos polinucleótidos mediante amplificación utilizando dichos cebadores.

La presente descripción proporciona un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las etiquetas se incluyen en adaptadores para la amplificación y dichos adaptadores se ligan a polinucleótidos en dichas muestras. Las realizaciones proporcionan un método a este respecto, en donde adicionalmente, dichos polinucleótidos ligados de ese modo a dichas etiquetas se amplifican a través de dichos adaptadores. Las realizaciones proporcionan un método a este respecto, en donde adicionalmente, dichos adaptadores comprenden un radical para la inmovilización. En realizaciones dicho radical es un ligando; en realizaciones es biotina. Las realizaciones proporcionan un método a este respecto, en donde adicionalmente, dichas etiquetas se incluyen en adaptadores para la amplificación en emulsión de cuentas. En realizaciones los adaptadores son adecuados para su uso en un sistema de secuenciación de 454 Life Sciences u otro sistema de secuenciación en el que se lleva a cabo la amplificación en emulsión de cuentas.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde el cebador para la amplificación comprende una secuencia para la amplificación mediante PCR, la amplificación lineal, la amplificación transcripcional, la replicación de círculo rodador, o la replicación QB.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde el cebador para la amplificación comprende una secuencia para la amplificación mediante PCR.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde cada

uno de dichos polinucleótidos se dispone individualmente en una cuenta aislado de los otros polinucleótidos.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde cada uno de dichos polinucleótidos se dispone individualmente sobre una cuenta aislado del resto de dichos polinucleótidos, se amplifica mientras se dispone en la misma, y sus productos de amplificación también se disponen sobre dicha cuenta.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde cada uno de dichos polinucleótidos se dispone individualmente sobre una cuenta aislado del resto de dichos polinucleótidos, se amplifica mientras se dispone en la misma, sus productos de amplificación también se disponen sobre dicha cuenta, y cada una de dichas cuentas se dispone de forma individual en un pocillo aislado del resto de dichas cuentas.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las secuencias se determinan mediante pirosecuenciación.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde dichas muestras son muestras biológicas, que comprenden cada una o más especies.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde al menos una secuencia de dichos polinucleótidos es específica de un organismo concreto.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde dichas secuencias comprenden una secuencia variable de ARNr 16S.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde dichas secuencias comprenden una secuencia variable de ARNr 18S, una secuencia variable de ARNr ITS, una secuencia mitocondrial, una secuencia de microsatélites, una secuencia enzimática metabólica, y/o una secuencia de enfermedad genética.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las muestras son muestras de la comunidad microbiana.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las muestras son muestras de la comunidad microbiana para el análisis clínico de un paciente.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las muestras son muestras ambientales de la comunidad microbiana.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las muestras son muestras de suelo de la comunidad microbiana.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las muestras son muestras de agua de la comunidad microbiana.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las muestras son muestras para el análisis de SNP.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde las muestras son muestras para la determinación del genotipo.

Las realizaciones proporcionan un método múltiplex de acuerdo con cualquiera de lo anterior o el siguiente para determinar secuencias de polinucleótidos de dos o más muestras, que comprende,
amplificar polinucleótidos de una primera muestra para producir primeros polinucleótidos amplificados que comprenden una primera secuencia etiqueta;
amplificar por separado polinucleótidos de una segunda muestra para producir segundos polinucleótidos amplificados que comprende una segunda secuencia etiqueta diferente de dicha primera secuencia etiqueta;
en donde los productos de amplificación resultantes de diferentes polinucleótidos individuales se separan espacialmente entre sí;
mezclar entre sí los amplicones de dichas primera y segunda muestras;
distribuir los amplicones en la mezcla en lugares espacialmente distintos; secuenciar los amplicones distribuidos de este modo utilizando uno o más cebadores que hibridan 5' con respecto a dichas secuencias etiqueta;
identificar dichas secuencias etiqueta en las secuencias de polinucleótidos determinadas de este modo; y
identificar mediante dichas etiquetas los polinucleótidos de dicha primera muestra y los polinucleótidos de dicha segunda muestra.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, que comprende,

(a) para cada muestra por separado: aislar los polinucleótidos que se van a secuenciar, ligar dichos polinucleótidos a un adaptador común que comprende una secuencia etiqueta, y capturar los polinucleótidos ligados individuales sobre cuentas individuales en condiciones que proporcionan predominantemente la inmovilización de 0 o 1 molécula por cuenta;
(b) después de eso mezclar entre sí dichas cuentas que comprenden dichos polinucleótidos.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, que comprende adicionalmente, amplificar polinucleótidos inmovilizados sobre cuentas en gotitas de una emulsión para amplificar clonalmente de este modo dichos polinucleótidos individuales sobre dichas cuentas, en donde la amplificación comprende la amplificación de dicha secuencia etiqueta.

Las realizaciones proporcionan un método de acuerdo con cualquiera de lo anterior o lo siguiente, que comprende adicionalmente, distribuir las gotitas individuales que contienen dichos polinucleótidos amplificados en pocillos en condiciones que proporcionan predominantemente 0 o 1 gota por pocillo, determinar en los pocillos individuales las secuencias de polinucleótidos que comprenden dichas secuencias etiqueta, y mediante dichas secuencias etiqueta identificar los polinucleótidos de dichas primera y segunda muestras.

En realizaciones la invención proporciona métodos, de acuerdo con cualquiera de lo anterior o lo siguiente, que son adecuados para una cualquiera o más de detectar, verificar, perfilar, pronosticar, y/o el diagnosticar un trastorno, enfermedad, o similares.

En realizaciones la invención proporciona métodos, de acuerdo con cualquiera de lo anterior o lo siguiente, que son adecuados para analizar la composición, diversidad, estabilidad, dinámica, y/o cambios en muestras agrícolas, alimentarias, de bioseguridad, veterinarias, clínicas, ecológicas, zoológicas, oceanográficas, y/o cualquier otra muestra que comprenda uno o más polinucleótidos.

La presente descripción proporciona kits que comprenden una pluralidad de dos o más cebadores, comprendiendo cada cebador de dicha pluralidad una secuencia etiqueta y una secuencia sonda específica de una secuencia diana, en donde:

(A) en cada uno de dichos cebadores la secuencia sonda se encuentra en posición 3' con respecto a la secuencia etiqueta, pero no necesariamente adyacente a la misma;
(B) en cada uno de dichos cebadores: la secuencia etiqueta es diferente de la secuencia etiqueta del resto en la pluralidad; la secuencia etiqueta no es la secuencia complementaria a ningún otra secuencia etiqueta en la pluralidad; la secuencia etiqueta no contiene ninguna secuencia de homodinucleótidos; las secuencias de empalme entre la secuencia etiqueta y las porciones adyacentes del cebador, si las hubiera, no es una secuencia de homodinucleotidos;
(C) en cada uno de dichos cebadores la secuencia sonda es complementaria a la secuencia diana y la secuencia diana se encuentra en posición 3' con respecto a una región genética variable, y
(D) cada uno de dichos cebadores se dispone por separado de los otros en recipientes en dicho kit.

La presente descripción proporciona kits de acuerdo con cualquiera de lo anterior o lo siguiente, en donde cada uno de dichos cebadores comprende adicionalmente una secuencia cebadora en posición 5' con respecto a la secuencia etiqueta pero no necesariamente adyacente a la misma, y la secuencia cebadora es la misma en la totalidad de dichos cebadores, comprendiendo adicionalmente dicho kit un cebador complementario y eficaz para la polimerización de dicha secuencia cebadora.

La presente descripción proporciona kits que comprenden una pluralidad de dos o más pares de cebadores, comprendiendo cada cebador en dicha pluralidad una secuencia etiqueta y una secuencia sonda específica para una secuencia diana, en donde:

(A) en cada uno de dicho cebador la secuencia sonda se encuentra en posición 3' con respecto a la secuencia etiqueta, pero no necesariamente adyacente a la misma;
(B) en cada uno de dichos cebadores: la secuencia etiqueta es diferente de la secuencia etiqueta del resto en la pluralidad; la secuencia etiqueta no es la secuencia complementaria a ningún otra secuencia etiqueta en la pluralidad; la secuencia etiqueta no contiene ninguna secuencia de homodinucleótidos; las secuencias de empalme entre la secuencia etiqueta y las porciones adyacentes del cebador, si las hubiera, no son una secuencia de homodinucleótidos;
(C) en cada uno de dichos cebadores la secuencia sonda es complementaria a la secuencia diana,
(D) en cada par de cebadores las secuencias sonda son específicas a secuencias diana que flanquean una región genética variable;
(E) cada uno de dichos cebadores se dispone por separado de los otros en dicho kit.

La presente descripción proporciona kits de acuerdo con cualquiera de lo anterior o lo siguiente, en donde, los

cebadores comprenden adicionalmente una secuencia cebadora en posición 5' con respecto a la secuencia etiqueta, pero no necesariamente adyacente a la misma, la secuencia cebadora o es la misma en todos los cebadores, o un miembro de cada par tiene la misma primera secuencia cebadora y el segundo miembro de cada par tiene la misma segunda secuencia cebadora, comprendiendo adicionalmente dicho kit dispuestos por separado el uno del otro en uno o más recipientes uno o más cebadores complementarios a y eficaces para la elongación de dicho cebado.

La presente descripción proporciona un kit útil en los métodos de acuerdo con cualquiera de lo anterior o lo siguiente, que comprende un conjunto de cebadores y/o adaptadores, en donde cada cebador y/o adaptador en dicho conjunto comprende una secuencia etiqueta y una secuencia cebadora. En realizaciones los cebadores y/o los adaptadores comprenden adicionalmente un radical para la inmovilización. En realizaciones los cebadores y/o los adaptadores comprenden biotina. En realizaciones los cebadores y/o adaptadores en el conjunto comprenden todas las secuencias etiqueta definidas por secuencias de polinucleótidos de 2, 3, 4, 5, 6, 7, u 8 bases, en donde cada uno de dichos cebadores y/o adaptadores están dispuestos en recipientes separados el uno del otro. En realizaciones hay 1-5, 3-10, 5-15, 10-25, 20-50, 25-75, 50-100, 50-150, 100-200, 150-500, 250-750, 100-1000, o más secuencias etiqueta diferentes dispuestas por separado las unas de las otras, con el fin de ser útiles para etiquetar de forma única dicho número de muestras diferentes. En realizaciones los cebadores y/o los adaptadores son adecuados para su uso como adaptadores y/o cebadores de amplificación de 454 Life Sciences. En realizaciones los cebadores y/o los adaptadores comprenden adicionalmente una cualquiera o más de una secuencia cebadora para una cualquiera o más de una secuencia de ARNr 16S, una secuencia de ARNr 18S, una secuencia ITS, una secuencia mitocondrial, una secuencia de microsatélite, una secuencia enzimática metabólica, una secuencia de enfermedad genética, y/o cualquier otra secuencia para su amplificación o análisis.

En realizaciones, la descripción proporciona un kit, de acuerdo con cualquiera de lo anterior o lo siguiente, que comprende un conjunto de cebadores y/o adaptadores para su uso en un método de acuerdo con cualquiera de lo anterior o lo siguiente, en donde cada cebador y/o adaptador en dicho conjunto comprende una secuencia etiqueta, la secuencia etiqueta de cada uno de dichos cebadores y/o adaptadores es diferente de la de los otros cebadores y/o adaptadores en dicho conjunto, los cebadores y/o adaptadores comprenden adicionalmente una secuencia cebadora que es la misma en todos los cebadores y/o adaptadores en dicho conjunto, las secuencias etiqueta están situadas en posición 5' con respecto a la secuencia cebadora y los diferentes cebadores y/o adaptadores que comprenden cada secuencia etiqueta diferente se disponen por separado los unos de los otros. En realizaciones las etiquetas tienen cualquier número de bases de longitud. En realizaciones las etiquetas tienen 2, 3, 4, 5, 6, 8, 10, 12 bases de longitud. En realizaciones las etiquetas tienen 4 bases. En realizaciones la secuencia cebadora es específica para cualquier polinucleótido diana de interés. En realizaciones la secuencia cebadora es específica para una secuencia en el ARNr 16S. En realizaciones las etiquetas difieren entre sí en al menos 2 bases. En las realizaciones de las etiquetas no contienen extensiones de polinucleótidos dentro de la etiqueta. En realizaciones las etiquetas no contienen extensiones de homopolinucleótidos dentro de o en el empalme de la etiqueta y el cebador de PCR. En realizaciones las etiquetas no contienen extensiones de polinucleótidos dentro de o en el empalme de la etiqueta y el adaptador de PCR en emulsión. En realizaciones, las etiquetas no son complementos inversos entre sí.

Breve descripción de las Figuras

La Figura 1 es un diagrama esquemático que muestra una realización general de la invención. En la parte superior de la Figura se muestra una pluralidad de muestras (S_1, S_2, a S_j). Cada muestra está compuesta de una pluralidad de polinucleótidos ($P_{1\text{-}1}$ a $P_{1\text{-}N1}$ en S_1; $P_{2\text{-}1}$ a $P_{2\text{-}N2}$ en S2; a $P_{j\text{-}1}$ a $P_{j\text{-}nj}$). Los polinucleótidos de cada muestra se etiquetan por separado con una secuencia de polinucleótidos etiqueta, estando etiquetados todos los polinucleótidos en una muestra dada (en esta ilustración) con una sola secuencia etiqueta, denominada en la figura T_1 para S_1, T_2 para S_2, T_j para S_j. Los polinucleótidos etiquetados individuales se denotan en consecuencia. Los polinucleótidos etiquetados en cada muestra se denominan colectivamente, para cada muestra, T_1S_1, T_2S_2, a T_jS_j. Los polinucleótidos etiquetados de las muestras se mezclan entre sí para formar una mezcla, denominada M_i. La mezcla se secuencia, típicamente mediante medio de un método de secuenciación masiva en paralelo. Las secuencias etiqueta se identifican en los datos así obtenidos. Las secuencias se agrupan por etiquetas. Las secuencias de las muestras individuales se identifican de este modo.

La Figura 2A es un diagrama que representa la etapa I en la secuenciación multietiqueta de muestras de la comunidad microbiana utilizando un par de cebador-conector directo e inverso 16S para la amplificación mediante PCR. (a) representa el cebador de ARNr 16S Directo con la Etiqueta *I* y el Conector de PCR en Emulsión, (b) representa la secuencia de ARNr 16S, (c) representa el cebador de ARNr 16S Inverso con la Etiqueta *j* y el Conector de PCR en Emulsión, (d) representa la Secuencia de ARNr 16S amplificada con las Etiquetas *ij* Directa e Inversa, (e) representa las cuentas de PCR en emulsión, (f) representa la lectura de pirosecuenciación, (g) representa el pocillo en la placa de picotitulación, (h) representa una etiqueta Única, (i) representa la Comunidad Amplificada 1, (j) representa la Comunidad Amplificada 2, y (k) representa la Comunidad Amplificada n. La etapa 1 implica la amplificación de la comunidad microbiana de cada muestra utilizando cebadores-conectores universales etiquetados de forma única. En la etapa 1, se amplifican diferentes muestras por separado, utilizando adaptador-etiquetas-cebadores específicos de ARNr 16S con una etiqueta diferente para cada muestra.

La Figura 2B es un diagrama que representa las cuentas para la reacción de PCR en emulsión dispuestas al azar en matrices en la placa picolitros. En la etapa 2 en el procedimiento, los productos de PCR de todas las

muestras se mezclan, se inmovilizan sobre cuentas, se distribuyen en los pocillos de la placa de picotitulación, y se amplifican mediante PCR en emulsión.

