(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 697 804**

(51) Int. Cl.:

C12Q 1/68 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA MODIFICADA TRAS OPOSICIÓN T5

(86) Fecha de presentación y número de la solicitud internacional: **22.05.2015 PCT/GB2015/051518**

(87) Fecha y número de publicación internacional: **26.11.2015 WO15177570**

(96) Fecha de presentación y número de la solicitud europea: **22.05.2015 E 15727042 (2)**

(97) Fecha y número de publicación de la concesión europea modificada tras oposición: **31.07.2024 EP 3146070**

(54) Título: **Proceso de secuenciación**

(30) Prioridad:

23.05.2014 GB 201409282

(45) Fecha de publicación y mención en BOPI de la traducción de la patente modificada:
25.11.2024

(73) Titular/es:

ILLUMINA SINGAPORE PTE. LTD. (100.0%)
29 WOODLANDS INDUSTRIAL PARK E1, Internal address #02-13/18
NORTHTECH 757716, SG

(72) Inventor/es:

BURKE, CATHERINE MAREE y
DARLING, AARON EARL

(74) Agente/Representante:

GONZÁLEZ PECES, Gustavo Adolfo

DESCRIPCIÓN

Proceso de secuenciación

Campo técnico

La presente invención se refiere a procedimientos para la generación de las secuencias de moléculas de ácidos nucleicos de molde, a procedimientos implementados por ordenador para la determinación de las secuencias de al menos dos moléculas de ácidos nucleicos de molde, a programas informáticos adaptados para llevar a cabo los procedimientos y a un medio legible por ordenador que almacena los programas informáticos.

Antecedentes

En general, existen dificultades en la secuenciación de secuencias de ácidos nucleicos largas (por ejemplo, aquellas mayores de 1 Kpb) de una forma efectiva y rápida. Actualmente, la tecnología de secuenciación puede producir bien unos grandes volúmenes de lecturas de secuencia corta (es decir, secuencias de moléculas cortas de ácidos nucleicos) o bien unas pequeñas cantidades de lecturas de secuencia larga. Actualmente es difícil secuenciar unas grandes cantidades de lecturas de secuencia larga.

El gen del ARNr 16S se usa para estudios filogenéticos, ya que está muy conservado entre las diferentes especies de bacterias y arqueas. Además de unos sitios de unión al cebador muy conservados, las secuencias del gen del ARNr 16S contienen regiones hipervariables que pueden proporcionar unas secuencias de firma específicas de la especie útiles para la identificación de bacterias. Como resultado, la secuenciación del gen del ARNr 16S se ha vuelto predominante en la microbiología médica como una alternativa rápida y barata a los procedimientos fenotípicos de identificación de bacterias. Además, aunque originalmente se usaba para identificar bacterias, posteriormente se averiguó que la secuenciación del 16S era capaz de reclasificar las bacterias en especies o incluso géneros completamente nuevos. También se ha transformado en uno de los criterios principales usados para identificar y describir nuevas especies de bacterias, tanto en cultivos de laboratorio como en muestras medioambientales no cultivadas. Sin embargo, el uso del análisis de la secuencia del ARNr 16S está obstaculizado debido a las dificultades asociadas con la secuenciación de grandes cantidades de moléculas de ácidos nucleicos mayores de 1 Kpb. Esto ha significado que, en general, la mayor parte de los investigadores que llevan a cabo un análisis de la secuencia del 16S tienden a centrarse en las regiones cortas, de hasta 500 pb, del gen 16S. La secuenciación de dichas regiones cortas da como resultado una escasez en la resolución taxonómica.

Además, los procedimientos de secuenciación generales tienden a carecer de precisión debido a los acontecimientos de recombinación que pueden producirse durante el proceso de secuenciación. La secuenciación implica las etapas de amplificar las moléculas de ácidos nucleicos que van a ser secuenciadas. Durante estas etapas de amplificación pueden producirse acontecimientos de recombinación. Esto puede significar que cuando las muestras de moléculas de ácidos nucleicos contienen genes con unas secuencias similares, los procedimientos de secuenciación generarán no solo las secuencias de los genes originales, sino también las secuencias de las moléculas de ácidos nucleicos producidas a través de la recombinación entre estos genes similares. Dado que los genes de ARNr 16S tienden a ser similares a lo largo de las diferentes especies, una molécula de un ácido nucleico de molde en una muestra de moléculas de ácidos nucleicos que comprende moléculas de ácidos nucleicos de múltiples genes diferentes del ARNr 16S puede recombinar durante la secuenciación. Dichos acontecimientos de recombinación se vuelven cada vez más frecuentes según crece la cantidad de amplificación requerida para el análisis de la muestra, especialmente a los niveles requeridos para analizar ciertas muestras de la microbiota asociada a un hospedador, y forenses. Por lo tanto, es beneficioso, cuando se van a secuenciar muestras de ADN que tienen genes del ARNr 16S, ser capaces de identificar y de eliminar las secuencias de ácidos nucleicos producidas a través de la recombinación.

Los procedimientos informáticos para la detección de una recombinación son limitados, sin embargo, debido a que solo pueden detectar los acontecimientos de recombinación que se producen entre dos moléculas parentales que son sustancialmente diferentes en su secuencia. La recombinación entre secuencias muy similares (por ejemplo, > 97 % de identidad) sigue siendo difícil de discriminar con respecto a una verdadera diversidad biológica mediante el uso de procedimientos informáticos. Actualmente no existen metodologías moleculares para potenciar la precisión de la detección informática de una recombinación.

Previamente se han descrito metodologías para potenciar la longitud de lectura de los instrumentos de secuenciación de alto rendimiento. Entre estas están las metodologías de reducción de la complejidad, tales como Moleculo de Illumina, que asigna unos códigos de barras únicos a grupos de 100s de moléculas de ADN, y procedimientos de etiquetado molecular, que añaden un código de barras único a cada molécula individual de una muestra. Ambas metodologías reconstruyen las moléculas de molde originales mediante el análisis de una colección de lecturas cortas pertenecientes a cada código de barras, reconstruyendo informáticamente una secuencia consenso de los moldes originales. Ambas metodologías dependen de la amplificación para crear muchas copias de los conjuntos con códigos de barras o de las moléculas individuales etiquetadas. Sin embargo, ninguna de estas metodologías previas emplea un sistema molecular para la detección de un error de recombinación *in vitro* introducido por la amplificación.

Sumario de la invención

Los presentes inventores han desarrollado una técnica que permite la secuenciación de largas secuencias de ácidos nucleicos de una forma rápida y precisa. Esta técnica puede usarse en muchas aplicaciones diferentes, pero es particularmente ventajosa para su uso en la secuenciación del gen del ARNr 16S, dado que puede usarse para generar grandes volúmenes de lecturas largas que abarcan la longitud total del gen de 1,5 Kpb. Por lo tanto, esta técnica puede usarse para la secuenciación de la totalidad del gen del ARNr 16S proporcionando una mayor resolución taxonómica que los procedimientos previos que implicaban la secuenciación de regiones más cortas del gen del ARNr 16S.

Además, los presentes inventores han desarrollado una técnica que permite identificar y descartar las secuencias de los productos de recombinación generados durante el proceso de secuenciación. Esto mejora la sensibilidad y la precisión de la secuenciación en general, y dicha precisión mejora la resolución taxonómica cuando se usa la técnica para estudios filogenéticos que usan la secuenciación del 16S.