La Figura 2C es un diagrama que representa el procedimiento de pirosecuenciación de cada adaptador externo en cada pocillo de la placa de picotitulación. Cada reacción lee la secuencia del adaptador, a través de las etiquetas únicas y la secuencia asociada de la muestra etiquetada.

La Figura 2D es un diagrama que representa la clasificación algorítmica de las lecturas de pirosecuenciación utilizando la secuencia etiqueta individual y una porción de la secuencia del cebador. (l) representa las lecturas de secuencia de la muestra 1, (m) representa las lecturas de secuencia de la muestra 2, y (n) representa las lecturas de secuencia de la muestra n.

La Figura 2E es un diagrama que representa la identificación de los taxones microbianos comparando las lecturas de secuencia para cada muestra frente a la base de datos de secuencias de ARNr 16S y a continuación normalizando la abundancia de cada taxón con respecto a las lecturas totales en esa muestra concreta. (o) representa el histograma normalizado de especies derivado de las lecturas de pirosecuenciación obtenidas a partir de la muestra 1, (p) representa el histograma normalizado de especies derivado de las lecturas de pirosecuenciación obtenidas a partir de la muestra 2, (q) representa el histograma normalizado de especies derivado de las lecturas de pirosecuenciación obtenidas a partir de la muestra n.

La Figura 3 es la distribución de especies en muestras de los Controles (A), Crohn (B), y Colitis ulcerosa (C) determinada mediante el procedimiento de pirosecuenciación de 454 Life Sciences. Cada barra del histograma representa la abundancia media normalizada de los taxones en cada estado de enfermedad. Cada muestra se realizó en un pocillo separado sobre la placa picolitros utilizando la máscara de 454 para 16 pocillos.

La Figura 4 es un ejemplo de la distorsión de los componentes de una mezcla compleja causada por ligación de los adaptadores de PCR en Emulsión en amplicones de PCR. La Figura 4A muestra la distribución del tamaño de los amplicones de PCR en la muestra 309 antes de la ligación y la Figura 4B muestra la distribución del tamaño de la muestra de 309 después de la ligación.

La Figura 5 es un ejemplo de la abundancia de taxones normalizados en muestras duplicadas determinada mediante Pirosecuenciación multietiqueta después de la ligación directa de los adaptadores de PCR en emulsión.

La Figura 6 muestra todas las posibles etiquetas de polinucleótidos hexaméricas dentro de las cuales no hay repeticiones de dinucleótidos y ninguna etiqueta es el complemento inverso de ninguna otra etiqueta.

La Figura 7 muestra 96 cebadores adaptadores etiquetados en los que no hay repeticiones de dinucleótidos en las etiquetas, ningún dinucleótido se repite en el empalme de las etiquetas y las etiquetas no son complementos inversos de ninguna otra. En cada caso también se pueden utilizar 5 bases del cebador para identificar las muestras. 7A y 7B muestran los cebadores directos. 7C y 7D muestran los cebadores inversos.

Glosario

Los significados atribuidos a los diversos términos y frases utilizados en la presente memoria se explican de forma ilustrativa a continuación.

"Un", o "uno" significa uno o más; al menos uno.

"Aproximadamente", según se utiliza en la presente memoria, significa groseramente, más o menos. Si se requiere una definición numérica precisa, "aproximadamente" significa +/- 25%.

"Adaptador" significa una secuencia de polinucleótidos utilizada para anclar fragmentos de polinucleótidos individuales a las cuentas y/o para cebar la reacción de PCR en emulsión y/o como molde para cebar reacciones de pirosecuenciación.

"ALH" se utiliza en la presente memoria para significar la heterogeneidad de la longitud del amplicón.

"Amplicón" se utiliza en la presente memoria para referirse a los productos de una reacción de amplificación.

"Amplificado clonalmente" se utiliza en la presente memoria para significar generalmente la amplificación de una sola molécula de partida. Típicamente también se refiere al agrupamiento conjunto de los productos de amplificación, aislados de otros moldes o productos de amplificación.

"ADNds" significa ADN de bicatenario.

Disbiosis significa un cambio en una de las especies y la abundancia de especies de una comunidad microbiana.

"Flanqueantes" se utiliza generalmente para significar en cada lado, tal como en el lado 5' y 3' de una región de un polinucleótido - con referencia a los extremos 5' y 3' de una u otra cadena un polinucleótido bicatenario. Los cebadores directo e inverso para amplificar una región de un polinucleótido mediante PCR, por ejemplo, flanquean la región que va a amplificarse.

"Muestra de la comunidad microbiana" se utiliza en la presente memoria para referirse a una muestra, generalmente de naturaleza biológica, que contiene dos o más microbios diferentes. Las muestras de la comunidad microbiana

incluyen, por ejemplo, muestras ambientales, así como muestras biológicas, tales como muestras para análisis clínicos. El término se aplica también a preparaciones, tales como preparaciones de ADN, derivadas de tales muestras.

"Secuenciación múltiplex" en la presente memoria se refiere a la secuenciación de dos o más tipos o muestras de polinucleótidos en una sola reacción o en un solo recipiente de reacción.

"PCO" significa análisis de coordenadas principales.

"PCA" significa análisis de componentes principales.

"Placa de picotitulación" significa una placa que tiene un gran número de pocillos que tienen un volumen relativamente pequeño, típicamente más pocillos que una placa de microtitulación de 96 pocillos, y volúmenes más pequeños que los de un placa de microtitulación típica de 96 pocillos también.

"Cebador" significa una secuencia de polinucleótidos que se utiliza para amplificar productos de PCR y/o para cebar reacciones de secuenciación.

"ADNss" significa ADN de monocatenario.

"Etiqueta", "Secuencia etiqueta" etc. significa típicamente una secuencia heteróloga, tal como una secuencia de polinucleótidos que identifica otra secuencia con la que está asociada por ser de un tipo dado o pertenecer a un grupo dado.

"Región genética variable" según se utiliza en la presente memoria significa una región genética que varía, por ejemplo entre individuos de una misma especie y entre especies. La frase no denota una longitud específica, sino, más bien se utiliza para denotar una región que comprende una variación de la longitud exacta de la cual puede variar y puede diferir en diferentes contextos. En cuanto a un polinucleótido bicatenario, el término incluye una o la otra y ambas cadenas de la región, y puede ser utilizado para referirse a una, la otra, o a ambas cadenas, y por lo general quedará claro por el contexto que se quiera significar. Un ejemplo específico de una región genética que varía entre los individuos, proporcionado solo con fines ilustrativos, es una región genética que contiene un sitio de SNP (polimorfismo de un solo nucleótido). Por región genética variable a este respecto se entiende una región que contiene el sitio del SNP. Diferentes secuencias de SNP a este respecto constituyen las variantes de la región genética variable. Un ejemplo específico de una región genética variable que difiere entre especies consiste en los genes para el ARN 16S que varían característicamente entre microbios y se puede utilizar para identificar microbios en muestras de la comunidad mixtas como se describe con mayor detalle en algunos de los ejemplos de la presente memoria.

Descripción de la Invención

En ciertos aspectos y realizaciones la invención se refiere a análisis de secuenciación múltiplex utilizando etiquetas. En diversos aspectos y realizaciones de la invención a este respecto la invención proporciona métodos para secuenciar dos o más muestras simultáneamente en una mezcla entre ellas, en donde cada muestra se conecta primero a una etiqueta de secuencia específica de la muestra, las muestras etiquetadas se mezclan y se secuencian, y las secuencias de cada muestra se identifican a continuación por sus respectivas etiquetas de secuencias específicas de la muestra.

La Figura 1 proporciona una representación general de varios aspectos y realizaciones de la invención a este respecto, y la figura se comenta a modo de ilustración a continuación con referencia a la secuenciación de ADN de diferentes muestras. En la parte superior de la Figura se muestra una pluralidad de muestras (S_1, S_2, a S_j). Cada muestra se compone de una pluralidad de polinucleótidos (P_{1-1} a P_{1-N1} en S_1; P_{2-1} a P_{2-N2} en S2; a P_{j-1} a P_{j-nj}). Los polinucleótidos de cada muestra se etiquetan por separado con una secuencia de polinucleótidos etiqueta, etiquetándose todos los polinucleótidos en una muestra dada (en esta ilustración) con una secuencia etiqueta única, designada en la figura como T_1 para S_1, T_2 para S_2, a T_j para S_j. Los polinucleótidos etiquetados individuales se denotan en consecuencia. Los polinucleótidos etiquetados en cada muestra se denominan colectivamente, para cada muestra, T_1S_1, T_2S_2 a T_jS_j. Los polinucleótidos etiquetados de las muestras se mezclan entre sí para formar una mezcla, denominada M_i. La mezcla se secuencia típicamente por medio de un método de secuenciación en paralelo. Las secuencias etiqueta se identifican en los datos obtenidos de este modo. Las secuencias se agrupan mediante etiquetas. Las secuencias de las muestras individuales se identifican de este modo.

En realizaciones las etiquetas son secuencias de 3 a 30, de 4 a 25, de 4 a 20 base de longitud. En realizaciones las etiquetas son de 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36 nucleótidos de longitud o cualquiera de sus combinaciones.

En realizaciones hay 1 - 6, 6 - 12, 10 - 15, 10 - 20, 15 - 25, 20 - 40, 25 - 50, 25 - 75, 50 - 100, 50 - 150, 100 - 200, 100 - 250, 50 - 250, 100 - 500, 500 -1000, 100 - 1.000, 500 - 5.000, 100 - 10.000, 1.000 - 25.000, 500 - 50.000, 100 - 100.000, 1 - 1.000.000 o más muestras, etiquetadas, respectivamente, con 1 - 6, 6 - 12, 10 - 15, 10 - 20, 15 - 25, 20 -

40, 25 - 50, 25 - 75, 50 - 100, 50 - 150, 100 - 200, 100 - 250, 50 - 250, 100 - 500, 500 - 1.000, 100 - 1.000, 500 - 5.000, 100 - 10.000, 1.000 - 25.000, 500 - 50.000, 100 - 100.000, 1-1.000.000 o más etiquetas diferentes.

En realizaciones las secuencias se determinan sin el uso de electroforesis en gel. En realizaciones las secuencias se determinan sin el uso de la transferencia de secuencias de un gel sobre una membrana o un filtro para la hibridación. En realizaciones, las secuencias se determinaron por un método de secuenciación en paralelo. En las realizaciones las secuencias se determinan por pirosecuenciación, secuenciación mediante síntesis, secuenciación por hibridación, secuenciación sustractiva, secuenciación por poros o la secuenciación de lectura directa.

En realizaciones de las etiquetas se incorporan a los polinucleótidos en muestras para la secuenciación mediante una etapa de ligación y/o una etapa de amplificación.

En realizaciones las etiquetas están incluidas en cebadores para la amplificación.

En realizaciones las etiquetas están incluidas en cebadores para la amplificación mediante PCR, amplificación mediante transcripción, amplificación por círculo rodador, o amplificación por replicasa Qß.

En realizaciones las etiquetas están incluidas en adaptadores de PCR en emulsión y cebadores para la amplificación.

En realizaciones de las etiquetas se incorporan mediante una etapa de clonación en un vector.

En realizaciones las muestras son muestras de la comunidad microbiana. En realizaciones las muestras son muestras clínicas. En realizaciones las muestras son muestras ambientales. En realizaciones las muestras son muestras para el análisis de SNP. En realizaciones las muestras son muestras para determinación del genotipo. En realizaciones las secuencias se determinan en una o más placas de picotitulación.

En realizaciones las muestras son ADN genómicos fragmentados. En realizaciones las muestras son ADN genómico bacteriano fragmentado, ADN genómico de Archaea, ADN genómico fúngico, ADN genómico eucariótico, el ADN de cloroplasto, y/o ADN mitocondrial. En realizaciones las muestras son ADNc. En realizaciones las muestras son de ADNc eucariótico, ADNc bacteriano, ADNc de Archaea, y/o de ADNc fúngico. En realizaciones las etiquetas se incorporan mediante una etapa ligación y/o de una etapa de amplificación.

En realizaciones las muestras son para uno cualquiera o más de detectar, verificar, perfilar, pronosticar, y/o diagnosticar un trastorno, enfermedad, o similares.

En realizaciones las muestras son para analizar la composición, diversidad, estabilidad, dinámica, y/o cambios en muestras agrícolas, alimentarias, de bioseguridad, veterinarias, clínicas, ecológicas, zoológicas, oceanográficas, y/o cualquier otra muestra que comprende uno o más polinucleótidos.

En realizaciones las secuencias se determinan en pocillos de una placa de titulación. En realizaciones las secuencias se determinan en una o más placas de picotitulación que tienen una máscara. En realizaciones las secuencias se determinan en una o más placas de picotitulación que tienen una máscara, en donde la máscara define 2, 4, 8, 16, 32, 64 o más compartimentos.

A modo de ilustración de una placa de picotitulación 454, en realizaciones existen aproximadamente 120.000 moldes/placa y la longitud de lectura promedia aproximadamente 250 bases por molde. En realizaciones relacionadas con las mismas existen 10 etiquetas de 4 bases por 1/16 de placa, 160 etiquetas en total, un promedio de aproximadamente 750 moldes por etiqueta (y por muestra), y alrededor de 187.500 bases secuenciadas por etiqueta (y por muestra).

En realizaciones existen aproximadamente 260.000 molde/placa y la longitud de lectura promedia aproximadamente 250 bases por molde. En realizaciones relacionadas con las mismas, existen 12 etiquetas de 4 bases por 1/8 de placa, 96 muestras en total, un promedio de alrededor de 2.708 moldes por etiqueta (y por muestra) y aproximadamente 677.083 bases de secuencia por etiqueta (y por muestra).

En realizaciones existe aproximadamente 400.000 moldes/placa y la longitud de lectura promedia aproximadamente 250 bases por molde. En realizaciones relacionadas con las mismas, existen 96 etiquetas de 6 bases para 96 muestras por placa, aproximadamente 4.166 moldes por etiqueta (y por muestra) y aproximadamente 1.041.666 bases de secuencia por etiqueta (y por muestra).

En realizaciones las etiquetas tienen secuencias de 10 bases de longitud, existen 192 etiquetas diferentes, y las muestras se analizan en un formato de placa de microtitulación.

En realizaciones la invención proporciona algoritmos para la desconvolución, a partir de una mezcla de secuencias de dos o más muestras, de las secuencias de las muestras de la mezcla identificando las etiquetas específicas de la

muestra en las secuencias, agrupando las secuencias mediante las etiquetas así identificadas, agrupando de este modo la secuencia de cada una de dichas muestras, separadas entre sí.

En realizaciones la invención proporciona algoritmos para la desconvolución, a partir de una mezcla de secuencias de dos o más muestras, de las secuencias de las muestras de la mezcla identificando las etiquetas específicas de la muestra en secuencias, de la siguiente manera:

1. Leyendo todas las lecturas de secuencias en una matriz;
2. Buscando en el comienzo de cada lectura de secuencia e identificando la etiqueta;
3. Construyendo una matriz asociativa que conecta la etiqueta con la lectura de secuencia;
4. Clasificando las claves de la matriz asociada;
5. Asociando cada clave con la muestra correspondiente;
6. Reuniendo todas las lecturas de secuencia para cada muestra;
7. Analizando cada muestra por separado.
8. Normalizando la abundancia de cada componente dentro de cada muestra con respecto al total de lecturas dentro de esa muestra.

En realizaciones, el algoritmo puede ser implementado en cualquier lenguaje de programación. En realizaciones del algoritmo es implementado en C, C++, Java, Fortran o Basic. En realizaciones del algoritmo se implementa como una secuencia de comandos "script" en Perl.

En realizaciones, la descripción proporciona kits para la secuenciación múltiplex como se describe en la presente memoria, que comprende un conjunto de cebadores y/o adaptadores, en donde cada cebador y/o adaptador en dicho conjunto comprende una secuencia etiqueta, una secuencia cebadora y/o un adaptador de PCR en emulsión. En realizaciones los cebadores y/o adaptadores comprenden adicionalmente un radical para la inmovilización. En realizaciones los cebadores y/o adaptadores comprenden biotina. En realizaciones los cebadores y/o adaptadores en el conjunto comprenden todas las secuencias etiqueta definidas por 2, 3, 4, 5, 6, 7, u 8 secuencias de polinucleótidos base, en donde dichos cebadores y/o adaptadores que comprenden diferentes secuencias etiqueta están dispuestos en recipientes separados entre sí. En realizaciones existen 1-5, 3-10, 5-15, 10-25, 20-50, 25-75, 50-100, 50-150, 100-200, 150-500, 250-750, 100-1000, o más secuencias etiqueta diferentes dispuestas por separado entre sí, con el fin de ser útiles para etiquetar de forma única dicho número de diferentes muestras. En realizaciones los cebadores y/o adaptadores son adecuados para su uso como adaptadores y/o cebadores de amplificación de 454 Life Sciences. En realizaciones los cebadores y/o adaptadores comprenden adicionalmente una cualquiera o más de una secuencia cebadora para uno cualquiera o más de una secuencia de ARNr 16S, una secuencia de ARNr 18S, una secuencia ITS, una secuencia mitocondrial, una secuencia de microsatélites, una secuencia de enzima metabólica, una secuencia de enfermedad genética, y/o cualquier otra secuencia para amplificación o análisis.

Ejemplos

La presente invención se describe adicionalmente por medio de los siguientes ejemplos ilustrativos, no limitantes.

EJEMPLO 1: Secuenciación utilizando el Sistema de pirosecuenciación 454

454 Life Sciences, una filial de Roche Diagnostics, proporciona un dispositivo para la pirosecuenciación de aproximadamente 100.000.000 de bases de aproximadamente 400.000 moldes diferentes en una sola ronda en una sola placa de picotitulación. La compañía también proporciona máscaras que permiten el procesamiento por 2, 4, 8 ó 16 muestras diferentes en una placa. En el máximo de capacidad de utilización de la placa enmascarada, el sistema proporciona alrededor de 1 millón de bases de datos de secuencias en aproximadamente 4.000 moldes para cada una de las 16 muestras.

El proceso general de secuenciación utilizando el sistema 454 es generalmente de la siguiente manera: aislar el ADN; opcionalmente fragmentar el ADN; opcionalmente convertir el ADN en bicatenario; ligar el ADN a los adaptadores; separar las cadenas del ADNds, unir el ADNss a cuentas en condiciones que den como resultado una preponderancia de las cuentas que no tengan una molécula de ADN unida a las mismas o una sola molécula de ADN unida las mismas; capturar las cuentas en gotitas individuales de una emulsión de una mezcla de reacción de PCR en aceite; llevar a cabo una reacción de PCR sobre los ADNs encapsulados en cuentas en emulsión (por medio de lo cual los productos de amplificación se capturan en las cuentas); distribuir los productos de amplificación en los pocillos de picotitulación de manera que o bien no haya ninguna cuenta o bien haya una cuenta en un pocillo; y llevar a cabo la pirosecuenciación en todas las cuentas en todos los pocillos en paralelo.

EJEMPLO 2: Pirosecuenciación múltiplex utilizando 96 adaptadores-cebadores de PCR etiquetados.

454 Life Sciences, una filial de Roche Diagnostics, proporciona un dispositivo para la pirosecuenciación de aproximadamente 100.000.000 millones de bases de secuencia de aproximadamente 400.000 moldes diferentes en una sola ronda en un sola placa de picotitulación. En el máximo de capacidad de utilización de la placa, el sistema proporciona alrededor de 10 millones de bases de datos de secuencia para cada uno de alrededor de 4.000 moldes para cada una de 96 muestras multietiquetadas. En este ejemplo, las 96 etiquetas tienen 6 bases de longitud y se utilizan junto con 6 bases del cebador directo o inverso para identificar las lecturas que pertenecen a cada una de las

96 muestras individuales (véase la Figura 2).

EJEMPLO 3: Análisis de pirosecuenciación multietiqueta de muestras de la comunidad microbiana

Diversos aspectos y realizaciones de la invención descrita en la presente memoria se ilustran por medio del siguiente ejemplo general con respecto al análisis "ecogenómico" de la diversidad microbiana en las muestras biológicas.

La capacidad de cuantificar el número y tipos de microorganismos dentro de una comunidad es fundamental para la comprensión de la estructura y función de un ecosistema, como se comentan, por ejemplo, Pace 1997 y Theron y Cloete 2000. Tradicionalmente, el análisis de las comunidades microbianas se ha realizado utilizando técnicas microbiológicas, pero estas técnicas son limitadas. Por ejemplo no son útiles para los muchos organismos que no pueden ser cultivados (Ritchie, Schutter et al 2000; Spring, Schulze et al. 2000). Incluso para aquellos organismos que se pueden cultivar, estas técnicas proporcionan poca información con la que identificar los microbios individuales o caracterizar sus rasgos fisiológicos (Morris, Bardin et al. 2002).

Los recientes avances en las técnicas moleculares han superado algunos de estos inconvenientes, y han permitido la identificación de muchos más taxones en las comunidades microbianas que las técnicas tradicionales microbianas. Estos avances han proporcionado una considerable penetración en la expresión de funciones clave en las especies en las comunidades microbianas. (Pace 1997; Suzuki 1998; Amann 2000; Frischer, Danforth et al. 2000; Ritchie, Schutter et al. 2000; Spring, Schulze et al. 2000). Entre estas técnicas moleculares se encuentran la Electroforesis en Gel de Gradiente Desnaturalizante por "Denaturing Gradient Gel Electrophoresis" (DGGE), la Electroforesis en Gel de Gradiente de Temperatura por "Temperature Gradient Gel Electrophoresis" (TGGE), la Electroforesis en Gel de Gradiente de Temperatura Temporal por "Temporal Temperature Gradient Gel Electrophoresis" (TTGE), el Polimorfismo de Longitud de Fragmentos de Restricción Terminal por "Terminal-Restriction Fragment Length Polymorphism" (T-RFLP), Polimirfismos de Conformación de Cadena Sencilla por "Single Strand Conformation Polymorphism" (SSCP), y PCR de Heterogeneidad de Longitud por "Length Heterogeneity PCR" (LH-PCR) (Frischer, Danforth et al. 2000; Theron y Cloete 2000; Mills, Fitzgerald et al. 2003; Seviour, Mino et al. 2003; Klaper y Thomas 2004).

Entre éstas, la LH-PCR es probablemente la mejor técnica de huella dactilar. Es poco costosa, rápida, y se puede utilizar de forma rutinaria para detectar varios cientos de muestras en un día. Es útil como una herramienta de estudio rutinaria que se puede utilizar para controlar la dinámica de las comunidades microbianas naturales del suelo, y para identificar rápidamente muestras de interés mediante análisis PCO. La LH-PCR se ha utilizado ampliamente para evaluar la variación natural en las comunidades bacterianas mediante perfilado de las regiones variables amplificadas de genes de ARNr 16S en muestras de poblaciones microbianas mixtas, utilizando PAGE (Véanse Mills 2000; Litchfield y Gillevet 2002; Lydell, Dowell et al. 2004). Los productos de la LH-PCR de las especies individuales en la población dan lugar a distintas bandas en los geles. El "área del pico" de cada banda es proporcional a la abundancia de la especie en la comunidad. La LH-PCR de regiones variables de ARNr 16S se ha utilizado con bastante éxito para estimar la diversidad de especies en comunidades de bacterioplancton, en particular. (Véase Suzuki, Rappe et al 1998; Ritchie, Schutter et al 2000).

La funcionalidad de la comunidad no se puede determinar directamente a partir de datos del clon de ARNr 16S, sin embargo, debe deducirse de los datos mediante análisis filogenético. Además, la LH-PCR y otras tecnologías de huellas dactilares, a la vez que las herramientas de gran alcance para controlar la dinámica de poblaciones, no pueden identificar las especies individuales en una comunidad. Para ello, se deben seguir investigaciones de huellas dactilares por la construcción, la clonación, la secuenciación y el análisis filogenético de bibliotecas (Fitzgerald 1999; McCraig 1999; Spring, Schulze et al. 2000; Theron y Cloete 2000; Litchfield y Gillevet 2002; Bowman y McCuaig 2003; Kang y Mills 2004; Eckburg, Bik et al. 2005). La identificación de especies de un estudio de huellas dactilares, por lo tanto, es una empresa considerable que es inconveniente, consume mucho tiempo, es costosa y está sujeta a limitaciones técnicas.

Las muestras de agrupamiento pueden, en cierta medida, reducir el coste, el tiempo y los gastos de tales análisis. Por ejemplo, el análisis PCO de los datos de LH-PCR se puede utilizar para agrupar muestras con perfiles similares para la clonación y la secuenciación por lotes. La combinación de las muestras de esta manera se reduce el tiempo, los gastos y el trabajo implicados en el análisis de las muestras. Se requiere la secuenciación de al menos 300 clones al azar para identificar los componentes bacterianos de la muestra agrupada hasta 1% de las poblaciones de bacterias totales en muestras típicas. Este nivel de resolución es similar al de la huella dactilar ALH. Originalmente un nuevo enfoque, agrupando muestras similares antes de la clonación y secuenciación ha demostrado ser seguro y eficaz.

En los estudios clásicos de la comunidad en la literatura (Eckburg, Bik et al. 2005), las muestras ambientales se analizan de forma independiente. A continuación, los datos de secuencia del clon de clases/grupos específicos se analizan estadísticamente utilizando usualmente algún tipo de promediación métrica. Los análisis de este tipo pueden ser extremadamente costosos, especialmente si se analizan de forma exhaustiva bibliotecas de clones, algo que implica típicamente la secuenciación de miles de clones. Por otra parte, para que el procedimiento de "promediación" sea válido, como se requiere para la comparación de las poblaciones mixtas, las muestras deben ser

agrupadas en proporciones iguales. Aunque simple en principio, en realidad, es difícil de lograr e, incluso si se lograra, sería imposible de verificar. Una nueva técnica, basada en la pirosecuenciación, ofrece ventajas que superan una variedad de estos inconvenientes de las tecnologías de la huella dactilar mencionadas anteriormente. El método se implementa en un instrumento comercializado por 454 Life Sciences, Inc., una subsidiaria de Curagen Sciences, Inc., utilizando reactivos proporcionados por la misma empresa. Además, 454 Life Sciences ofrece un servicio a medida para la pirosecuenciación.

En esta tecnología, las moléculas de ADN individuales se amplifican sobre cuentas mediante PCR en gotitas individuales en una emulsión de aceite-en-agua. La cuentas se depositan a continuación individualmente en pocillos de una placa de picotitulación. Las secuencias de todos los ADN en los pocillos se determinan en paralelo mediante pirosecuenciación. (Véanse Venter, Levy et al. 2003; Margulies, Egholm et al. 2005; Poinar, Schwarz et al. 2006). En una ronda típica, hay aproximadamente 200.000 moldes por placa, una longitud de lectura media de aproximadamente 100 bases a partir de cada molde, y una ronda con una sola placa genera unos 20 millones de bases de la secuencia en una sola ronda de cuatro horas.

Aunque la tecnología aumenta en gran medida el rendimiento sobre los métodos anteriores, es costoso. En particular, el coste por placa es demasiado alto para que sea económicamente práctico llevar a cabo muchos análisis. Para reducir costes, se pueden utilizar máscaras que dividen una placa en 16 zonas de muestra independientes, de manera que se puede utilizar una placa para procesar 16 muestras diferentes, ya sea al mismo tiempo o de forma independiente. Cada zona 1/16 proporciona aproximadamente 1.000.000 bases de datos de secuencias de aproximadamente 10.000 moldes diferentes. Aunque esto reduce el coste por muestra, los gastos asociados con el uso de esta tecnología todavía son indeseablemente altos.

Diversos aspectos y realizaciones de la presente invención se pueden utilizar para reducir aún más el coste por muestra de esta tecnología (así como otras técnicas, tal como se describe en otra parte en la presente memoria). El uso de técnicas de multietiquetado (referido, entre otros como "Proceso Multietiqueta") para el análisis genómico de las poblaciones bacterianas de acuerdo con ciertos aspectos y realizaciones de la invención, notablemente la secuenciación de alta cobertura de las comunidades bacterianas, se denomina en la presente memoria "Ecogenómica Multietiqueta" y también "Análisis Ecogenómico Multietiqueta".

(Algunas publicaciones utilizan el término "Pirosecuenciación múltiplex" (Pourmand, Elahi et al. 2002) para referirse a la generación de una señal compuesta de múltiples dianas que se leen como una firma para una muestra específica. El término no se utiliza para referirse a la multiplexación basada en etiquetas en la que se determinan las secuencias de diferentes muestras en una mezcla y después se someten a desconvolución partir de los datos de secuenciación mixtos utilizando etiquetas específicas de la muestra, incorporadas durante las reacciones de amplificación.)

Como se describe a continuación el Proceso Multietiqueta en una serie de etapas relativamente sencillas logra todo lo que de otro modo requeriría no solo el análisis de las huellas dactilares de la comunidad, sino también todos los procesos de clonación y secuenciación previamente requeridos para el Análisis Ecogénomico de alta cobertura utilizando técnicas convencionales.

A modo de ilustración, el siguiente ejemplo describe el uso de Análisis Ecogenómico Multietiqueta de regiones variables de genes comunes utilizando cebadores universales etiquetados para el análisis de alta cobertura de varias muestras de la comunidad microbiana, todo al mismo tiempo. El análisis se lleva a cabo como se ha descrito en general anteriormente, y se elabora adicionalmente en detalle a continuación.

En pocas palabras, se añaden etiquetas cortas a los extremos 5' de los cebadores de PCR directo e inverso normalmente utilizados para el análisis de la comunidad. Estas etiquetas se pueden colocar entre los adaptadores de PCR en Emulsión y los cebadores de PCT (véase la Figura 2). Una etiqueta diferente se ancla a los cebadores para cada una de las muestras que se van a combinar. Por ejemplo se pueden utilizar cebadores que abarcan una región variable de genes de ARNr 16S para el análisis de las comunidades bacterianas y arqueales. Los cebadores específicos de ARNr 16S con etiquetas de 4 bases se establecen en la Tabla 1 a continuación. Asimismo se pueden utilizar cebadores que abarcan una región variable de un gen ITS para el análisis de las comunidades fúngicas. Se apreciará que la elección de estos cebadores específicos no es exclusiva, y que se puede emplear una amplia variedad de otros cebadores adecuados para otras regiones diana para la amplificación de la misma manera que se ha descrito en la presente memoria para los genes 16S e ITS. Por lo tanto, se puede utilizar cualquier gen de interés que proporcione sitios de cebadores conservados a través de una comunidad, y una variación suficiente en la región entre los cebadores para la resolución deseada de las especies individuales. Así, por ejemplo, los genes específicos de las rutas funcionales tales como la oxidación de metano anaeróbica, o la reducción de azufre pueden servir como dianas para la reacción de amplificación, así como las secuencias de ARNr 16S.