En un primer aspecto de la presente invención se proporciona un procedimiento para la generación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo individual que es mayor de 1 Kpb en tamaño que comprende:

a) proporcionar al menos una muestra de moléculas de ácidos nucleicos que comprende al menos dos moléculas de ácidos nucleicos de molde objetivo que son mayores de 1 Kpb en tamaño;

b) introducir una primera etiqueta molecular en un extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo y una segunda etiqueta molecular en el otro extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo para proporcionar al menos dos moléculas de ácidos nucleicos de molde etiquetadas, en las que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular;

c) amplificar las al menos dos moléculas de ácidos nucleicos de molde etiquetadas para proporcionar múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas;

d) aislar una fracción de las múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas y fragmentar las moléculas de ácidos nucleicos de molde etiquetadas de la fracción para proporcionar múltiples moléculas de ácidos nucleicos de molde fragmentadas;

e) secuenciar las regiones de las múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular y la segunda etiqueta molecular;

f) secuenciar las múltiples moléculas de ácidos nucleicos de molde fragmentadas; y

g) reconstruir una secuencia consenso para al menos una de las al menos dos moléculas de ácidos nucleicos de molde objetivo de las secuencias que comprenden al menos un subconjunto de las secuencias producidas en la etapa f), en donde la etapa g) comprende:

(i) identificar agregados de secuencias de regiones de las múltiples copias de las al menos dos moléculas de ácido nucleico de molde etiquetadas que es probable que correspondan a la misma molécula de ácido nucleico de molde objetivo individual mediante la asignación de secuencias que comprenden primeras secuencias de etiqueta molecular que son homólogas entre sí y segundas secuencias de etiqueta molecular que son homólogas entre sí con respecto al mismo agregado;

(ii) analizar las secuencias de las múltiples moléculas de ácido nucleico de molde fragmentadas para identificar secuencias de las múltiples moléculas de ácido nucleico de molde fragmentadas que comprenden una primera etiqueta molecular que es homóloga a la primera etiqueta molecular de las secuencias de un primer agregado y secuencias de las múltiples moléculas de ácido nucleico de molde fragmentadas que comprenden una segunda etiqueta molecular que es homóloga a la segunda etiqueta molecular de las secuencias del primer agregado;

(iii) reconstruir la secuencia de una primera molécula de ácido nucleico de molde alineando secuencias que comprenden al menos un subconjunto de las secuencias de las múltiples moléculas de ácido nucleico de molde fragmentadas identificadas en la etapa (ii) y definir una secuencia de consenso a partir de estas secuencias; y

(iv) llevar a cabo las etapas (i) a (iii) con respecto a una segunda molécula de ácido nucleico de molde y/o a una adicional.

En un segundo aspecto de la presente invención se proporciona un procedimiento implementado por ordenador para la determinación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo individual que comprende las siguientes etapas:

(a) obtener los datos que comprenden las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas en los que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprende una primera etiqueta molecular en un extremo y una segunda etiqueta molecular en el otro extremo, en los que cada molécula de ácido nucleico de molde objetivo está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular y en los que las regiones comprenden la primera etiqueta molecular y la segunda etiqueta molecular;

(b) analizar los datos que comprenden las secuencias de las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular y la segunda etiqueta molecular para identificar los agregados de las secuencias que es probable que se correspondan con la misma molécula de ácido nucleico de molde objetivo individual mediante la asignación de unas secuencias que comprenden unas primeras etiquetas moleculares que son homólogas entre sí y unas segundas etiquetas moleculares que son homólogas entre sí al mismo agregado;

(c) obtener los datos que comprenden las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en los que cada uno de los fragmentos comprende bien la primera etiqueta molecular o bien la segunda etiqueta molecular;

(d) analizar las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas para identificar las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular que es homóloga de la primera etiqueta molecular de las secuencias de un primer agregado, o la segunda etiqueta molecular que es homóloga de la segunda etiqueta molecular de las secuencias del primer agregado;

(e) reconstruir la secuencia de una primera molécula de ácido nucleico de molde objetivo mediante la alineación de las secuencias que comprenden al menos un subconjunto de las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas identificadas en la etapa (d) y definir una secuencia consenso a partir de estas secuencias; y

(f) llevar a cabo las etapas (c) hasta (e) con respecto a una segunda molécula de ácido nucleico de molde y/o a una adicional.

En un tercer aspecto de la invención se proporciona un procedimiento implementado por ordenador para la determinación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo que comprende las siguientes etapas:

(a) obtener los datos que comprenden agregados de secuencias en los que:

(i) cada agregado comprende las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas en los que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprende una primera etiqueta molecular en un extremo y una segunda etiqueta molecular en el otro extremo, en los que cada una de las al menos dos ácidos nucleicos de molde objetivo está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular y en los que las regiones comprenden la primera etiqueta molecular y la segunda etiqueta molecular;

(ii) cada agregado comprende las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que cada uno de los fragmentos comprende bien la primera etiqueta molecular o bien la segunda etiqueta molecular;

(iii) las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas en cada agregado comprenden la primera etiqueta molecular y la segunda etiqueta molecular que son homólogas entre sí

(iv) las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprenden la primera etiqueta molecular que es homóloga de la primera etiqueta molecular de las secuencias de las regiones de las múltiples copias de al menos dos moléculas de ácidos nucleicos de molde objetivo en ese agregado, o la segunda etiqueta molecular que es homóloga de la segunda etiqueta molecular de las secuencias de las regiones de múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en ese agregado;

(b) reconstruir la secuencia de una primera molécula de ácido nucleico de molde mediante la alineación de las secuencias que comprenden al menos un subconjunto de las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en un primer agregado y definir una secuencia

consenso a partir de estas secuencias; y

(c) llevar a cabo la etapa (b) con respecto a una segunda molécula de ácido nucleico de molde y/o a una adicional.

En un cuarto aspecto de la invención se proporciona un procedimiento para la generación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo individual que comprende:

a) proporcionar al menos una muestra de las moléculas de ácidos nucleicos que comprende al menos dos moléculas de ácidos nucleicos de molde;

b) introducir una primera etiqueta molecular en un extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo y una segunda etiqueta molecular en el otro extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo para proporcionar al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular;

c) amplificar las al menos dos moléculas de ácidos nucleicos de molde etiquetadas, lo que proporciona múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas;

d) secuenciar las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular y la segunda etiqueta molecular; y

e) reconstruir una secuencia consenso para al menos una de las al menos dos moléculas de ácidos nucleicos de molde objetivo, en el que la etapa e) comprende

(i) identificar los agregados de las secuencias de las regiones de las múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que es probable que se correspondan con la misma molécula de ácido nucleico de molde objetivo mediante la asignación de unas secuencias que comprenden unas primeras secuencias de etiqueta molecular que son homólogas entre sí y unas segundas secuencias de etiqueta molecular que son homólogas entre sí al mismo agregado;

(ii) seleccionar al menos un agregado de secuencias en el que las secuencias de los agregados seleccionados comprenden una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre si que con una primera etiqueta molecular o una segunda etiqueta molecular diferente;

(iii) reconstruir una secuencia consenso de una primera molécula de ácido nucleico de molde objetivo mediante la alineación de las secuencias de las al menos dos moléculas de ácidos nucleicos de molde en el agregado seleccionado en la etapa (ii) y definir una secuencia consenso a partir de estas secuencias; y

(iv) llevar a cabo las etapas (ii) hasta (iii) con respecto a una segunda molécula de ácido nucleico de molde y/o a una adicional.

En un quinto aspecto de la invención se proporciona un procedimiento implementado por ordenador para la determinación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo individual que comprende las siguientes etapas:

(a) obtener los datos que comprenden las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas en los que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprende una primera etiqueta molecular en un extremo y una segunda etiqueta molecular en el otro extremo, en los que cada molécula de ácido nucleico de molde objetivo está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular y en los que las regiones comprenden la primera etiqueta molecular y la segunda etiqueta molecular;

(b) analizar los datos que comprenden las secuencias de las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular y la segunda etiqueta molecular para identificar los agregados de las secuencias que es probable que se correspondan con la misma molécula de ácido nucleico de molde mediante la asignación de unas secuencias que comprenden unas primeras etiquetas moleculares que son homólogas entre sí y unas segundas etiquetas moleculares que son homólogas entre sí al mismo agregado;

(c) seleccionar al menos un agregado de secuencias en el que las secuencias de los agregados seleccionados comprenden una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre si que con una primera etiqueta molecular o una segunda etiqueta molecular diferente;

(d) reconstruir una secuencia consenso de una primera molécula de ácido nucleico de molde mediante la alineación de al menos un subconjunto de las moléculas de las secuencias del agregado seleccionado en la

etapa (c) y definir una secuencia consenso a partir de estas secuencias; y

(e) llevar a cabo las etapas (c) hasta (d) con respecto a una segunda molécula de ácido nucleico de molde y/o a una adicional.

En un sexto aspecto de la invención se proporciona un procedimiento implementado por ordenador para la determinación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo que comprende

(a) obtener los datos que comprenden un agregado de secuencias;

(b) reconstruir una secuencia consenso de una primera molécula de ácido nucleico de molde mediante la alineación de las secuencias de al menos un subconjunto de las secuencias del agregado seleccionado;

en el que las secuencias del agregado seleccionado comprenden las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas, en las que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprende una primera etiqueta molecular en un extremo y una segunda etiqueta molecular en el otro extremo, en las que cada molécula de ácido nucleico de molde objetivo está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular y en las que las regiones comprenden la primera etiqueta molecular y la segunda etiqueta molecular; y cada secuencia del agregado seleccionado

(i) comprende una primera etiqueta molecular que es homóloga de la primera etiqueta molecular de las otras secuencias del mismo, y la segunda etiqueta molecular que es

(ii) comprende una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre si que con una primera etiqueta molecular o una segunda etiqueta molecular diferente.