Tabla 1

Nombre	Etiqueta	Secuencia Directa Compartida
		AGCTAGAGTTTGATCMTGGCTCAG
L27FA	AGCT	AGCTAGAGTTTGATCMTGGCTCAG
L27FB	AGTC	AGTCAGAGTTTGATCMTGGCTCAG
L27FC	GATC	GATCAGAGTTTGATCMTGGCTCAG
L27FD	GACT	GACTAGAGTTTGATCMTGGCTCAG
L27FE	CTGC	CTGCAGAGTTTGATCMTGGCTCAG
L27FF	CTAG	CTAGAGAGTTTGATCMTGGCTCAG
L27FG	ATGC	ATGCAGAGTTTGATCMTGGCTCAG
L27FH	ATAG	ATAGAGAGTTTGATCMTGGCTCAG
L27FM	ATCT	ATCTAGAGTTTGATCMTGGCTCAG
L27FO	ATAT	ATATAGAGTTTGATCMTGGCTCAG
Nombre	**Etiqueta**	**Secuencia Inversa Compartida**
		AGCTGCTGCCTCCCGTAGGAGT
355RA	AGCT	AGCTGCTGCCTCCCGTAGGAGT
355RB	AGTC	AGTCGCTGCCTCCCGTAGGAGT
355RC	GATC	GATCGCTGCCTCCCGTAGGAGT
355RD	GACT	GACTGCTGCCTCCCGTAGGAGT
355RE	CTGC	CTGCGCTGCCTCCCGTAGGAGT
355RF	CTAT	CTATGCTGCCTCCCGTAGGAGT
355RG	ATGC	ATGCGCTGCCTCCCGTAGGAGT
355RH	ATAT	ATATGCTGCCTCCCGTAGGAGT
355RM	ATCT	ATCTGCTGCCTCCCGTAGGAGT
355RO	ATAC	ATACGCTGCCTCCCGTAGGAGT

La Tabla 1 muestra un cebador específico de ARNr 16S con una variedad de secuencias etiqueta de 4 bases ancladas. Como se describe en la presente memoria tales cebadores son útiles para amplificar ARNr 16S en varias muestras que pueden ser secuenciados juntos a continuación. El ARNr 16S en cada muestra se amplifica utilizando una etiqueta diferente, pero la misma secuencia de cebador 16S. Las secuencias de ARNr amplificadas a partir de las muestras se combinan y se secuencian juntas. Las secuencias de ARNr de las diferentes muestras se identifican y clasifican a continuación por su secuencia etiqueta de 4 bases más las primeras 4 bases de cada cebador. Se debe apreciar que las secuencias aguas abajo de la secuencia compartida del cebador 16S diferirán entre las muestras, así como la secuencia etiqueta.

En cada caso, las muestras se amplifican individualmente. Los amplicones resultantes comprenden las secuencias de los cebadores incluyendo las etiquetas. Puesto que se utilizan etiquetas únicas para cada muestra, las etiquetas en los amplicones de cada muestra serán diferentes. Los ADN amplificados se reúnen y se secuencian mediante

pirosecuenciación como se ha descrito anteriormente. Los datos de secuencia de una ronda se analizan, en parte, agrupando todas las secuencias que tienen la misma etiqueta. De esta manera, las secuencias de cada muestra se desmultiplexan partir de los datos de secuenciación obtenidos a partir de la mezcla.

El funcionamiento de la invención a este respecto se ilustra mediante la siguiente simulación, llevada a cabo utilizando los datos de población convencionalmente obtenidos a partir de muestras de derrames fríos. El algoritmo para el análisis de secuencia utiliza una secuencia de comandos en PERL para extraer las primeras 100 bases de la secuencia. A continuación, analiza la totalidad de las secuencias de 100 bases utilizando una secuencia de comandos de PERL RDP a medida. La secuencia de comandos funciona de la siguiente manera:

1. Leer todas las lecturas de secuencias en una matriz asociada (Encriptación "Hash" 1);
2. Extraer 100 subsecuencias de bases desde el principio de cada lectura de secuencia;
3. Crear una matriz asociativa (Encriptación 2) de las secuencias;
4. Realizar una búsqueda BLAST de la base de datos RDP con la Encriptación 1;
5. Realizar una búsqueda BLAST de la base de datos RDP con la Encriptación 2;
6. Comparar las identificaciones de la secuencia original (Encriptación 1) y la subsecuencia (Encriptación 2);
7. Compilar una lista de las identificaciones similares para la Encriptación 1 y la Encriptación 2;
8. Compilar una lista de las identificaciones diferentes para la Encriptación 1 y la Encriptación 2;
9. Calcular el porcentaje de identificaciones similares.

Como se muestra a continuación, no existe prácticamente ninguna diferencia a nivel de clase en la diversidad microbiana generada por la simulación de la secuencia y la que se deriva directamente de las secuencias de ARNr 16S en la base de datos.

Tabla 2

Clase RDP	Primer segmento de 100 unidades	ARNr 16S
SUBDIVISION_ALFA	3,6%	3,6%
HALÓFILOS_ANAEROBIOS	3,6%	3,6%
SUBDIVISION_BACILLUS-LACTOBACILLUS_STREPTOCOCCUS	3,6%	3,6%
BACTEROIDES_Y_CYTOPHAGA	7,1%	7,1%
SUBDIVISION_CHLOROFLEXUS	3,6%	3,6%
GRUPO_CY.AURANTIACA	7,1%	7,1%
CYANOBACTERIA	7,1%	7,1%
SUBDIVISION DELTA	14,3%	14,3%
SUBGRUPO_CLON_MEDIOAMBIENTAL_WCHB1-41_	7,1%	7,1%
GRUPO_FLX.LITORALIS	3,6%	3,6%
SUBDIVISIÓN_GAMMA	10,7%	10,7%
BACTERIAS CON ALTO CONTENIDO DE _G + C_	7,1%	7,1%
GRUPO_LEPTOSPIRILLUM	3,6%	3,6%
MYCOPLASMA_Y_RELACIONADOS	3,6%	3,6%
GRUPO_PIRELLULA	3,6%	3,6%
GRUPO_SPHINGOBACTERIUM	3,6%	3,6%
SUBDIVISION_SPIROCHAETA-TREPONEMA-BORRELIA	3,6%	3,6%
THERMOANAEROBACTER_Y_RELACIONADOS	3,6%	3,6%

EJEMPLO 3: Análisis de pirosecuenciación multietiqueta de disbiosis en EII

Las enfermedades inflamatorias intestinales (EII o las EII), a saber, la colitis ulcerosa (CU) y la enfermedad de Crohn (EC), son enfermedades crónicas, de por vida, con recaídas, que afectan a aproximadamente 1 millón de estadounidenses y cuestan aproximadamente dos mil millones de dólares al año al sistema de salud de EE.UU. Se desconoce la causa de las EII, no tienen cura, y su incidencia está aumentando. El curso natural de estas enfermedades se caracteriza por períodos de quiescencia (enfermedad inactiva) alternados con brotes (enfermedad activa). Ahora es ampliamente aceptado que los brotes de la EII se deben, sin embargo, a una reacción inflamatoria mal regulada a una disbiosis anormal de la microflora intestinal).

Los cambios específicos en la microflora de los pacientes con EII que podrían causar estas enfermedades siguen siendo desconocidos. La restricción de las búsquedas, para un solo patógeno que causa la EII ha sido infructuosa (Véase Guarner y Malagelada 2003). Los estudios de pequeños grupos de bacterias han producido resultados ambiguos (Véase Schultz y Sartor 2000). Solo recientemente se han intentado estudios de grandes conjuntos de la flora bacteriana (Véase Eckburg, Bik, et al. 2005). La mejora del conocimiento de los autores de la presente invención sobre la microflora del tracto gastrointestinal tiene el potencial de revolucionar el tratamiento de la EII. El desarrollo de métodos en tiempo real para el estudio de los cambios en la microflora puede dar lugar a herramientas de diagnóstico para predecir los brotes, y para dirigir, tratamientos seguros para la EII.

El requisito clave para la comprensión de la disbiosis en las enfermedades polimicrobianas es un método para interrogar ampliamente la microflora en numerosas muestras de control y de enfermedad para identificar las tendencias dinámicas en la composición de especies asociadas a la salud y la progresión de la enfermedad. En estudios comunitarios clásicos (Eckburg, Bik, et al. 2005) se someten a ensayo de forma independiente muestras ambientales y a continuación se analizan estadísticamente los datos de secuencia del clon de clases/grupos específicos por lo general con algún tipo de promediación métrica. Esto puede ser extremadamente costoso, especialmente si se analizan de forma exhaustiva las bibliotecas de clones (es decir, 10.000 clones por muestra).

Para mejorar el rendimiento y reducir el coste, se ha utilizado PCR de Heterogeneidad de Longitud del Amplicón (ALH-PCR) para estudiar la microflora intestinal. Ésta ofrece una forma rápida de escrutinio de comunidades microbianas complejas, permitiendo una fácil toma de huellas dactilares de los cambios en la microflora. La toma de huellas dactilares de LH-PCR es económica y rápida, con la capacidad de detectar varios cientos de muestras de un día. Se puede utilizar como una herramienta de estudio rutinaria para controlar la dinámica de las comunidades microbianas del suelo naturales o para identificar rápidamente muestras de interés mediante el análisis PCO. El análisis PCO se ha utilizado para las muestras del grupo con perfiles similares, lo que les permite ser agrupadas para la clonación y secuenciación. Esto reduce en gran medida el coste del análisis de múltiples muestras, en particular cuando el análisis requiere la secuenciación de al menos 300 clones al azar para identificar componentes bacterianos de la muestra hasta una representación del 1% en la población total (que es el límite de resolución de la huellas dactilar de ALH). El agrupamiento de muestras similares antes de la clonación y secuenciación ha demostrado ser bastante robusto. Sin embargo, debe agruparse la misma cantidad de producto de PCR de cada muestra o los resultados serán sesgados.

La pirosecuenciación multietiqueta es una nueva tecnología de pirosecuenciación que permite que se secuencien juntas muchas muestras de la comunidad con una alta cobertura sin la necesidad de técnicas de huella dactilar, clonación, o purificación y separación requeridas por métodos los convencionales para el análisis de las comunidades microbianas, como se ha descrito anteriormente en la presente memoria. La secuenciación multietiqueta es más eficaz, rápida y menos costosa que otros métodos.

A modo de ilustración, la pirosecuenciación multietiqueta se puede llevar a cabo utilizando un conjunto de etiquetas específicas en el extremo de cebadores de ARNr de la subunidad ribosomal pequeña ("SSU") convencional universal (Véase la Tabla 1). Se utiliza un conjunto diferente de los cebadores etiquetados para amplificar el ARNr SSU en cada muestra ambiental diferente (Figura 2-Etapa 1). Los amplicones de PCR de todas las muestras se agrupan. Se lleva a cabo la PCR en emulsión y los amplicones derivados de cada molécula se capturan en sus respectivas cuentas. Después de la amplificación, las cuentas se distribuyen en los pocillos de una placa de picotitulación (Figura 2-Etapa 2). Las secuencias, incluyendo las secuencias etiquetadas, de los amplicones en cada cuenta se determinan mediante pirosecuenciación (Figura 2-Etapa 3). Se utiliza una secuencia de comandos de PERL u otro programa adecuado para clasificar la información de la secuencia utilizando las etiquetas y la secuencia del cebador como clave. Las secuencias con las mismas etiquetas se identifican de este modo con su respectiva muestra. Las especies de bacterias de cada muestra se identifican a continuación emparejando las secuencias de ARNr SSU con las entradas en la base de datos del Ribosomal Database Project (RDP 8.1 o RDP 9.0). La frecuencia normalizada con la que se identifica de este modo una bacteria en una muestra dada es indicativa de su representación relativa en la comunidad microbiana. Los histogramas basados en estas determinaciones de frecuencia se pueden utilizar para el análisis no paramétrico de los desplazamientos disbióticos implicados en los estados de enfermedad.

Por ejemplo, la Figura 3 representa los resultados de tal experimento en el que se analizaron seis muestras de mucosa de control, diez de Crohn, y ocho de colitis ulcerosa por medio de Pirosecuenciación Multietiqueta. Cada uno de los segmentos en las barras apiladas del histograma representa la abundancia normalizada de los taxones específico en una muestra específica. En este experimento se realizó la identificación de los taxones utilizando análisis BLAST de la base de datos RDP 8.1. Se puede observar que algunos taxones (es decir, el subgrupo de Bacillus fragilis y el subgrupo de Rumanococcus gnavus) están presentes en la misma abundancia en los estados tanto de control como de enfermedad. Otros taxones, tales como Clostridium leptum son más dominantes en la Colitis ulcerosa, mientras que otros (es decir, el subgrupo Gloeothece gloeocapsa) son indicadores de disbiosis en el estado de enfermedad.

Sin embargo, en el procedimiento convencional de 454 Life Science se utiliza una etapa de ligación para conectar

los adaptadores de PCR en emulsión a los amplicones de PCR y produce numerosos artefactos en la cuantificación de la abundancia de cada taxón en las muestras. En los resultados que se muestran en la Figura 3, los autores de la presente invención eliminaron algorítmicamente las quimeras, las lecturas inversas y los productos truncados y filtraron los datos para eliminar todos los taxones representados por menos de 5% de la abundancia. Sólo los autores de la presente invención fueron capaces de observar una correlación con el estado de enfermedad y los taxones microbianos específicos.

EJEMPLO 4: Distorsión de la distribución de los componentes de una comunidad microbiana mediante la ligación directa de adaptadores de PCR en Emulsión en amplicones de PCR.

En un experimento los autores de la presente invención utilizaron cebadores de PCR marcados para amplificar los componentes en muestras de la comunidad microbiana duplicadas, ligaron los adaptadores de PCR en emulsión a estas muestras, y a continuación sometieron estas muestras a rondas de pirosecuenciación separadas. Los amplicones se ejecutan de forma rutinaria en un sistema Agilent Bioanalyzer antes y después de la ligación para cuantificar la mezcla antes de PCR en emulsión. La Figura 4 representa una ronda de la muestra en el Bioanalizador antes y después de la ligación directa y muestra claramente que la etapa de ligación ha alterado drásticamente la distribución de los amplicones.

Adicionalmente, los autores de la presente invención compararon la abundancia normalizadas de los taxones componentes identificados por medio del procedimiento multietiqueta después de la ligación directa de los adaptadores de PCR en emulsión. En este experimento, la identificación de los taxones se realizó utilizando un análisis Bayesiano de la base de datos de RDP 9.0. En la Figura 5 se puede observar que la abundancia de cebadores directo e inverso para varios taxones es diferente dentro de una muestra y entre muestras duplicadas. En varios casos, se pierden familias completas en la comparación entre los duplicados. La Tabla 3 resume las diferencias entre los cebadores directos y los cebadores inversos de las muestras duplicadas y esto es claramente estocástico sin un patrón predecible. Los autores de la presente invención postulan que esta eficacia de ligación diferencial podría ser debida a diversos factores tales como la estructura interna de los amplicones o los sesgos en el nucleótido terminal del adaptador o del amplicón.