En un séptimo aspecto de la invención se proporciona un programa informático adaptado para llevar a cabo los procedimientos o las etapas del procedimiento de la invención cuando dicho programa es ejecutado en un dispositivo electrónico.

En un octavo aspecto de la invención se proporciona un medio legible por ordenador que almacena el programa informático de la invención.

Se divulga un kit que comprende:

(i) cebadores que comprenden una porción que comprende una primera etiqueta molecular o una segunda etiqueta molecular y una porción que tiene una secuencia que es capaz de hibridar con al menos dos moléculas de ácidos nucleicos de molde;

(ii) instrucciones que describen como llevar a cabo el procedimiento de la invención.

Se divulga un kit que comprende

(i) cebadores que comprenden una porción que comprende una primera etiqueta molecular o una segunda etiqueta molecular y una porción que tiene una secuencia que es capaz de hibridar con al menos dos moléculas de ácidos nucleicos de molde;

(ii) el medio legible por ordenador que almacena el programa informático de la invención.

Descripción de las figuras

Figura 1. Flujo de trabajo informático completamente automatizado usado para el procesado de las lecturas de un único nanoanálisis MiSeq para secuenciar moldes de longitud completa del ARNr 16S. Se secuenció un conjunto de moléculas que contienen tanto los moldes de longitud completa como los fragmentos "rellenados" con el instrumento, y se proceso informáticamente usando las etapas mostradas.

Figura 2. Descripción de las secuencias de **Figura 3.** Abundancia de los agregados con código de barras identificados como posibles recombinantes (columna izquierda), junto con las abundancias de las moléculas progenitoras que producen las formas recombinantes (derecha). Los moldes parentales son de media 28-35x más abundantes que las posibles formas recombinantes.

Figura 4. Grafica que ilustra la distribución de la longitud de las secuencias ensambladas del 16S.

Figura 5. Grafica que compara las asignaciones taxonómicas a nivel de filo de las OTU usando secuencias largas y cortas. La barra blanca representa el valor medio a lo largo de las 12 muestras para el procedimiento de secuenciación corto. La barra negra representa el valor medio a lo largo de las 12 muestras para el procedimiento "largo". La barra gris representa el valor medio a lo largo de las 12 muestras de la región V4 ensamblada del procedimiento de secuenciación "largo".

Figura 6. Gráfica que compara las asignaciones taxonómicas a nivel de género de las OTUs usando secuencias largas y cortas. La barra blanca representa el valor medio a lo largo de las 12 muestras para el procedimiento de secuenciación corto. La barra negra representa el valor medio a lo largo de las 12 muestras para el procedimiento "largo". La barra gris representa el valor medio a lo largo de las 12 muestras de la región V4 ensamblada del procedimiento de secuenciación "largo".

Figura 7. Curvas de acumulación que muestran el número de códigos de barras aleatorios observados frente al número de moléculas de molde secuenciadas. Se predijo que las diluciones de 50x y de 100x tendrán un nivel apropiado de redundancia en los moldes para permitir la reconstrucción de la molécula de molde de longitud completa mediante una secuenciación de relleno en un Illumina MiSeq. La línea superior representa una dilución de 1 a 10, la segunda línea desde la parte superior representa una dilución de 1 a 50. La tercera línea desde la parte superior representa una dilución de 1 a 100. La cuarta línea desde la parte superior representa una dilución de 1 a 500 y la línea inferior representa una dilución de 1 a 1.000.

Figuras 8-11. Diagramas de flujo que representan los procedimientos de la invención.

Descripción detallada de la invención

Generación o determinación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo individual

La presente invención se refiere a un procedimiento para la generación de, o a un procedimiento implementado por ordenador para la determinación de, las secuencias de al menos una moléculas de ácidos nucleicos de molde objetivo individual.

El termino 'molécula de ácido nucleico de molde objetivo' se refiere a una molécula de ácido nucleico que el operario del procedimiento pretende secuenciar. Una 'molécula de ácido nucleico de molde' puede comprender parte de una molécula de ácido nucleico mayor, tal como un cromosoma. Una 'molécula de ácido nucleico de molde' puede comprender un gen, múltiples genes o un fragmento de un gen. Una 'molécula de ácido nucleico de molde' puede aislarse mediante el uso de cebadores que son capaces de hibridar con la molécula de ácido nucleico de molde.

Hay al menos dos moléculas de ácidos nucleicos de molde objetivo en la muestra de moléculas de ácidos nucleicos. En el caso de la secuenciación del 16S, las al menos dos moléculas de ácidos nucleicos de molde objetivo podrían incluir múltiples moléculas, codificando cada una un ARNr 16S diferente. Por ejemplo, las al menos dos moléculas de ácidos nucleicos de molde objetivo podrían incluir ácidos nucleicos que codifican el ARNr 16S de diferentes bacterias, ácidos nucleicos que codifican diferentes moléculas del ARNr 16s de la misma bacteria, o ambos. Alternativamente, las al menos dos moléculas de ácidos nucleicos de molde objetivo pueden comprender múltiples copias del mismo gen. Las 'moléculas de ácidos nucleicos de molde objetivo' pueden comprender un fragmento del ARNr 16s, sin embargo, es preferente que el fragmento tenga una longitud de al menos 1 Kpb. Esto es debido a que los inventores han demostrado que cuando se usa la secuenciación del 16S para estudios filogenéticos, cuanto más larga sea la hebra del ARNr 16s que se está secuenciando, mayor será el nivel de resolución taxonómica que puede obtenerse.

En una realización de la invención, la al menos una molécula de ácido nucleico de molde objetivo es mayor de 1 Kpb, mayor de 1,2 Kpb, mayor de 1,3 Kpb o mayor de 1,5 Kpb en tamaño. En una realización adicional de la invención, la al menos una molécula de ácido nucleico de molde objetivo es menor de 100 Kpb, menor de 50 Kpb, menor de 25 Kpb, menor de 15 Kpb, menor de 10 Kpb, menor de 5 Kpb, menor de 3 Kpb o menor de 2 Kpb en tamaño.

En una realización adicional de la invención, el procedimiento es un procedimiento de alto rendimiento para la generación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo.

Proporcionar al menos una muestra de ácidos nucleicos

Algunos aspectos de la invención requieren una etapa de proporcionar al menos una muestra de ácidos nucleicos que comprenda al menos dos moléculas de los ácidos nucleicos de molde objetivo. Opcionalmente, las al menos dos moléculas de ácidos nucleicos de molde objetivo son mayores de 1 Kpb en tamaño.

En general el término 'que comprende' pretende indicar que incluye, pero no se limita a, por ejemplo, la expresión 'que comprende las siguientes etapas' indica que el procedimiento incluye esas etapas, pero que también pueden llevarse a cabo etapas adicionales. En algunas realizaciones de la invención, la expresión 'que comprende' puede ser sustituida por la expresión 'que consiste en'. El termino 'que consiste en' pretende ser limitante, por ejemplo, si un procedimiento 'consiste en las siguientes etapas' el procedimiento incluye esas etapas y no otras.

La muestra puede ser cualquier muestra de ácidos nucleicos. La muestra de ácidos nucleicos puede ser una muestra de ácidos nucleicos procedente de un ser humano, por ejemplo, una muestra extraída a partir de un frotis cutáneo de un paciente humano. Alternativamente, la muestra de ácidos nucleicos puede proceder de cualquier otra fuente, tal como una muestra de un suministro de agua. Dicha muestra podría contener billones de moléculas de

ácidos nucleicos de molde. Sería posible secuenciar cada una de esos billones de moléculas de ácidos nucleicos de molde simultáneamente usando el procedimiento de la invención, por lo que no hay ningún límite superior con respecto a las moléculas de ácidos nucleicos de molde que podrían usarse en el procedimiento de la invención.

En una realización adicional de la invención, el procedimiento comprende proporcionar múltiples muestras de ácidos nucleicos, por ejemplo, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 15, 20, 25, 50, 75 o 100 muestras. Opcionalmente, se proporcionan menos de 100, 75, 50, 25, 20, 15, 11, 10, 9, 8, 7, 6, 5 o 4 muestras de ácidos nucleicos. En una realización adicional, se proporcionan entre 2 y 100, entre 2 y 75, entre 2 y 50, entre 2 y 25, entre 5 y 15 o entre 7 y 15 muestras.

Introducción de una primera etiqueta molecular y de una segunda etiqueta molecular y amplificación de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas

Algunos de los procedimientos de la invención implican la introducción de una primera etiqueta molecular en un extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde y una segunda etiqueta molecular en el otro extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde para proporcionar al menos dos moléculas de ácidos nucleicos de molde etiquetadas. Algunos de los procedimientos de la invención implican la amplificación de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas para proporcionar múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde.

Con objeto de permitir la secuenciación de las moléculas de ácidos nucleicos de molde de la muestra, las moléculas de ácidos nucleicos de molde deben ser amplificadas opcionalmente mediante una PCR con objeto de proporcionar múltiples copias de cada molécula de ácido nucleico de molde (es decir, para asegurar que las moléculas de ácidos nucleicos de molde están a una concentración suficiente para la reacción de secuenciación). En una realización, la amplificación se lleva a cabo mediante una reacción en cadena de la polimerasa (PCR). La etapa de amplificación también asegura que las moléculas de ácidos nucleicos de molde objetivo están enriquecidas con respecto a los otros ácidos nucleicos en la muestra. La etapa de amplificación usa unos cebadores que hibridan con las moléculas de ácidos nucleicos de molde objetivo, amplificando así únicamente el ácido nucleico de molde objetivo y aumentando la concentración de las moléculas de ácidos nucleicos de molde objetivo con respecto a los otros ácidos nucleicos de la muestra (enriquecimiento). Sin embargo, dado que las muestras generalmente contienen múltiples moléculas de ácidos nucleicos de molde objetivo, esta etapa de amplificación puede amplificar múltiples moléculas de ácidos nucleicos de molde objetivo. Por ejemplo, en la secuenciación del 16S, una muestra puede contener moldes del ADN 16S de múltiples bacterias, los cebadores usados en la etapa de amplificación pueden hibridar con todas estas secuencias del gen 16S y por lo tanto se amplificaran todos estos moldes de ADN. Esto puede conseguirse mediante el uso de cebadores degenerados, que pueden variar ligeramente en su secuencia de forma que un grupo de cebadores degenerados pueda hibridar con (o sea complementario de) secuencias de ácidos nucleicos de molde objetivo similares, pero no idénticas.

Es ventajoso ser capaces de determinar cuáles de las secuencias generadas en las etapas de secuenciación se originaron a partir de la misma molécula de ácido nucleico de molde original. Consecuentemente, el termino 'molécula de ácido nucleico de molde etiquetada' se refiere a una molécula que comprende una 'molécula de ácido nucleico de molde objetivo' y una etiqueta en cada extremo. Esto permite la determinación de la secuencia consenso para cada molécula original de los ácidos nucleicos de molde. Esto puede conseguirse mediante la adición de etiquetas moleculares en ambos extremos (los extremos 5' y 3') de cada una de las moléculas de ácidos nucleicos de molde originales (Lundberg et al; Nature Methods 10: 999-1002) para producir moléculas de ácidos nucleicos de molde etiquetadas. Se considerara que la primera y/o segunda etiqueta molecular ha sido introducida en los extremos de las moléculas de ADN de molde siempre que estén cerca, en la secuencia, de los nucleótidos terminales (el primer o el ultimo nucleótido de la secuencia) de las moléculas de ADN de molde. En una realización hay menos de 50, 40, 30, 25, 20, 15, 10 o 5 nucleótidos entre un nucleótido terminal y la primera etiqueta molecular. En una realización adicional hay menos de 50, 40, 30, 25, 20, 15, 10 o 5 nucleótidos entre un nucleótido terminal y la segunda etiqueta molecular.

Los procedimientos de la invención requieren que la primera etiqueta molecular y la segunda etiqueta molecular sean únicas. En este caso, el termino 'única' se refiere a unas etiquetas moleculares que comprenden una secuencia aleatoria de pares de bases, asumiendo que se usan las suficientes secuencias de nucleótidos aleatorias, cada primera etiqueta molecular y cada segunda etiqueta molecular tendrán una secuencia diferente con respecto a cualquier otra etiqueta que se genere. Sin embargo, en algunas realizaciones la misma secuencia de etiqueta puede aparecer más de una vez, en esta realización la primera etiqueta molecular y la segunda etiqueta molecular todavía serán consideradas como 'únicas'. En una realización adicional, cada primera etiqueta molecular y cada segunda etiqueta molecular comprenden unas secuencias de nucleótidos que son diferentes de las secuencias de nucleótidos de cualquier otra primera etiqueta molecular y segunda etiqueta molecular. En una realización adicional, al menos el 90 % de las primeras etiquetas moleculares y de las segundas etiquetas moleculares comprenden secuencias de nucleótidos que son diferentes de las secuencias de nucleótidos de cualquier otra primera etiqueta molecular y segunda etiqueta molecular. Esto significa que es probable que las secuencias de moléculas de ácidos nucleicos que comparten el mismo par de primera y segunda etiqueta molecular única se hayan originado en la misma molécula de ácido nucleico de molde original (paradoja del cumpleaños). Además, también es probable que las secuencias de los fragmentos de ácidos nucleicos que comprenden bien la primera etiqueta molecular o bien la

segunda etiqueta molecular asociadas a una molécula de ácido nucleico de molde objetivo se hayan originado a partir de esa molécula de ADN de molde objetivo. El uso de dos etiquetas moleculares únicas también permite identificar y descartar las secuencias que son generadas mediante una recombinación durante los procedimientos de la invención.

La secuencia de la primera etiqueta molecular y de la segunda etiqueta molecular también puede comprender unos pocos nucleótidos de la secuencia del ácido nucleico de molde objetivo, por ejemplo, menos de 50, 40, 35, 30, 25, 20, 15 o 10 pares de bases de la secuencia de la molécula del ácido nucleico de molde objetivo.

En una realización, la primera etiqueta molecular y la segunda etiqueta molecular son mayores de 5 pb, mayores de 6 pb o mayores de 7 pb en tamaño. En una realización adicional, la primera etiqueta molecular y la segunda etiqueta molecular son menores de 20 pb, menores de 18 pb, menores de 15 pb o menores de 10 pb en tamaño.

Dichas etiquetas moleculares únicas pueden ser introducidas usando una diversidad de técnicas que incluyen una PCR, una tagmentación y un cizallamiento físico o una digestión de restricción de los ácidos nucleicos objetivo, combinadas con una posterior ligación de un adaptador (opcionalmente una ligación de extremos adhesivos). Por ejemplo, la PCR puede llevarse a cabo con las al menos dos moléculas de ácidos nucleicos de molde objetivo usando un primer conjunto de cebadores capaz de hibridar con (opcionalmente complementario de) las al menos dos moléculas de ácidos nucleicos de molde objetivo. En una realización de la invención, la primera etiqueta molecular y la segunda etiqueta molecular son introducidas en cada una de las al menos dos moléculas de ácidos nucleicos de molde mediante una PCR que usa unos cebadores que comprenden una porción (una porción del extremo 5') que comprende la primera etiqueta molecular o la segunda etiqueta molecular, y una porción (una porción del extremo 3') que tiene una secuencia que es capaz de hibridar con (opcionalmente complementaria de) las al menos dos moléculas de ácidos nucleicos de molde objetivo. Dichos cebadores hibridaran con una molécula de ácido nucleico de molde objetivo, la extensión con el cebador de la PCR proporcionara entonces una molécula de ácido nucleico que comprende bien la primera etiqueta molecular o bien la segunda etiqueta molecular. Una ronda adicional de la PCR con estos cebadores proporcionara unas moléculas de ácidos nucleicos de molde etiquetadas que comprenden una primera etiqueta molecular en un extremo y una segunda etiqueta molecular en el otro extremo. En una realización adicional, los cebadores están degenerados, es decir, las porciones de los extremos 3' de los cebadores son similares, pero no idénticas entre si Por ejemplo, si el procedimiento de la invención se usa durante la secuenciación del 16S ribosómico, la porción del extremo 3' de los cebadores puede variar ligeramente entre un cebador y otro cebador, pero cada una de las porciones de los extremos 3' será complementaria de la secuencia del 16S en al menos un organismo. Esto permite la secuenciación de una secuencia del 16S cuyo origen es desconocido, permitiendo por lo tanto la secuenciación de cualquier secuencia de ARNr 16S independientemente de su origen (por ejemplo, la bacteria de la que procede). Después, dichas secuencias pueden ser usadas en estudios filogenéticos. En una realización en la que las al menos dos moléculas de ácidos nucleicos de molde objetivo son genes del ARNr 16S, los cebadores adecuados pueden tener una porción del extremo 3' que comprende las secuencias del cebador bacteriano 27F (Weisberg et al, J Bacteriol. Enero de 1991; 173 (2): 697-703) o 1391R (Turner *et al,* 1999).