Tabla 3 Análisis de muestras duplicadas

FAMILIA RDP 9.0	PROPORCIONES DE CEBADORES DIRECTOS	PROPORCIONES DE CEBADORES INVERSOS
Acidaminococcaceae	**544,6%**	**195,0%**
Actinomycetales	**144,0%**	**116,5%**
Bacteroidaceae	**119,9%**	**124,5%**
Clostridiaceae	**97,5%**	**99,4%**
Comamonadaceae	**198,0%**	
Coriobacteriales	**181,5%**	**141,5%**
Enterobacteriaceae	**4,2%**	
Eubacteriaceae	**88,0%**	**87,5%**
Flavobacteriaceae	**34,9%**	
Incertae sedis 9	**106,4%**	**143,0%**
Lachnospiraceae	**176,8%**	**113,1%**
Peptococcaceae		**91,0%**
Peptostreptococcaceae	**94,7%**	**115,4%**
Porphyromonadaceae	**99,0%**	**97,3%**
Prevotellaceae	**264,0%**	**88,1%**
Rikenellaceae	**212,2%**	**106,1%**
Streptococcaceae	**74,3%**	**60,7%**

Literatura Citada

Amann, R. (2000). "Who is out there? Microbial Aspects of Biodiversity." System. Appl. Microbiol. 23: 1-8.

Bowman, J. P. y R. D. McCuaig (2003). "Biodiversity, Community Structural Shifts, and Biogeography of Prokaryotes within Antarctic Continental Shelf Sediment." Appl. Environ. Microbiol. 69(5): 2463-2483.

Eckburg, P. B., E. M. Bik, et al. (2005). "Diversity of the human intestinal microbial flora." Science 308: 1635-1638.
Fitzgerald, K. M. (1999). Microbial Community Dynamics During the Bench-Scale Bioremediation of Petroleum-Contaminated Soil. Department of Biology. Fairfax, VA, George Mason University: 73.
Frischer, A. E., J. M. Danforth, et al. (2000). "Whole-cell versus total RNA extraction for analysis of microbial community structure with 16S rRNA-targeted oligonucleotide probes in salt marsh sediments." Appl. Environ. Microbiol. 66(7): 3037-3043.
Guarner, F., y J.R. Malagelada. (2003). "Gut flora in health and disease." Lancet 361: 512-9.
Kang, S. y A. L. Mills (2004). "Soil Bacterial Community Changes Following Disturbance of the Overlying Plant Community." Soil Science 169: 55-65.
Klaper, R. y M. Thomas (2004). "At the crossroads of genomics and ecology: the promise of a canary on a chip." BioScience 54: 403-412.
Litchfield, C. D. y P. M. Gillevet (2002). "Microbial diversity and complexity in hypersaline environments: A preliminary assessment." Journal of Industrial Microbiology & Biotechnology 28(1):48-55.
Lydell, C., L. Dowell, et al. (2004). "A population survey of members of the phylum Bacteroidetes isolated from salt marsh sediments along the east coast of the United States." Microbial ecology 48(2): 263-73.
Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature, 2005 Sep 15, 437(7057):376-80. Epub: 2005 Jul 31.
McCraig, A. E., L. Glover, J.I. Prosser (1999). "Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures." Appl. Environ. Microbiol. 65: 1721-1730.
Mills, D. (2000). Molecular Monitoring of Microbial Populations during Bioremediation of Contaminated Soils. Environmental Sciences and Public Policy/Biology. Fairfax, VA, George Mason University: 217.
Mills, D. K., K. Fitzgerald, et al. (2003). "A Comparison of DNA Profiling Techniques for Monitoring Nutrient Impact on Microbial Community Composition during Bioremediation of Petroleum Contaminated Soils." J. Microbiol. Method 54: 57-74.
Morris, C. E., M. Bardin, et al. (2002). "Microbial biodiversity: approaches to experimental design and hypothesis testing in primary scientific literature from 1975 to 1999." Microbiology and Molecular Biology Reviews 66: 592-616.
Pace, N. R. (1997). "A Molecular View of Microbial Diversity and the Biosphere." Science 276: 734-739.
Poinar, H. N., C. Schwarz, et al. (2006). "Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA." Science, 2006 Jan 20, 311(5759):392-4. Epub: 2005 Dec 20.
Pourmand, N., E. Elahi, et al. (2002). "Multiplex Pyrosequencing." Nucleic acids research 30(7): 31.
Ritchie, N. J., M. E. Schutter, et al. (2000). "Use of Length Heterogeneity PCR and Fatty Acid Methyl Ester Profiles to Characterize Microbial Communities in Soil." Applied and Environmental Microbiology 66(4): 1668-1675.
Schultz, M., y R.B. Sator. (2000). "Probiotics and inflammatory bowel disease." Am. J. of Gastroenterology 2000 Jan. 95 (1 Suppl): S19-21.
Seviour, R. J., T. Mino, et al. (2003). "The microbiology of biological phosphorus removal in activated sludge systems." FEMS Microbiology Reviews 27: 99-127.
Spring, S., R. Schulze, et al. (2000). "Identification and characterization of ecologically significant prokaryotes in the sediment of freshwater lakes: molecular and cultivation studies." FEMS Microbiology Reviews 24: 573-590.
Suzuki, M., M. S. Rappe, et al. (1998). "Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity." Applied and Environmental Microbiology [Appl. Environ. Microbiol.]. 64(11): 4522-4529.
Suzuki, M. T. (1998). The Effect of Protistan Bacterivory on Bacterioplankton Community Structure: Dissertation Abstracts International Part B Science and Engineering [Diss. Abst. Int. Pt. B - Sci. & Eng.]. Vol. 59, no. 2, [np]. Aug 1998.
Theron, J. y T. E. Cloete (2000). "Molecular techniques for determining microbial diversity and community structure in natural environment." Critical Reviews in Microbiology 26: 37-57.
Venter, J. C., S. Levy, et al. (2003). "Massive parallelism, randomness and genomic advances." Nature genetics, 2003 Mar, 33 Suppl: 219-27.

REIVINDICACIONES

1. Un método múltiplex para determinar cuantitativamente polinucleótidos diana de dos o más muestras, que comprende:

anclar una primera secuencia etiqueta específica de la muestra a uno o más polinucleótidos de una primera muestra;
anclar una segunda secuencia etiqueta específica de la muestra diferente de dicha primera secuencia etiqueta a uno o más polinucleótidos de una segunda muestra;
mezclar entre sí las muestras etiquetadas; y
secuenciar dichos polinucleótidos que comprenden dicha primera y dicha segunda etiquetas;
en donde las etiquetas específicas de la muestra están incluidas en los cebadores para la amplificación mediante PCR y las etiquetas específicas de la muestra se anclan a los polinucleótidos por medio de amplificación mediante PCR, y en donde los cebadores para la amplificación mediante PCR comprenden, en orden 5' a 3', un radical para la inmovilización y/o una secuencia para la amplificación mediante PCR, la etiqueta de secuencia específica de la muestra y una secuencia sonda específica para una secuencia diana localizada en posición 3' con respecto a una región genética variable;
comprendiendo el método adicionalmente las etapas de:

a partir de las secuencias etiqueta incluidas en las secuencias de polinucleótidos determinadas de este modo identificar la muestra en la que se produjeron las secuencias de polinucleótidos;
a partir de las secuencias de la región genética variable incluida en las secuencias de polinucleótidos determinadas de este modo identificar las variantes concretas de dicho elemento genético variable;
a partir de esta información determinar el número de veces que una o más variantes dadas aparecen en cada muestra; y
a partir del número de cada variante en los polinucleótidos determinados de este modo, cuantificar dichos polinucleótidos en dichas muestras.

2. El método de la reivindicación 1, en donde
el anclaje de la primera secuencia etiqueta específica de la muestra por medio de amplificación mediante PCR comprende:

hibridar un primer cebador a los polinucleótidos en una primera muestra, comprendiendo dicho primer cebador una primera secuencia etiqueta y una primera secuencia sonda específica para una primera secuencia diana, en donde dicha primera secuencia diana se encuentra en posición 3' con respecto a una región genética variable; y
elongar los moldes de los cebadores formados de esta manera para formar una primera población de polinucleótidos etiquetados que comprende: dicho primer cebador que incluye dicha primera secuencia etiqueta y secuencias de dicha región genética variable;

y en donde el anclaje de la segunda etiqueta específica de la muestra por medio de amplificación mediante PCR comprende:

hibridar un segundo cebador a los polinucleótidos en una segunda muestra, comprendiendo dicho segundo cebador una segunda secuencia etiqueta y una segunda secuencia sonda específica para una segunda secuencia diana, en donde dicha segunda secuencia diana se encuentra en posición 3' con respecto a la misma región genética variable como dicha primera secuencia diana, en donde adicionalmente dicha segunda secuencia sonda puede ser la misma que, o diferente de, dicha primera secuencia sonda; y
elongar los moldes de los cebadores formados de esta manera para formar una segunda población de polinucleótidos etiquetados que comprende: dicho segundo cebador que incluye dicha segunda secuencia etiqueta; y secuencias de dicha región genética variable;

y adicionalmente en donde la etapa de mezcla de la muestra etiquetado comprende mezclar entre sí dicha primera y segunda poblaciones.

3. El método de la reivindicación 1 o 2, en donde las secuencias se determinan cuantitativamente y sin transferencia de Southern y/o sin separar por tamaño los productos de extensión del cebador y/o sin electroforesis.

4. El método de cualquier reivindicación precedente, en donde:

(A) en cada uno de dichos cebadores utilizados en el método la secuencia sonda se encuentra en posición 3' con respecto a la secuencia etiqueta, pero no necesariamente adyacente a la misma;
(B) en cada uno de los cebadores utilizados en el método, la secuencia etiqueta es diferente de la secuencia etiqueta del otro cebador utilizado en el método; la secuencia etiqueta no es la secuencia complementaria a ningún otra secuencia etiqueta utilizada en el método; la secuencia etiqueta no contiene ninguna secuencia de homodinucleótido; las secuencias de empalme entre la secuencia etiqueta y las porciones adyacentes del

cebador, si las hubiera, no es una secuencia de homodinucleótido;
(C) en cada uno de los cebadores utilizados en el método la secuencia sonda es complementaria a la secuencia diana y la secuencia diana se encuentra localizada en posición 3' con respecto a una región genética variable; y
(D) cada uno de dichos cebadores se dispone por separado de los otros utilizados en el método.

5. Un método de cualquier reivindicación precedente, que comprende adicionalmente normalizar el número de apariciones determinadas para una secuencia diana dada.

6. Un método de la reivindicación 5, en donde el número de apariciones se normaliza dividiendo el número de apariciones determinadas para la secuencia de polinucleótidos dada por el número total de apariciones de secuencias de polinucleótidos en la muestra.

7. Un método de acuerdo con cualquier reivindicación precedente, en donde las secuencias etiqueta tienen 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, o 36 nucleótidos de longitud o cualquiera de sus combinaciones.

8. Un método de cualquier reivindicación precedente, en donde la secuenciación se lleva a cabo utilizando amplificación en emulsión de cuentas.

9. Un método de acuerdo con cualquier reivindicación precedente, en donde cada uno de dichos polinucleótidos se dispone individualmente en una cuenta aislado de otros polinucleótidos.

10. Un método de acuerdo con cualquier reivindicación precedente, en donde los polinucleótidos en dichas muestras se disponen aislados entre sí, se amplifican mientras están dispuestos de este modo, y los productos de amplificación de cada polinucleótido formado de esta manera se disponen por separado de los productos de amplificación de otros polinucleótidos.

11. Un método de acuerdo con cualquier reivindicación precedente, en donde los polinucleótidos en dichas muestras se inmovilizan individualmente sobre cuentas, los polinucleótidos inmovilizados de este modo se amplifican mediante PCR en emulsión de cuentas, la emulsión se resuelve, y las cuentas se disponen después separadamente entre sí para la secuenciación.

12. Un método de acuerdo con cualquier reivindicación precedente, en donde las secuencias de las variantes de una región genética variable se secuencian y dichas variantes son específicas de los organismos concretos.

13. Un método según la reivindicación 12, en donde dicha región genética variable comprende una o más de una secuencia de ARNr 16S variable, una secuencia de ARNr 18S variable, una secuencia de ARNr ITS variable, una secuencia mitocondrial o una secuencia de microsatélites.

14. Un método de acuerdo con cualquier reivindicación precedente, en donde las muestras son muestras de la comunidad microbiana y las secuencias de dichos polinucleótidos se utilizan para análisis clínico, análisis ambiental, análisis del suelo, o análisis de agua.

15. Un método de acuerdo con una cualquiera de las reivindicaciones 1 a 12, en donde las muestras son ADN genómico bacteriano fragmentado, ADN genómico de Archaea, ADN genómico fúngico, ADN genómico eucariótico, ADN de cloroplasto y/o ADN mitocondrial.

16. Un método de acuerdo con cualquier reivindicación precedente, en donde las secuencias se determinan mediante pirosecuenciación.

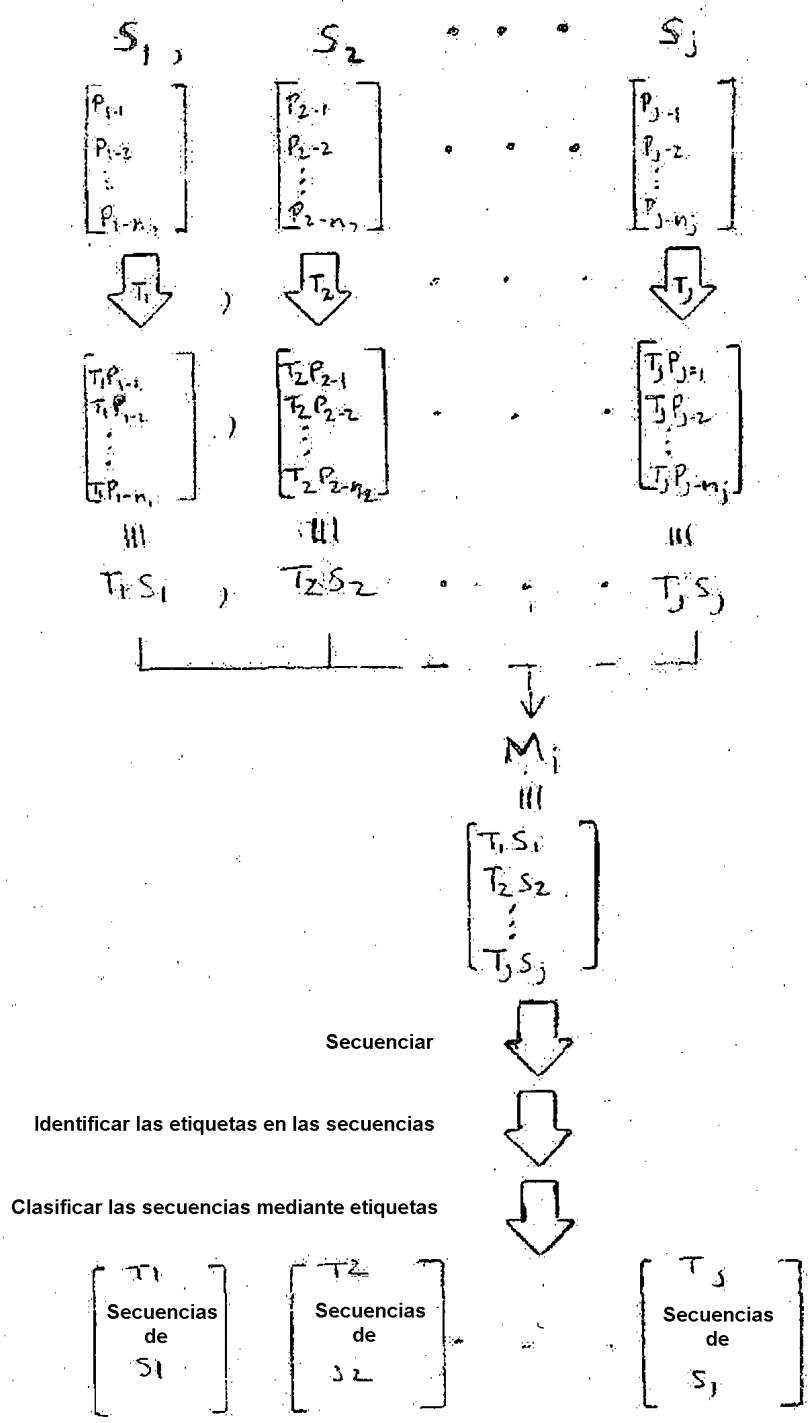
Secuenciar
Identificar las etiquetas en las secuencias
Clasificar las secuencias mediante etiquetas
Secuencias de S1
Secuencias de S2
Secuencias de Sj