En una realización individual de la invención, la primera etiqueta molecular y la segunda etiqueta molecular pueden ser introducidas usando una tagmentación. En una realización en la que la primera etiqueta molecular y la segunda etiqueta molecular son introducidas usando una tagmentación, pueden ser introducidas usando una tagmentación directa, o mediante la introducción de una secuencia definida por la tagmentación, seguida de dos rondas de la PCR usando unos cebadores que comprenden una porción capaz de hibridar con la secuencia definida y una porción que comprende la primera etiqueta molecular o la segunda etiqueta molecular. En una realización adicional de la invención, la primera etiqueta molecular y la segunda etiqueta molecular pueden ser introducidas mediante una digestión de restricción de los ácidos nucleicos originales, seguida de una ligación de los ácidos nucleicos que comprenden la primera o la segunda etiqueta molecular. La digestión de restricción de los ácidos nucleicos originales debería llevarse a cabo de tal forma que la digestión de como resultado una molécula que comprenda la región que va a ser secuenciada (la al menos una molécula de ácido nucleico de molde objetivo).

En una realización en la que la primera etiqueta molecular y la segunda etiqueta molecular son introducidas en las al menos dos moléculas de ácidos nucleicos de molde objetivo mediante una PCR, los cebadores usados pueden comprender una porción adicional que comprende una 'secuencia de lengüeta' constante. Esta secuencia de lengüeta constante está preferentemente en 5' de la etiqueta molecular única. En esta realización, las moléculas de ácidos nucleicos de molde etiquetadas proporcionadas comprenderán una secuencia de lengüeta.

En una realización en la que se proporcionan múltiples muestras de ácidos nucleicos, el procedimiento comprende una etapa adicional de introducir un código de barras de la muestra en uno de los extremos de las moléculas de ácidos nucleicos de molde objetivo en cada muestra. Esta etapa adicional tiene lugar antes o durante la etapa de introducción de una primera etiqueta molecular en un extremo y una segunda etiqueta molecular en el otro extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo en los procedimientos de la presente invención. Estos códigos de barras de la muestra pueden ser introducidos de una forma similar a la de la introducción de la primera etiqueta molecular y de la segunda etiqueta molecular, por ejemplo, puede llevarse a cabo una ronda de PCR con cada muestra por separado en la que los cebadores usados hibridan con (o son complementarios de) las al menos dos moléculas de ácidos nucleicos de molde objetivo y comprenden una porción

(opcionalmente una porción en 3') que comprende el código de barras de la muestra. Opcionalmente, en una realización en la que la primera etiqueta molecular y la segunda etiqueta molecular son introducidas en las al menos dos moléculas de ácidos nucleicos de molde mediante una PCR, los cebadores usados para la introducción de las etiquetas pueden comprender una porción adicional que comprende un código de barras específico de la muestra. En esta realización se lleva a cabo una primera ronda de la PCR con cada muestra de los ácidos nucleicos por separado. La primera ronda de la PCR puede usar unos cebadores que comprendan la primera etiqueta molecular o la segunda etiqueta molecular, un código de barras específico de la muestra que es idéntico para todas las moléculas de ácidos nucleicos de molde de la muestra, una región que hibrida con las moléculas de ácidos nucleicos de molde y opcionalmente una región de lengüeta. Las muestras de ácidos nucleicos pueden entonces agruparse y someterse a rondas adicionales de la PCR usando unos cebadores (que opcionalmente son capaces de hibridar con, o son complementarios de, la región de 'lengüeta') que no comprenden un código de barras especifico de la muestra. Opcionalmente, se lleva a cabo una segunda ronda de la PCR usando un cebador que comprende un segundo código de barras específico de la muestra, en esta realización las muestras de los ácidos nucleicos no son agrupadas hasta después de la segunda ronda de la PCR.

La etapa de amplificar los al menos dos ácidos nucleicos de molde etiquetados puede implicar una PCR que usa un segundo conjunto de cebadores que son capaces de hibridar con los extremos de las moléculas de ácidos nucleicos de molde etiquetadas de tal forma que la extensión de los cebadores dará como resultado múltiples copias de las moléculas de ácidos nucleicos de molde etiquetadas y mantendrá la primera etiqueta molecular y la segunda etiqueta molecular. En una realización en la que el primer conjunto de cebadores comprende una secuencia de lengüeta, el segundo conjunto de cebadores puede comprender una región que es capaz de hibridar con la secuencia de lengüeta de las moléculas de ácidos nucleicos de molde etiquetadas.

Aislamiento de una fracción de las moléculas de ácidos nucleicos de molde amplificadas y fragmentación de las moléculas de ácidos nucleicos de molde amplificadas en la fracción

El procedimiento puede comprender el aislamiento de una fracción de las moléculas de ácidos nucleicos de molde amplificadas y la fragmentación de las moléculas de ácidos nucleicos de molde amplificadas en la fracción para proporcionar múltiples moléculas de ácidos nucleicos de molde fragmentadas.

Por el termino 'fragmento' nos estamos refiriendo a un segmento corto de una molécula de ácido nucleico, es decir, a una cadena de nucleótidos que forma parte de una secuencia 'completa'. Los fragmentos según la invención tendrán una longitud de al menos 10, 15, 20, 50, 100, 200, 250 o 500 pares de bases. Opcionalmente, los fragmentos según la invención tendrán una longitud menor de 2.500, 2.200, 2.000 o 1.500 pares de bases.

La fragmentación puede llevarse a cabo usando cualquier procedimiento apropiado. Por ejemplo, la fragmentación puede llevarse a cabo usando una digestión de restricción o usando una PCR con cebadores complementarios de al menos una región interna de las moléculas de ácidos nucleicos de molde etiquetadas. Preferentemente, la fragmentación se lleva a cabo usando un procedimiento que produzca unos fragmentos arbitrarios. El termino "fragmento arbitrario" se refiere a un fragmento generado de forma aleatoria, por ejemplo, un fragmento generado mediante una tagmentación. Los fragmentos generados mediante el uso de enzimas de restricción no son "arbitrarios", ya que la digestión de restricción se produce en unas secuencias específicas del ADN definidas por la enzima de restricción que se usa. Incluso más preferentemente, la fragmentación se lleva a cabo mediante una tagmentación. Si la fragmentación se lleva a cabo mediante una tagmentación, la reacción de tagmentación opcionalmente introduce una región adaptadora en las moléculas de ácidos nucleicos de molde fragmentadas. Esta región adaptadora es una secuencia corta de ADN que puede codificar, por ejemplo, adaptadores que permiten que las moléculas de ácidos nucleicos de molde fragmentadas sean secuenciadas mediante el uso de la tecnología Illumina MiSeq.

En una realización típica, esta etapa puede comprender una etapa adicional de enriquecimiento en las múltiples moléculas de molde fragmentadas para aumentar la proporción de las múltiples moléculas de ácidos nucleicos de molde fragmentadas que comprenden la primera etiqueta molecular o la segunda etiqueta molecular. En esta realización preferente, la etapa de enriquecimiento en las múltiples moléculas de ácidos nucleicos de molde fragmentadas se lleva a cabo preferentemente mediante una PCR. Preferentemente, la PCR se lleva a cabo usando unos cebadores que son capaces de hibridar con (opcionalmente complementarios de) cualquiera de la primera o la segunda etiqueta molecular, y unos cebadores que son capaces de hibridar con (opcionalmente complementarios de) las regiones internas de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas. Dicha etapa de la PCR aumentara la concentración de los fragmentos que comprenden la primera etiqueta molecular o la segunda etiqueta molecular.

En una realización en la que la fragmentación se lleva a cabo mediante una tagmentación, y la tagmentación introduce una región adaptadora en las moléculas de ácidos nucleicos de molde fragmentadas, el enriquecimiento puede llevarse a cabo mediante una PCR usando unos cebadores que sean capaces de hibridar con (opcionalmente complementarios de) cualquiera de la primera o la segunda etiqueta molecular, y unos cebadores que sean capaces de hibridar con (opcionalmente complementarios de) la secuencia adaptadora.

Secuenciación de las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas y/o

secuenciación de las múltiples moléculas de ácidos nucleicos de molde fragmentadas

En general, las etapas de secuenciación pueden llevarse a cabo usando cualquier procedimiento de secuenciación. Algunos ejemplos de posibles procedimientos de secuenciación incluyen una secuenciación de Maxam Gilbert, una secuenciación de Sanger o una secuenciación que comprende una PCR de puente. En una realización típica, las etapas de secuenciación implican una PCR de puente, opcionalmente la etapa de la PCR de puente se lleva a cabo usando un tiempo de extensión mayor de 5, de 10, de 15 o de 20 segundos. Un ejemplo del uso de una PCR de puente es en los secuenciadores analizadores de Illumina Genome.