Figura 1

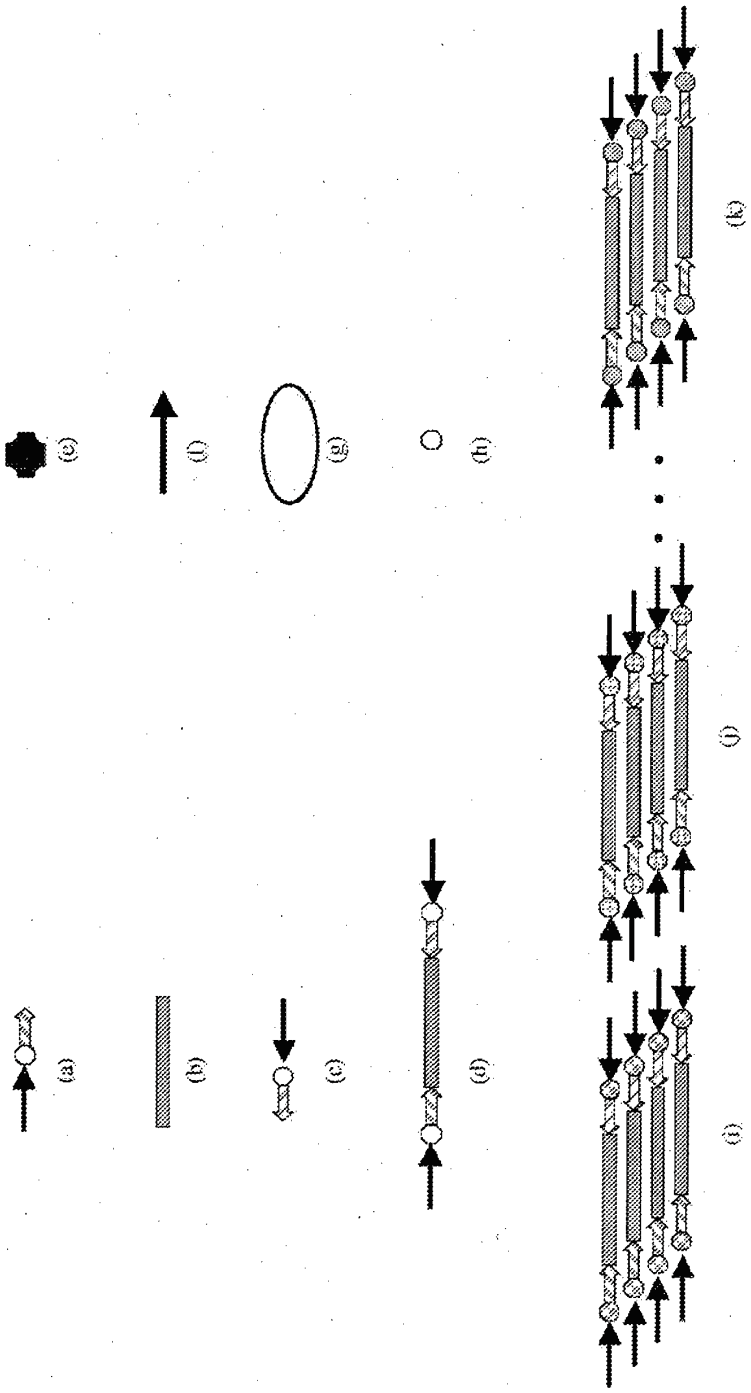

Figura 2A

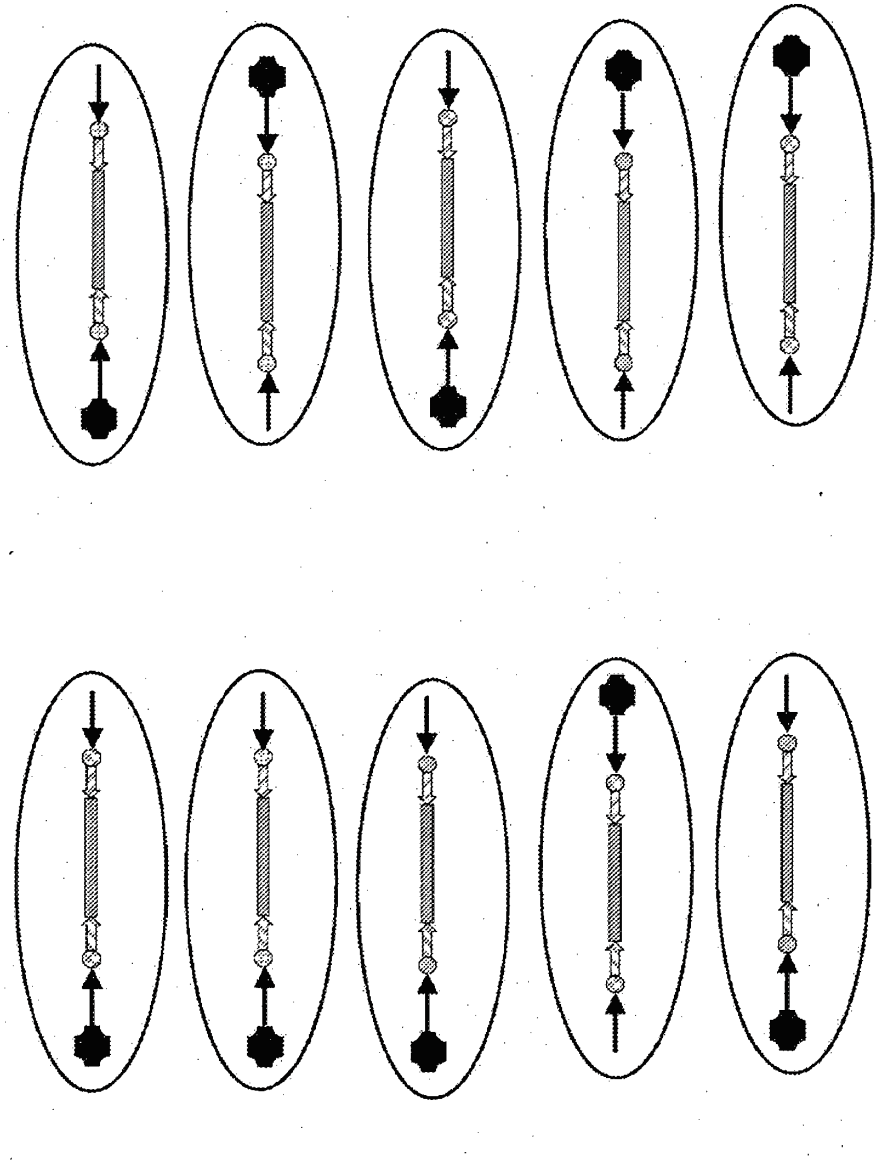

Figura 2B

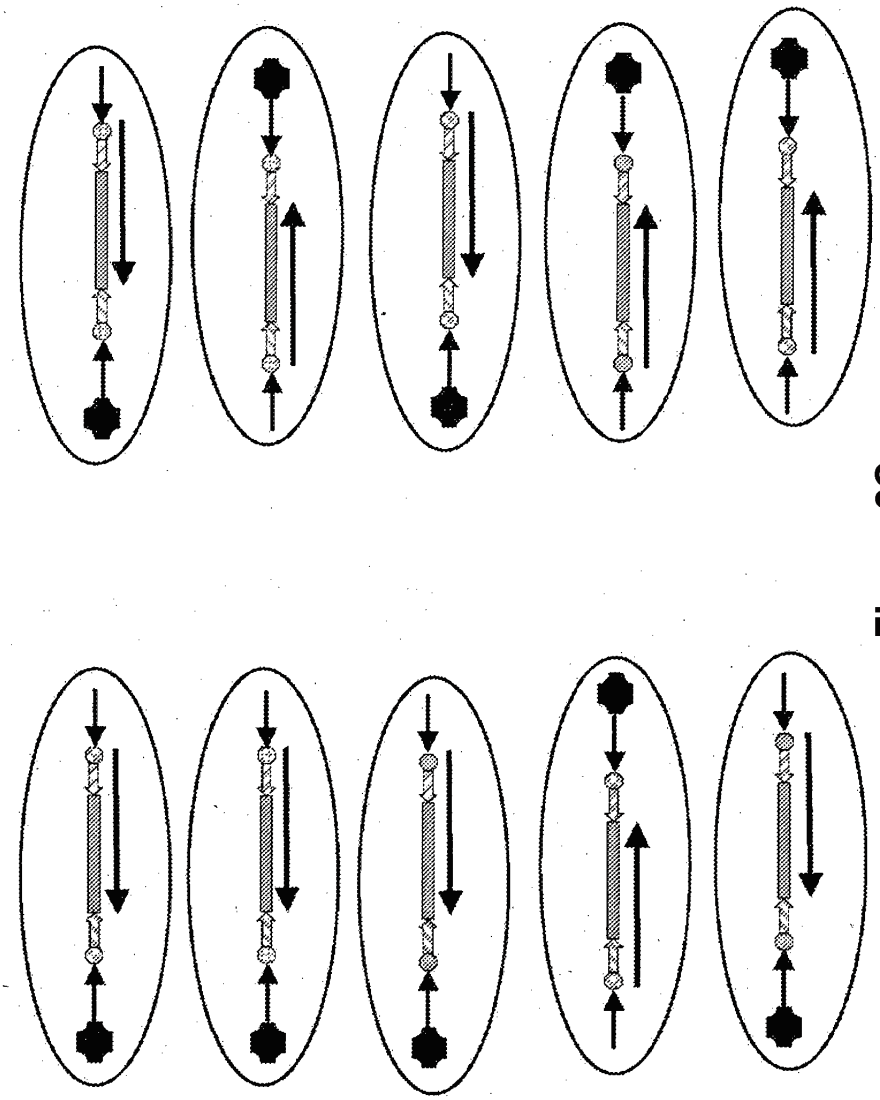

Figura 2C

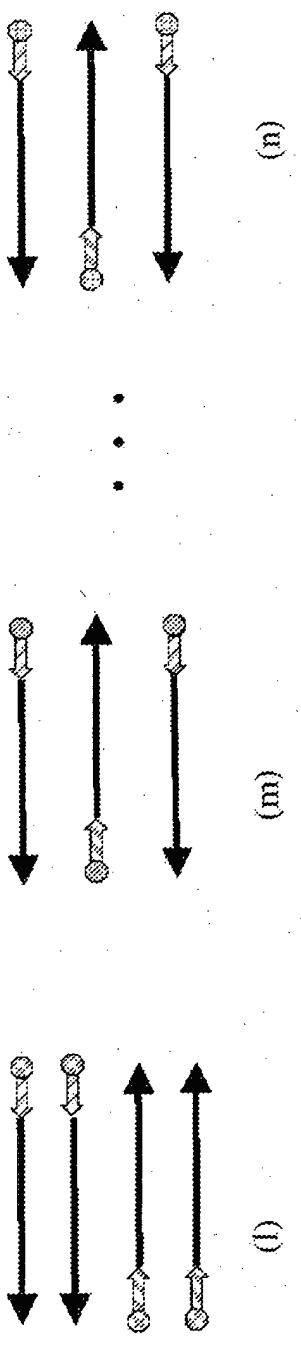

Figura 2D

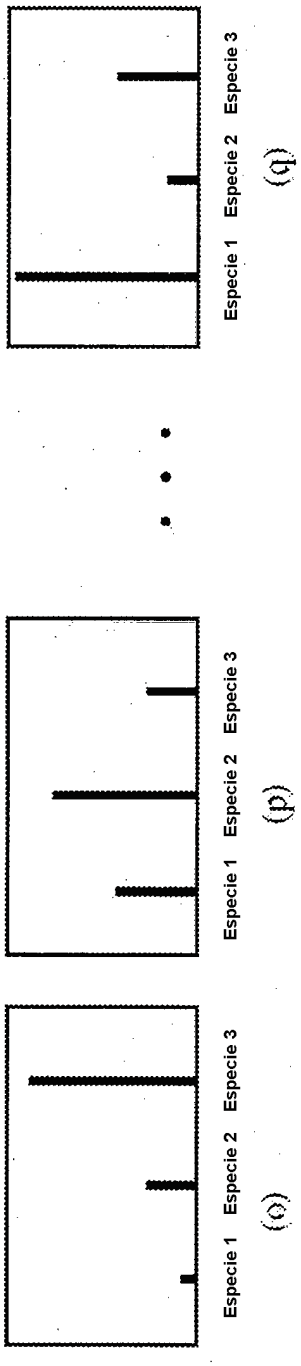

Figura 2E

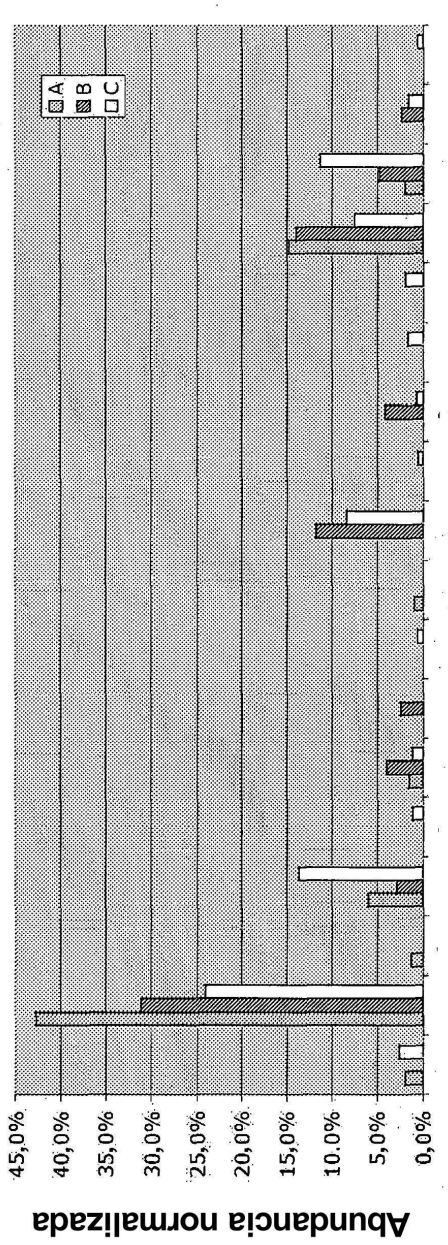
Abundancia normalizada
45,0%
40,0%
35,0%
30,0%
25,0%
20,0%
15,0%
10,0%
5,0%
0,0%
A
B
C
GRUPO_ALICYCLOBACILLUS
SUBGRUPO_BAC.FRAGILIS
SUBGRUPO_BTV.FIBRISOLVENS
SUBGRUPO_C.LEPTUM
SUBGRUPO_C.POLYSACCHAROLYTICUM
SUBGRUPO_C.XYLANOLYTICUM
SUBGRUPO_ESCHERICHIA
SUBGRUPO_EUB.CYLINDROIDES
GRUPO_FUSOBACTERIA
GRUPO_GLOEOTHECE-GLOEOCAPSA
SUBGRUPO_NSS.MULTIFORMIS
SUBGRUPO_PPM.MACCAE
SUBGRUPO_PRV.BRYANTII
SUBGRUPO_PRV.NIGRESCENS
SUBGRUPO_RUC.GNAVUS
SUBGRUPO_RUC.HANSENII
GRUPO_SPR.VOLUTANS
SUBGRUPO_VLN.PARVULA