El procedimiento de la invención puede comprender una etapa de secuenciar las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas. Como se ha descrito anteriormente, el procedimiento de la invención requiere que se introduzca una primera y una segunda etiqueta molecular en las al menos dos moléculas de ácidos nucleicos de molde objetivo, y que cada una de las al menos dos moléculas de ácidos nucleicos de molde este etiquetada con una única etiqueta. Dado que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprende una única etiqueta, entonces, a pesar de que se producen múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde después de la etapa de amplificación, es posible ver que secuencias se corresponden con que molécula de ácido nucleico de molde individual objetivo. Con objeto de conseguir esto, el operario debe ser capaz de determinar la secuencia de la primera y de la segunda etiqueta molecular única asociada con cada molécula de ácido nucleico de molde objetivo original. Esto se consigue mediante la secuenciación de las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas, en las que las regiones comprenden la primera etiqueta molecular y la segunda etiqueta molecular. Esta etapa puede comprender la secuenciación de la longitud total de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas, o normalmente comprende la secuenciación únicamente de los extremos de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas.

El procedimiento de la invención puede comprender una etapa de secuenciación de las múltiples moléculas de ácidos nucleicos de molde fragmentadas. En una realización en la que el procedimiento comprende una etapa de secuenciación de múltiples moléculas de ácidos nucleicos de molde fragmentadas, esto puede llevarse a cabo en el mismo análisis de secuenciación que en el análisis de la secuenciación en el que se secuenciaron las al menos dos moléculas de ácidos nucleicos de molde etiquetadas. Por otro lado, puede ser más eficaz y preciso secuenciar las múltiples moléculas de ácidos nucleicos de molde fragmentadas en un análisis de secuenciación individual de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas.

Reconstrucción de una secuencia consenso para al menos una de las al menos dos moléculas de ácidos nucleicos de molde objetivo

Los procedimientos de la invención pueden comprender una etapa de reconstrucción de una secuencia consenso para al menos una de las al menos dos moléculas de ácidos nucleicos de molde.

Opcionalmente, la etapa de reconstrucción de una secuencia consenso comprende una etapa de identificación de los agregados de las secuencias de múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que es probable que se correspondan con la misma molécula de ácido nucleico de molde mediante la asignación de unas secuencias que comprenden unas primeras secuencias de etiqueta molecular que son homólogas entre sí y unas segundas secuencias de etiqueta molecular que son homólogas entre sí al mismo agregado (por ejemplo, la etapa S2). Para los propósitos de la presente invención, la expresión "homólogas entre sí' requiere que dos secuencias tengan más del 75 %, del 80 %, del 85 %, del 90 %, del 95 %, del 98 %, del 99 % o del 100 % de identidad de la secuencia entre si a lo largo de la longitud total de las secuencias más largas. Por ejemplo, si las secuencias comprenden unas etiquetas moleculares de 10 pb, dos etiquetas moleculares serán idénticas entre si al 90 % si las etiquetas difieren únicamente en un par de bases. Esta diferencia puede ser una sustitución o una deleción de un par de bases. Esto puede ser determinado mediante la alineación de las secuencias de las etiquetas moleculares y su comparación mediante el uso del algoritmo 'uclust' o de cualquier algoritmo de agregación de secuencias similar, tal como el CD-HIT.

Opcionalmente, la etapa de reconstrucción de una secuencia consenso comprende una etapa de análisis de las secuencias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas y/o de las múltiples moléculas de ácidos nucleicos de molde fragmentadas para identificar las secuencias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas y/o de las múltiples moléculas de ácidos nucleicos de molde fragmentadas que comprenden una primera etiqueta molecular o una segunda etiqueta molecular que es homóloga de la primera etiqueta molecular o de la segunda etiqueta molecular de las secuencias de un primer agregado (por ejemplo, la etapa S4 o S7). Esto puede comprender una etapa de determinación de una secuencia consenso para la primera secuencia de etiqueta molecular y la segunda secuencia de etiqueta molecular de un agregado. Como se ha descrito anteriormente, las secuencias serán asignadas al mismo agregado si las primeras secuencias de etiquetas moleculares y las segundas secuencias de etiquetas moleculares son homólogas entre sí. Las secuencias de la primera etiqueta molecular y de la segunda etiqueta molecular pueden ser ligeramente diferentes entre si incluso cuando las secuencias se han originado a partir de la misma molécula de ácido nucleico de molde objetivo individual, debido a errores introducidos en la secuencia durante el procedimiento de la invención. Por lo tanto, puede definirse una secuencia consenso procedente de estas secuencias de la primera etiqueta molecular y de la

segunda etiqueta molecular. Es muy probable que esta secuencia consenso represente la secuencia de la etiqueta, ya que se introdujo en la molécula de ácido nucleico de molde objetivo. Una vez que se ha definido una secuencia consenso para la primera etiqueta molecular y la segunda etiqueta molecular para un agregado, pueden identificarse las secuencias de las múltiples moléculas de ácidos nucleicos de molde fragmentadas que comprenden una primera etiqueta molecular o una segunda etiqueta molecular que es homóloga de una de estas secuencias consenso. Esto proporciona una mayor precisión en la identificación de las múltiples moléculas de ácidos nucleicos de molde fragmentadas que se corresponden con una molécula de ácido nucleico de molde original en particular.

Como se ha descrito anteriormente cada molécula de ácido nucleico de molde etiquetada comprende una primera etiqueta molecular y una segunda etiqueta molecular. Estas moléculas de ácidos nucleicos de molde etiquetadas se copian, y las copias se fragmentan. Cada fragmento tendrá la misma secuencia en forma de una porción de la molécula de ácido nucleico de molde objetivo individual (a pesar de la posibilidad de algún error en la replicación durante las etapas de amplificación de la PCR), y por lo tanto puede considerarse que 'se corresponden' con una porción de la molécula de ácido nucleico de molde individual objetivo original. Una porción de estos fragmentos comprenderá la primera etiqueta molecular o la segunda etiqueta molecular. Una vez secuenciado, por lo tanto, puede identificarse con que molécula de ácido nucleico de molde objetivo individual se corresponde el fragmento.

Opcionalmente, la etapa de reconstruir una secuencia consenso comprende una etapa de reconstruir la secuencia de una primera molécula de ácido nucleico de molde mediante la alineación de al menos un subconjunto de las secuencias de las múltiples moléculas de ácidos nucleicos de molde fragmentadas identificadas como que comprenden una primera etiqueta molecular o una segunda etiqueta molecular homóloga de la primera etiqueta molecular o de la segunda etiqueta molecular de las secuencias del primer agregado, y definir unas secuencias consenso a partir de estas secuencias (por ejemplo, la etapa S4, S6 o S7).

Como se ha descrito anteriormente, la naturaleza de la primera etiqueta molecular o de la segunda etiqueta molecular asociada con cada fragmento permite al operador determinar con que molécula de ácido nucleico de molde original se corresponde el fragmento. Habrá fragmentos múltiples producidos que se correspondan con la misma molécula de ácido nucleico de molde original. Las secuencias de cada uno de estos fragmentos se corresponderán con una región diferente (potencialmente solapante) de la molécula de ácido nucleico de molde. La secuencia del molde puede ser reconstruida mediante la alineación de estos fragmentos y el cálculo de una secuencia consenso a partir de los fragmentos alineados.

El termino 'alineación' se refiere a la disposición de las secuencias de los fragmentos de tal forma que se alineen las áreas de las secuencias que comparten una secuencia común. Esto puede llevarse a cabo usando un soporte lógico tal como Clustal W2, IDBAUD o SOAPdenovo. Una vez que las secuencias están alineadas, puede determinarse la secuencia consenso. Como se ha descrito anteriormente, durante la reacción de secuenciación pueden introducirse mutaciones en las secuencias, sin embargo, estas secuencias mutadas estarán a una concentración menor que las secuencias precisas. Por esta razón se define una 'secuencia consenso'. El termino 'secuencia consenso' puede considerarse que, en el contexto de la presente invención, se refiere a la secuencia más probable para al menos una molécula de ácido nucleico de molde objetivo individual cuando se considera la secuencia de todos los fragmentos correspondientes a esa al menos una molécula de ácido nucleico de molde.