Figura 3

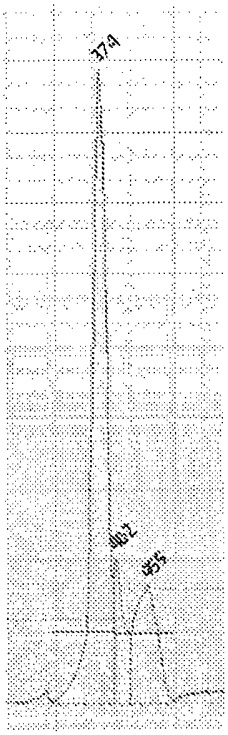

Figura 4A

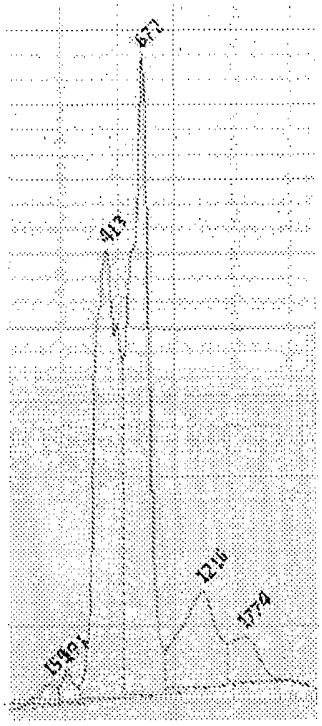

Figura 4B

FAMILIA	Cebador directo	Cebador inverso	RAZÓN	Cebador directo	Cebador ivnerso	RAZÓN
	ALD309FO	ALD309RO	Directo/Inverso	ALD309FP	ALD309RP	Directo/Inverso
Acidaminococcaceae	0,8%	0,9%	**88,5%**	0,2%	0,5%	**31,7%**
Actinomycetales	2,4%	2,0%	**120,7%**	1,7%	1,7%	**97,6%**
Bacteroidaceae	58,4%	62,8%	**93,1%**	48,7%	50,5%	**96.6%**
Brucellaceae				7,5%	3,1%	**239,1%**
Burkholderiaceae				0,4%	0,2%	**184.9%**
Clostridiaceae	14,5%	9,6%	**150,7%**	14,9%	9,7%	**153,7%**
Comamonadaceae	0,3%			0,2%	0,3%	**55,5%**
Coriobacteriales	0,8%	0,9%	**94,8%**	0,5%	0,6%	**73,9%**
Enterobacteriaceae	0,2%			3,6%	2,9%	**124,1%**
Eubacteriaceae	1,8%	1,6%	**115,9%**	2,1%	1,8%	**115,2%**
Flavobacteriaceae	0,5%			1,3%	2,1%	**62,9%**
Incertae sedis 9	3,3%	4,2%	**78,6%**	3,1%	2,9%	**105,6%**
Lachnospiraceae	5,7%	8,7%	**65,6%**	3,2%	7,7%	**42,0%**
Lactobacillaceae				1,3%	0,5%	**269,4%**
Peptococcaceae		0,3%		0,2%	0,3%	**83,2%**
Peptostreptococcaceae	3,3%	3,3%	**102,1%**	3,5%	2,8%	**124,4%**
Phyllobacteriaceae				1,6%	6,2%	**26,2%**
Porphyromonadaceae	2,1%	2,0%	**109,0%**	2,2%	2,0%	**107,1%**
Prevotellaceae	3,0%	2,0%	**155,7%**	1,2%	2,2%	**52,0%**
Rikenellaceae	1,1%	0,9%	**129,3%**	0,5%	0,8%	**64,7%**
Streptococcaceae	0,5%	0,3%	**181,1%**	0,6%	0,4%	**147,9%**

Figura 5

1	25	49	73	97	121	145	169	193
AGACAG	AGTGAG	ATGCAG	CATACA	CGCTAG	GCGCAG	GTCACA	TACTAG	TCAGAG
AGACGT	AGTGCA	ATGCGT	CATAGT	CGCTCA	GCGCGT	GTCAGT	TACTCA	TCAGCA
AGACTC	AGTGTC	ATGCTC	CATATC	CGCTGT	GCGCTC	GTCATC	TACTGT	TCAGTC
AGAGAG	ATACAG	ATGTAG	CATCAG	CGTACA	GCGTAG	GTCGAG	TAGACA	TCATAG
AGAGCA	ATACGT	ATGTCA	CATCGT	CGTAGT	GCGTCA	GTCGCA	TAGAGT	TCATCA
AGAGTC	ATACTC	ATGTGT	CATCTC	CGTATC	GCGTGT	GTCGTC	TAGATC	TCATGT
AGATAG	ATAGAG	CACACA	CATGAG	CGTCAG	GCTACA	GTCTAG	TAGCAG	TCGACA
AGATCA	ATAGCA	CACAGT	CATGCA	CGTCGT	GCTAGT	GTCTCA	TAGCGT	TCGAGT
AGATGT	ATAGTC	CACATC	CATGTC	CGTCTC	GCTATC	GTCTGT	TAGCTC	TCGATC
AGCACA	ATATAG	CACGAG	CGACAG	CGTGAG	GCTCAG	GTGACA	TAGTAG	TCGCAG
AGCAGT	ATATCA	CACGCA	CGACGT	CGTGCA	GCTCGT	GTGAGT	TAGTCA	TCGCGT
AGCATC	ATATGT	CACGTC	CGACTC	CGTGTC	GCTCTC	GTGATC	TAGTGT	TCGCTC
AGCGAG	ATCACA	CACTAG	CGAGAG	GCACAG	GCTGAG	GTGCAG	TATACA	TCGTAG
AGCGCA	ATCAGT	CACTCA	CGAGCA	GCACGT	GCTGCA	GTGCGT	TATAGT	TCGTCA
AGCGTC	ATCATC	CACTGT	CGAGTC	GCACTC	GCTGTC	GTGCTC	TATATC	TCGTGT
AGCTAG	ATCGAG	CAGACA	CGATAG	GCAGAG	GTACAG	GTGTAG	TATCAG	TCTACA
AGCTCA	ATCGCA	CAGAGT	CGATCA	GCAGCA	GTACGT	GTGTCA	TATCGT	TCTAGT
AGCTGT	ATCGTC	CAGATC	CGATGT	GCAGTC	GTACTC	GTGTGT	TATCTC	TCTATC
AGTACA	ATCTAG	CAGCAG	CGCACA	GCATAG	GTAGAG	TACACA	TATGAG	TCTCAG
AGTAGT	ATCTCA	CAGCGT	CGCAGT	GCATCA	GTAGCA	TACAGT	TATGCA	TCTCGT
AGTATC	ATCTGT	CAGCTC	CGCATC	GCATGT	GTAGTC	TACATC	TATGTC	TCTCTC
AGTCAG	ATGACA	CAGTAG	CGCGAG	GCGACA	GTATAG	TACGAG	TCACAG	TCTGAG
AGTCGT	ATGAGT	CAGTCA	CGCGCA	GCGAGT	GTATCA	TACGCA	TCACGT	TCTGCA
AGTCTC	ATGATC	CAGTGT	CGCGTC	GCGATC	GTATGT	TACGTC	TCACTC	TCTGTC
24	48	72	96	120	144	168	192	216

FIGURA 6

FIGURA 7A

CEBADOR ADAPTADOR ETIQUETADO DIRECTO

Par de cebadores	ADAPTADOR A	ETIQUETA	Cebador de ARNr 16S Directo
1	GCCTCCCTCGCGCCATCAG	AGACGT	AGAGTTTGATCMTGGCTCAG
2	GCCTCCCTCGCGCCATCAG	AGACTC	AGAGTTTGATCMTGGCTCAG
3	GCCTCCCTCGCGCCATCAG	AGAGTC	AGAGTTTGATCMTGGCTCAG
4	GCCTCCCTCGCGCCATCAG	AGATGT	AGAGTTTGATCMTGGCTCAG
5	GCCTCCCTCGCGCCATCAG	AGCAGT	AGAGTTTGATCMTGGCTCAG
6	GCCTCCCTCGCGCCATCAG	AGCATC	AGAGTTTGATCMTGGCTCAG
7	GCCTCCCTCGCGCCATCAG	AGCGTC	AGAGTTTGATCMTGGCTCAG
8	GCCTCCCTCGCGCCATCAG	AGCTGT	AGAGTTTGATCMTGGCTCAG
9	GCCTCCCTCGCGCCATCAG	AGTAGT	AGAGTTTGATCMTGGCTCAG
10	GCCTCCCTCGCGCCATCAG	AGTATC	AGAGTTTGATCMTGGCTCAG
11	GCCTCCCTCGCGCCATCAG	AGTCGT	AGAGTTTGATCMTGGCTCAG
12	GCCTCCCTCGCGCCATCAG	AGTCTC	AGAGTTTGATCMTGGCTCAG
13	GCCTCCCTCGCGCCATCAG	AGTGTC	AGAGTTTGATCMTGGCTCAG
14	GCCTCCCTCGCGCCATCAG	ATACGT	AGAGTTTGATCMTGGCTCAG
15	GCCTCCCTCGCGCCATCAG	ATACTC	AGAGTTTGATCMTGGCTCAG
16	GCCTCCCTCGCGCCATCAG	ATAGTC	AGAGTTTGATCMTGGCTCAG
17	GCCTCCCTCGCGCCATCAG	ATATGT	AGAGTTTGATCMTGGCTCAG
18	GCCTCCCTCGCGCCATCAG	ATCAGT	AGAGTTTGATCMTGGCTCAG
19	GCCTCCCTCGCGCCATCAG	ATCATC	AGAGTTTGATCMTGGCTCAG
20	GCCTCCCTCGCGCCATCAG	ATCGTC	AGAGTTTGATCMTGGCTCAG
21	GCCTCCCTCGCGCCATCAG	ATCTGT	AGAGTTTGATCMTGGCTCAG
22	GCCTCCCTCGCGCCATCAG	ATGAGT	AGAGTTTGATCMTGGCTCAG
23	GCCTCCCTCGCGCCATCAG	ATGATC	AGAGTTTGATCMTGGCTCAG
24	GCCTCCCTCGCGCCATCAG	ATGCGT	AGAGTTTGATCMTGGCTCAG
25	GCCTCCCTCGCGCCATCAG	ATGCTC	AGAGTTTGATCMTGGCTCAG
26	GCCTCCCTCGCGCCATCAG	ATGTGT	AGAGTTTGATCMTGGCTCAG
27	GCCTCCCTCGCGCCATCAG	CACAGT	AGAGTTTGATCMTGGCTCAG
28	GCCTCCCTCGCGCCATCAG	CACATC	AGAGTTTGATCMTGGCTCAG
29	GCCTCCCTCGCGCCATCAG	CACGTC	AGAGTTTGATCMTGGCTCAG
30	GCCTCCCTCGCGCCATCAG	CACTGT	AGAGTTTGATCMTGGCTCAG
31	GCCTCCCTCGCGCCATCAG	CAGAGT	AGAGTTTGATCMTGGCTCAG
32	GCCTCCCTCGCGCCATCAG	CAGATC	AGAGTTTGATCMTGGCTCAG
33	GCCTCCCTCGCGCCATCAG	CAGCGT	AGAGTTTGATCMTGGCTCAG
34	GCCTCCCTCGCGCCATCAG	CAGCTC	AGAGTTTGATCMTGGCTCAG
35	GCCTCCCTCGCGCCATCAG	CAGTGT	AGAGTTTGATCMTGGCTCAG
36	GCCTCCCTCGCGCCATCAG	CATAGT	AGAGTTTGATCMTGGCTCAG
37	GCCTCCCTCGCGCCATCAG	CATATC	AGAGTTTGATCMTGGCTCAG
38	GCCTCCCTCGCGCCATCAG	CATCGT	AGAGTTTGATCMTGGCTCAG
39	GCCTCCCTCGCGCCATCAG	CATCTC	AGAGTTTGATCMTGGCTCAG
40	GCCTCCCTCGCGCCATCAG	CATGTC	AGAGTTTGATCMTGGCTCAG
41	GCCTCCCTCGCGCCATCAG	CGACGT	AGAGTTTGATCMTGGCTCAG
42	GCCTCCCTCGCGCCATCAG	CGACTC	AGAGTTTGATCMTGGCTCAG
43	GCCTCCCTCGCGCCATCAG	CGAGTC	AGAGTTTGATCMTGGCTCAG
44	GCCTCCCTCGCGCCATCAG	CGATGT	AGAGTTTGATCMTGGCTCAG
45	GCCTCCCTCGCGCCATCAG	CGCAGT	AGAGTTTGATCMTGGCTCAG
46	GCCTCCCTCGCGCCATCAG	CGCATC	AGAGTTTGATCMTGGCTCAG
47	GCCTCCCTCGCGCCATCAG	CGCGTC	AGAGTTTGATCMTGGCTCAG
48	GCCTCCCTCGCGCCATCAG	CGCTGT	AGAGTTTGATCMTGGCTCAG
49	GCCTCCCTCGCGCCATCAG	CGTAGT	AGAGTTTGATCMTGGCTCAG
50	GCCTCCCTCGCGCCATCAG	CGTATC	AGAGTTTGATCMTGGCTCAG