En una realización, cada una de las secuencias de las múltiples moléculas de ácidos nucleicos de molde fragmentadas que fueron identificadas como que comprenden una primera etiqueta molecular o una segunda etiqueta molecular homóloga de la primera etiqueta molecular o de la segunda etiqueta molecular de las secuencias del primer agregado, son alineadas y usadas para definir la secuencia consenso (la secuencia consenso que se define no comprende la primera etiqueta molecular ni la segunda etiqueta molecular). En una realización adicional, al menos un subconjunto, pero no todas las secuencias de la molécula de ácido nucleico de molde identificadas con fragmentos múltiples son alineadas y usadas para definir la secuencia consenso. En una realización adicional, el 90 %, el 92 %, el 95 %, el 98 %, el 99 % o el 100 % de las secuencias de la molécula de ácido nucleico de molde identificadas con fragmentos múltiples son alineadas y usadas para definir la secuencia consenso. En una realización adicional, las secuencias de la al menos una molécula de ácido nucleico de molde etiquetada de longitud completa también son incluidas en la alineación y usadas para definir la secuencia consenso.

Opcionalmente, el procedimiento de la invención comprende la realización de las etapas necesarias para la reconstrucción de una secuencia consenso para una segunda o una molécula de ácido nucleico de molde adicional. Generalmente, esto implicara la repetición de las etapas para un segundo agregado de secuencias que tiene unas primeras etiquetas moleculares que son homólogas entre sí y unas segundas etiquetas moleculares que son homólogas entre sí.

Opcionalmente, estas etapas de reconstrucción de una secuencia consenso para al menos una de las moléculas de ácidos nucleicos de molde objetivo se realizan a través de un ordenador. En un aspecto más de la invención, se proporciona un programa informático capaz de llevar a cabo estas etapas de reconstruir una secuencia consenso para al menos una de las moléculas de ácidos nucleicos de molde objetivo opcionalmente almacenadas en un medio legible por ordenador.

Descarte de las secuencias de los productos de recombinación

En un aspecto de la presente invención se proporciona un procedimiento para la generación de las secuencias que comprende o que comprende adicionalmente la selección de al menos un agregado de secuencias en el que las secuencias de los agregados seleccionados comprenden una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre si (por ejemplo, al menos 2 veces, al menos 5 veces, al menos 8 veces o al menos 10 veces más habitualmente) que con una primera etiqueta molecular o una segunda etiqueta molecular diferente.

Opcionalmente, esta etapa de selección de al menos un agregado consiste en la identificación de los grupos de agregados de las secuencias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en los que las secuencias de los agregados de cada grupo tienen unas primeras etiquetas moleculares que son homólogas entre sí o la identificación de los grupos de agregados de las secuencias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en los que las secuencias de los agregados de cada grupo tienen unas segundas etiquetas moleculares que son homólogas entre sí. Dicho procedimiento puede comprender adicionalmente la selección de un agregado entre un grupo de agregados de secuencias en los que el agregado que se selecciona contiene el mayor número de secuencias; en los que la secuencia de la primera molécula de ácido nucleico de molde es reconstruida a partir de las secuencias del agregado que fue seleccionado. Esto permite la detección de los productos de recombinación. Dicha recombinación puede dar como resultado moléculas de ácidos nucleicos que comprenden una secuencia que se corresponde con una parte de una molécula de ácido nucleico de molde original y una secuencia que se corresponde con la producción de una parte de una molécula de ácido nucleico de molde original diferente. Sin embargo, dichos productos de recombinación pueden ser detectados si se introduce una primera y una segunda etiqueta molecular única en las moléculas de ácidos nucleicos de molde. Si se produce un acontecimiento de recombinación, el par de etiquetas moleculares únicas no será el mismo que cualquiera de los pares de etiquetas moleculares únicas de cualquiera de las moléculas de ácidos nucleicos de molde etiquetadas originales. Esto significa que, aunque se podría esperar la identificación de un único agregado de secuencias en el que todas las secuencias comprenden la misma primera etiqueta molecular o segunda etiqueta molecular, si se ha producido una pequeña cantidad de recombinación, puede haber más de un agregado que tenga la misma primera etiqueta molecular, pero que empareja esta primera etiqueta molecular con al menos dos segundas etiquetas moleculares diferentes. Sin embargo, estos agregados contendrán menos secuencias que el agregado que tiene el mismo par de primera etiqueta molecular y segunda etiqueta molecular que la molécula de ácido nucleico de molde original, ya que tiende a haber presente un menor número de copias de los productos de recombinación que en el ácido nucleico de molde original.

De hecho, es posible usar los procedimientos de la invención para determinar la velocidad a la que se está produciendo la recombinación (o el número de recombinantes que se están generando en un proceso de secuenciación). Por ejemplo, pueden identificarse los agregados que comprenden unas secuencias que tienen una primera etiqueta molecular y una segunda etiqueta molecular que lo más habitualmente están asociadas entre si Es probable que otros agregados que comprenden unas secuencias que tienen la misma primera etiqueta molecular pero una segunda etiqueta molecular diferente o la misma segunda etiqueta molecular pero una primera etiqueta molecular diferente sean el resultado de acontecimientos de recombinación, y estos agregados pueden denominarse agregados del producto de recombinación. Puede cuantificarse la cantidad de secuencias en estos agregados del producto de la recombinación. Puede calcularse la proporción de estas secuencias (que son el resultado de la recombinación) en comparación con el número total de secuencias.

Un procedimiento de la invención puede comprender las etapas de:

a) proporcionar al menos una muestra de moléculas de ácidos nucleicos que comprende al menos dos moléculas de ácidos nucleicos de molde objetivo;

b) introducir una primera etiqueta molecular en un extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo y una segunda etiqueta molecular en el otro extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo para proporcionar al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que cada molécula de ácido nucleico de molde etiquetada está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular;

c) amplificar las al menos dos moléculas de ácidos nucleicos de molde etiquetadas para proporcionar múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas;

d) secuenciar las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular y la segunda etiqueta molecular; y

e) identificar y descartar las secuencias que sean el producto de acontecimientos de recombinación.

La etapa e) puede comprender una etapa de identificación de los agregados de las secuencias de múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que es probable que se correspondan con la misma molécula de ácido nucleico de molde mediante la asignación de unas secuencias que comprenden unas primeras secuencias de etiqueta molecular que son homólogas entre sí y unas segundas secuencias de etiqueta molecular que son homólogas entre sí al mismo agregado. La etapa e) puede comprender adicionalmente la

selección de los agregados de las secuencias en los que las secuencias de los agregados seleccionados comprenden una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre si que con una primera etiqueta molecular o una segunda etiqueta molecular diferente. La etapa e) puede comprender adicionalmente el descarte de cualquier secuencia que no este presente en uno de estos agregados seleccionados.

Opcionalmente, dicho procedimiento comprende adicionalmente una etapa de determinación de una secuencia consenso a partir de uno de los agregados seleccionados. Este procedimiento también puede comprender una etapa f) de determinación de la velocidad a la que se reproduce la recombinación o del porcentaje de la cantidad total de ADN que es el resultado de un acontecimiento de recombinación. Con objeto de llevar a cabo dicha etapa f), se debería determinar el número total de secuencias presente y el número de secuencias que se ha desechado. El porcentaje del ADN total que es el resultado de un acontecimiento de recombinación será igual al número de secuencias que han sido descartadas / el número de secuencias totales x 100. Cuando se genera una secuencia consenso, puede aplicarse la velocidad de recombinación estimada del agregado para eliminar las lecturas que divergen de la mayoría consenso en la que la secuencia divergente se produce a la velocidad esperada para los fragmentos recombinantes. Normalmente puede desecharse una secuencia que se produce a una de las siguientes frecuencias: menor del 30 %, menor del 20 %, menor del 15 %, menor del 12 % o menor del 11 %. La velocidad de recombinación estimada para un agregado puede ser notificada en forma de una métrica de calidad para la secuencia.

Procedimientos implementados por ordenador para la determinación de la secuencia de al menos una molécula de ácido nucleico de molde

La invención también proporciona procedimientos implementados por ordenador para la determinación de las secuencias de al menos dos moléculas de ácidos nucleicos de molde.

En dicho procedimiento se obtienen datos/entradas (S1, S3 o S5), por ejemplo, los datos que comprenden las secuencias de al menos dos moléculas de ácidos nucleicos de molde y/o los datos que comprenden las secuencias de las regiones de las al menos dos moléculas de ácidos nucleicos de molde que comprenden la primera etiqueta molecular y la segunda etiqueta molecular, pueden obtenerse usando las etapas del procedimiento descrito anteriormente.

Este procedimiento es llevado a cabo por un ordenador. En un aspecto más, se proporciona un programa informático adaptado para llevar a cabo los procedimientos de la invención cuando el programa es ejecutado en un dispositivo electrónico. En un aspecto más, se proporciona un medio legible por ordenador que almacena el programa informático de la invención.