FIGURA 7B

CEBADOR ADAPTADOR ETIQUETADO DIRECTO

Par de cebadores	ADAPTADOR A	ETIQUETA	Cebador de ARNr 16S Directo
51	GCCTCCCTCGCGCCATCAG	CGTCGT	AGAGTTTGATCMTGGCTCAG
52	GCCTCCCTCGCGCCATCAG	CGTCTC	AGAGTTTGATCMTGGCTCAG
53	GCCTCCCTCGCGCCATCAG	CGTGCA	AGAGTTTGATCMTGGCTCAG
54	GCCTCCCTCGCGCCATCAG	CGTGTC	AGAGTTTGATCMTGGCTCAG
55	GCCTCCCTCGCGCCATCAG	TACAGT	AGAGTTTGATCMTGGCTCAG
56	GCCTCCCTCGCGCCATCAG	TACATC	AGAGTTTGATCMTGGCTCAG
57	GCCTCCCTCGCGCCATCAG	TACGTC	AGAGTTTGATCMTGGCTCAG
58	GCCTCCCTCGCGCCATCAG	TACTGT	AGAGTTTGATCMTGGCTCAG
59	GCCTCCCTCGCGCCATCAG	TAGAGT	AGAGTTTGATCMTGGCTCAG
60	GCCTCCCTCGCGCCATCAG	TAGATC	AGAGTTTGATCMTGGCTCAG
61	GCCTCCCTCGCGCCATCAG	TAGCGT	AGAGTTTGATCMTGGCTCAG
62	GCCTCCCTCGCGCCATCAG	TAGCTC	AGAGTTTGATCMTGGCTCAG
63	GCCTCCCTCGCGCCATCAG	TAGTGT	AGAGTTTGATCMTGGCTCAG
64	GCCTCCCTCGCGCCATCAG	TATAGT	AGAGTTTGATCMTGGCTCAG
65	GCCTCCCTCGCGCCATCAG	TATATC	AGAGTTTGATCMTGGCTCAG
66	GCCTCCCTCGCGCCATCAG	TATCGT	AGAGTTTGATCMTGGCTCAG
67	GCCTCCCTCGCGCCATCAG	TATCTC	AGAGTTTGATCMTGGCTCAG
68	GCCTCCCTCGCGCCATCAG	TATGTC	AGAGTTTGATCMTGGCTCAG
69	GCCTCCCTCGCGCCATCAG	TCACGT	AGAGTTTGATCMTGGCTCAG
70	GCCTCCCTCGCGCCATCAG	TCACTC	AGAGTTTGATCMTGGCTCAG
71	GCCTCCCTCGCGCCATCAG	TCAGTC	AGAGTTTGATCMTGGCTCAG
72	GCCTCCCTCGCGCCATCAG	TCATGT	AGAGTTTGATCMTGGCTCAG
73	GCCTCCCTCGCGCCATCAG	TCGAGT	AGAGTTTGATCMTGGCTCAG
74	GCCTCCCTCGCGCCATCAG	TCGATC	AGAGTTTGATCMTGGCTCAG
75	GCCTCCCTCGCGCCATCAG	TCGCGT	AGAGTTTGATCMTGGCTCAG
76	GCCTCCCTCGCGCCATCAG	TCGCTC	AGAGTTTGATCMTGGCTCAG
77	GCCTCCCTCGCGCCATCAG	TCGTGT	AGAGTTTGATCMTGGCTCAG
78	GCCTCCCTCGCGCCATCAG	TCTAGT	AGAGTTTGATCMTGGCTCAG
79	GCCTCCCTCGCGCCATCAG	TCTATC	AGAGTTTGATCMTGGCTCAG
80	GCCTCCCTCGCGCCATCAG	TCTCGT	AGAGTTTGATCMTGGCTCAG
81	GCCTCCCTCGCGCCATCAG	TCTCTC	AGAGTTTGATCMTGGCTCAG
82	GCCTCCCTCGCGCCATCAG	TCTGTC	AGAGTTTGATCMTGGCTCAG
83	GCCTCCCTCGCGCCATCAG	AGACAG	AGAGTTTGATCMTGGCTCAG
84	GCCTCCCTCGCGCCATCAG	AGAGAG	AGAGTTTGATCMTGGCTCAG
85	GCCTCCCTCGCGCCATCAG	AGATAG	AGAGTTTGATCMTGGCTCAG
86	GCCTCCCTCGCGCCATCAG	AGCGAG	AGAGTTTGATCMTGGCTCAG
87	GCCTCCCTCGCGCCATCAG	AGCTAG	AGAGTTTGATCMTGGCTCAG
88	GCCTCCCTCGCGCCATCAG	AGTCAG	AGAGTTTGATCMTGGCTCAG
89	GCCTCCCTCGCGCCATCAG	AGTGAG	AGAGTTTGATCMTGGCTCAG
90	GCCTCCCTCGCGCCATCAG	ATACAG	AGAGTTTGATCMTGGCTCAG
91	GCCTCCCTCGCGCCATCAG	ATAGAG	AGAGTTTGATCMTGGCTCAG
92	GCCTCCCTCGCGCCATCAG	ATATAG	AGAGTTTGATCMTGGCTCAG
93	GCCTCCCTCGCGCCATCAG	ATCGAG	AGAGTTTGATCMTGGCTCAG
94	GCCTCCCTCGCGCCATCAG	ATCTAG	AGAGTTTGATCMTGGCTCAG
95	GCCTCCCTCGCGCCATCAG	ATGCAG	AGAGTTTGATCMTGGCTCAG
96	GCCTCCCTCGCGCCATCAG	ATGTAG	AGAGTTTGATCMTGGCTCAG

FIGURA 7C

CABADOR ADAPTADOR ETIQUETADO INVERSO

Par de cebadores	ADAPTADOR B	ETIQUETA	Cebador de ARNr 16S Inverso
1	GCCTTGCCAGCCCGCTCAG	AGACGT	GCTGCCTCCCGTAGGAGT
2	GCCTTGCCAGCCCGCTCAG	AGACTC	GCTGCCTCCCGTAGGAGT
3	GCCTTGCCAGCCCGCTCAG	AGAGTC	GCTGCCTCCCGTAGGAGT
4	GCCTTGCCAGCCCGCTCAG	AGATGT	GCTGCCTCCCGTAGGAGT
5	GCCTTGCCAGCCCGCTCAG	AGCAGT	GCTGCCTCCCGTAGGAGT
6	GCCTTGCCAGCCCGCTCAG	AGCATC	GCTGCCTCCCGTAGGAGT
7	GCCTTGCCAGCCCGCTCAG	AGCGTC	GCTGCCTCCCGTAGGAGT
8	GCCTTGCCAGCCCGCTCAG	AGCTGT	GCTGCCTCCCGTAGGAGT
9	GCCTTGCCAGCCCGCTCAG	AGTAGT	GCTGCCTCCCGTAGGAGT
10	GCCTTGCCAGCCCGCTCAG	AGTATC	GCTGCCTCCCGTAGGAGT
11	GCCTTGCCAGCCCGCTCAG	AGTCGT	GCTGCCTCCCGTAGGAGT
12	GCCTTGCCAGCCCGCTCAG	AGTCTC	GCTGCCTCCCGTAGGAGT
13	GCCTTGCCAGCCCGCTCAG	AGTGTC	GCTGCCTCCCGTAGGAGT
14	GCCTTGCCAGCCCGCTCAG	ATACGT	GCTGCCTCCCGTAGGAGT
15	GCCTTGCCAGCCCGCTCAG	ATACTC	GCTGCCTCCCGTAGGAGT
16	GCCTTGCCAGCCCGCTCAG	ATAGTC	GCTGCCTCCCGTAGGAGT
17	GCCTTGCCAGCCCGCTCAG	ATATGT	GCTGCCTCCCGTAGGAGT
18	GCCTTGCCAGCCCGCTCAG	ATCAGT	GCTGCCTCCCGTAGGAGT
19	GCCTTGCCAGCCCGCTCAG	ATCATC	GCTGCCTCCCGTAGGAGT
20	GCCTTGCCAGCCCGCTCAG	ATCGTC	GCTGCCTCCCGTAGGAGT
21	GCCTTGCCAGCCCGCTCAG	ATCTGT	GCTGCCTCCCGTAGGAGT
22	GCCTTGCCAGCCCGCTCAG	ATGAGT	GCTGCCTCCCGTAGGAGT
23	GCCTTGCCAGCCCGCTCAG	ATGATC	GCTGCCTCCCGTAGGAGT
24	GCCTTGCCAGCCCGCTCAG	ATGCGT	GCTGCCTCCCGTAGGAGT
25	GCCTTGCCAGCCCGCTCAG	ATGCTC	GCTGCCTCCCGTAGGAGT
26	GCCTTGCCAGCCCGCTCAG	ATGTGT	GCTGCCTCCCGTAGGAGT
27	GCCTTGCCAGCCCGCTCAG	CACAGT	GCTGCCTCCCGTAGGAGT
28	GCCTTGCCAGCCCGCTCAG	CACATC	GCTGCCTCCCGTAGGAGT
29	GCCTTGCCAGCCCGCTCAG	CACGTC	GCTGCCTCCCGTAGGAGT
30	GCCTTGCCAGCCCGCTCAG	CACTGT	GCTGCCTCCCGTAGGAGT
31	GCCTTGCCAGCCCGCTCAG	CAGAGT	GCTGCCTCCCGTAGGAGT
32	GCCTTGCCAGCCCGCTCAG	CAGATC	GCTGCCTCCCGTAGGAGT
33	GCCTTGCCAGCCCGCTCAG	CAGCGT	GCTGCCTCCCGTAGGAGT
34	GCCTTGCCAGCCCGCTCAG	CAGCTC	GCTGCCTCCCGTAGGAGT
35	GCCTTGCCAGCCCGCTCAG	CAGTGT	GCTGCCTCCCGTAGGAGT
36	GCCTTGCCAGCCCGCTCAG	CATAGT	GCTGCCTCCCGTAGGAGT
37	GCCTTGCCAGCCCGCTCAG	CATATC	GCTGCCTCCCGTAGGAGT
38	GCCTTGCCAGCCCGCTCAG	CATCGT	GCTGCCTCCCGTAGGAGT
39	GCCTTGCCAGCCCGCTCAG	CATCTC	GCTGCCTCCCGTAGGAGT
40	GCCTTGCCAGCCCGCTCAG	CATGTC	GCTGCCTCCCGTAGGAGT
41	GCCTTGCCAGCCCGCTCAG	CGACGT	GCTGCCTCCCGTAGGAGT
42	GCCTTGCCAGCCCGCTCAG	CGACTC	GCTGCCTCCCGTAGGAGT
43	GCCTTGCCAGCCCGCTCAG	CGAGTC	GCTGCCTCCCGTAGGAGT
44	GCCTTGCCAGCCCGCTCAG	CGATGT	GCTGCCTCCCGTAGGAGT
45	GCCTTGCCAGCCCGCTCAG	CGCAGT	GCTGCCTCCCGTAGGAGT
46	GCCTTGCCAGCCCGCTCAG	CGCATC	GCTGCCTCCCGTAGGAGT
47	GCCTTGCCAGCCCGCTCAG	CGCGTC	GCTGCCTCCCGTAGGAGT
48	GCCTTGCCAGCCCGCTCAG	CGCTGT	GCTGCCTCCCGTAGGAGT
49	GCCTTGCCAGCCCGCTCAG	CGTAGT	GCTGCCTCCCGTAGGAGT
50	GCCTTGCCAGCCCGCTCAG	CGTATC	GCTGCCTCCCGTAGGAGT

FIGURA 7D

CEBADOR ADAPTADOR ETIQUETADO INVERSO

Par de cebadores	ADAPTADOR B	ETIQUETA	Cebador de ARNr 16S Inverso
51	GCCTTGCCAGCCCGCTCAG	CGTCGT	GCTGCCTCCCGTAGGAGT
52	GCCTTGCCAGCCCGCTCAG	CGTCTC	GCTGCCTCCCGTAGGAGT
53	GCCTTGCCAGCCCGCTCAG	CGTGCA	GCTGCCTCCCGTAGGAGT
54	GCCTTGCCAGCCCGCTCAG	CGTGTC	GCTGCCTCCCGTAGGAGT
55	GCCTTGCCAGCCCGCTCAG	TACAGT	GCTGCCTCCCGTAGGAGT
56	GCCTTGCCAGCCCGCTCAG	TACATC	GCTGCCTCCCGTAGGAGT
57	GCCTTGCCAGCCCGCTCAG	TACGTC	GCTGCCTCCCGTAGGAGT
58	GCCTTGCCAGCCCGCTCAG	TACTGT	GCTGCCTCCCGTAGGAGT
59	GCCTTGCCAGCCCGCTCAG	TAGAGT	GCTGCCTCCCGTAGGAGT
60	GCCTTGCCAGCCCGCTCAG	TAGATC	GCTGCCTCCCGTAGGAGT
61	GCCTTGCCAGCCCGCTCAG	TAGCGT	GCTGCCTCCCGTAGGAGT
62	GCCTTGCCAGCCCGCTCAG	TAGCTC	GCTGCCTCCCGTAGGAGT
63	GCCTTGCCAGCCCGCTCAG	TAGTGT	GCTGCCTCCCGTAGGAGT
64	GCCTTGCCAGCCCGCTCAG	TATAGT	GCTGCCTCCCGTAGGAGT
65	GCCTTGCCAGCCCGCTCAG	TATATC	GCTGCCTCCCGTAGGAGT
66	GCCTTGCCAGCCCGCTCAG	TATCGT	GCTGCCTCCCGTAGGAGT
67	GCCTTGCCAGCCCGCTCAG	TATCTC	GCTGCCTCCCGTAGGAGT
68	GCCTTGCCAGCCCGCTCAG	TATGTC	GCTGCCTCCCGTAGGAGT
69	GCCTTGCCAGCCCGCTCAG	TCACGT	GCTGCCTCCCGTAGGAGT
70	GCCTTGCCAGCCCGCTCAG	TCACTC	GCTGCCTCCCGTAGGAGT
71	GCCTTGCCAGCCCGCTCAG	TCAGTC	GCTGCCTCCCGTAGGAGT
72	GCCTTGCCAGCCCGCTCAG	TCATGT	GCTGCCTCCCGTAGGAGT
73	GCCTTGCCAGCCCGCTCAG	TCGAGT	GCTGCCTCCCGTAGGAGT
74	GCCTTGCCAGCCCGCTCAG	TCGATC	GCTGCCTCCCGTAGGAGT
75	GCCTTGCCAGCCCGCTCAG	TCGCGT	GCTGCCTCCCGTAGGAGT
76	GCCTTGCCAGCCCGCTCAG	TCGCTC	GCTGCCTCCCGTAGGAGT
77	GCCTTGCCAGCCCGCTCAG	TCGTGT	GCTGCCTCCCGTAGGAGT
78	GCCTTGCCAGCCCGCTCAG	TCTAGT	GCTGCCTCCCGTAGGAGT
79	GCCTTGCCAGCCCGCTCAG	TCTATC	GCTGCCTCCCGTAGGAGT
80	GCCTTGCCAGCCCGCTCAG	TCTCGT	GCTGCCTCCCGTAGGAGT
81	GCCTTGCCAGCCCGCTCAG	TCTCTC	GCTGCCTCCCGTAGGAGT
82	GCCTTGCCAGCCCGCTCAG	TCTGTC	GCTGCCTCCCGTAGGAGT
83	GCCTTGCCAGCCCGCTCAG	AGAGCA	GCTGCCTCCCGTAGGAGT
84	GCCTTGCCAGCCCGCTCAG	AGATCA	GCTGCCTCCCGTAGGAGT
85	GCCTTGCCAGCCCGCTCAG	AGCACA	GCTGCCTCCCGTAGGAGT
86	GCCTTGCCAGCCCGCTCAG	AGCGCA	GCTGCCTCCCGTAGGAGT
87	GCCTTGCCAGCCCGCTCAG	AGCTCA	GCTGCCTCCCGTAGGAGT
88	GCCTTGCCAGCCCGCTCAG	AGTACA	GCTGCCTCCCGTAGGAGT
89	GCCTTGCCAGCCCGCTCAG	AGTGCA	GCTGCCTCCCGTAGGAGT
90	GCCTTGCCAGCCCGCTCAG	ATAGCA	GCTGCCTCCCGTAGGAGT
91	GCCTTGCCAGCCCGCTCAG	ATATCA	GCTGCCTCCCGTAGGAGT
92	GCCTTGCCAGCCCGCTCAG	ATCACA	GCTGCCTCCCGTAGGAGT
93	GCCTTGCCAGCCCGCTCAG	ATCGCA	GCTGCCTCCCGTAGGAGT
94	GCCTTGCCAGCCCGCTCAG	ATCTCA	GCTGCCTCCCGTAGGAGT
95	GCCTTGCCAGCCCGCTCAG	ATGACA	GCTGCCTCCCGTAGGAGT
96	GCCTTGCCAGCCCGCTCAG	ATGTCA	GCTGCCTCCCGTAGGAGT