Como se ha analizado, los aspectos de los procedimientos analizados en el presente documento, incluyendo los procedimientos ilustrados en las Figuras 8 hasta 11, son implementados mediante un ordenador. Es bien conocido que un ordenador individual puede comprender los elementos de soporte físico convencionales tales como CPU, RAM, dispositivos de almacenamiento, etc. También es bien conocido que pueden conectarse varios ordenadores entre si y pueden cooperar de forma que ejecuten tareas de computación colectivamente (en forma de un sistema de procesado distribuido). Por lo tanto, se apreciara que las referencias a los procedimientos implementados por ordenador pretenden incluir, pero no se limitan a, los procedimientos que usan un sistema de procesamiento de datos (ordenador) que puede realizar una o más de las etapas descritas independientemente, o un sistema de procesamiento distribuido. Un PC de sobremesa que trabaje con un sistema de computación en la nube a través de una conexión a Internet es un ejemplo de un sistema de procesamiento distribuido. Haciendo referencia a la Figura 8, por ejemplo, los datos que se van a introducir en las etapas S1 y S3 podrían estar almacenados en un servidor central en un sistema de computación en la nube (esto puede denominarse sistema de almacenamiento en la nube) y al que accede un ordenador de sobremesa que está configurado para llevar a cabo las etapas de análisis S2, S4 y S5. Alternativamente, los datos que se van a introducir en las etapas S1 y S3 podrían ser proporcionados por el ordenador de sobremesa, y el sistema de computación en la nube podría estar configurado para llevar a cabo las etapas de análisis S2, S4 y S5 y devolver los resultados al ordenador de sobremesa. Se apreciara que podría adoptarse cualquier otra distribución de almacenamiento de los datos y de las tareas de procesamiento de los datos entre diferentes ordenadores, según las necesidades de una aplicación en particular.

Desarrollos adicionales

Los procedimientos de la invención pueden ser modificados para secuencias incluso más largas. Por ejemplo, en un procedimiento que comprende la fragmentación de las moléculas de ácidos nucleicos de molde, puede llevarse a cabo una etapa adicional de introducción de etiquetas moleculares adicionales (por ejemplo, una tercera y una cuarta etiqueta molecular) en las moléculas de ácidos nucleicos de molde fragmentadas. Esto permite que las moléculas de ácidos nucleicos de molde fragmentadas sean adicionalmente fragmentadas, y que las moléculas de ácidos nucleicos de molde adicionalmente fragmentadas sean secuenciadas. El uso de una tercera etiqueta molecular y de una cuarta etiqueta molecular permite la reconstrucción de la secuencia de la secuencia de longitud completa de las moléculas de ácidos nucleicos de molde adicionalmente fragmentadas.

Los procedimientos de la invención pueden usarse para la secuenciación de múltiples genes diferentes de una muestra de ácidos nucleicos. Por ejemplo, el procedimiento de la invención podría usarse para la secuenciación de la totalidad o de una gran proporción del genoma de un organismo de interés, tal como un patógeno médicamente relevante, mediante el uso de un conjunto de cebadores capaces de hibridar con ácidos nucleicos que comprenden múltiples genes. En una realización, estos cebadores son fijados sobre la superficie de un soporte sólido o acoplados a un marcador seleccionable tal como la biotina.

Kits

Se divulgan kits. Opcionalmente, estos kits comprenden uno o más de los siguientes:

(i) cebadores que comprenden una porción que comprende una primera etiqueta molecular o una segunda etiqueta molecular y una porción que tiene una secuencia que es capaz de hibridar con una molécula de ácido nucleico de molde objetivo; opcionalmente en los que los cebadores comprenden una 'región de lengüeta';

(ii) cebadores que comprenden una porción capaz de hibridar con los cebadores de (i), por ejemplo, cebadores que comprenden una región complementaria de la 'región de lengüeta';

(iii) un componente capaz de fragmentar una molécula de ácido nucleico de molde objetivo, por ejemplo, una transposasa, enzimas de restricción o cebadores adicionales que son complementarios de las regiones internas de la molécula de ácido nucleico de molde objetivo;

(iv) cebadores que comprenden una porción capaz de hibridar con una molécula de ácido nucleico de molde objetivo fragmentada;

(v) reactivos para llevar a cabo la amplificación, por ejemplo, mediante una reacción en cadena de la polimerasa;

(vi) instrucciones que describen como llevar a cabo los procedimientos de la invención; y/o

(vii) un medio legible por ordenador que almacena un programa informático de la invención.

Ejemplo 1

Extracción de ADN microbiano a partir de la piel del pie

Se extrajo el ADN a partir de frotis cutáneos tomados del pie de 6 individuos sanos diferentes. En total se recogieron 12 muestras. Los frotis cutáneos se recogieron mediante un frotis del área del tercio anterior del pie o del talón del pie izquierdo o derecho con un hisopo de rayón humedecido con una solución de NaCl 0,15 M y Tween 20 al 0,1 %. El hisopo se frotó firmemente sobre la piel durante aproximadamente 30 segundos. Las cabezas de los hisopos se cortaron en tubos batidores de microesferas, y se extrajo el ADN de los hisopos usando el kit de aislamiento de ADN BiOstic Bacteriemia (Mo-Bio), según las instrucciones de los fabricantes. El ADN fue cuantificado con un Qubit con un ensayo de ADNbc HS (Life Technologies).

Ejemplo 2

Preparación de colecciones de 16S de lectura corta para la secuenciación con Illumina

Se preparó una colección de la región V4 del gen 16S para una secuenciación mediante Illumina a partir de las muestras de ADN microbiano de la piel del pie usando un procedimiento publicado previamente (Caporaso et al, 2012, ISME 6 (8)). En resumen, las muestras se amplificaron usando unos cebadores basados en el diseño de Caporaso, que fueron modificados para que incluyeran unos códigos de barras de la muestra de 8 pb en lugar de 12 pb, e incluían un código de barras tanto en el cebador directo como en el inverso (las secuencias del cebador se describen en Figura 2). La región V4 se amplificó a partir de 500 pg de ADN de molde usando 10 ciclos de PCR con los cebadores de Caporaso modificados (Caporaso_forward y Caporaso reverse), usando cebadores con diferentes códigos de barras para cada muestra. Después de la eliminación del exceso de cebador a través de una limpieza con microesferas magnéticas (Agencourt), las muestras se agruparon y se sometieron a 20 ciclos adicionales de PCR para un enriquecimiento en los amplicones que contienen los adaptadores de Illumina, usando los cebadores Illumina_E_1 e Illumina_E_2 (véase la Figura 2 para los detalles de los cebadores). Las PCR se llevaron a cabo con un kit de PCR Taq core (Qiagen), en las condiciones descritas en Caporaso et al, (2012, ISME 6 (8)). Los amplicones se secuenciaron usando una celda de nanoflujo y un kit de 500 ciclos de V2 en un Illumina MiSeq, siguiendo el procedimiento descrito en Caporaso et al (2012, ISME 6 (8)). En lo sucesivo este procedimiento se denominará "secuenciación corta" y los datos producidos con este procedimiento, datos "V4".

Ejemplo 3

Preparación de colecciones de 16S de longitud completa para la secuenciación con Illumina con etiquetas moleculares únicas

Los cebadores para la amplificación del gen 16S contenían las secuencias del cebador bacteriano 27F (Weisberg et al, J Bacteriol. Enero de 1991;173 (2): 697-703) o 1391R (Turner et al, Journal of Eukaryotic Microbiology, 1999, 46: 327-338), una secuencia de código de barras de 8 pb, una etiqueta molecular única de 10 pb y las secuencias adaptadoras parciales de Illumina PE. Las secuencias del cebador (Long_forward y Long reverse) se muestran en la Figura 2. El uso de una etiqueta molecular única de 10 pb en ambos cebadores directo e inverso (10 billones de posibles etiquetas únicas en cada extremo) nos permitió etiquetar de forma única cada molécula de 16S de nuestro conjunto, usando un procedimiento similar al de Lundberg et al (Nature Methods, 2013, 10: 999-1002). El ADN de molde se sometió a un ciclo de la PCR con el cebador directo, seguido de una limpieza con microesferas para eliminar el exceso de cebador, después otro ciclo de la PCR con el cebador inverso, seguido de otra limpieza con microesferas. La primera PCR lleva a cabo la extensión del gen 16S a partir del cebador directo, que introduce unas etiquetas moleculares únicas en cada molécula del molde de 16S diferente en la reacción. La segunda PCR usa los productos de la extensión de la primera PCR como molde, y produce moléculas con unas etiquetas moleculares únicas en ambos extremos. Aunque las moléculas de 16S originales también pueden actuar como molde en la segunda reacción de la PCR, estos productos únicamente contendrán una secuencia adaptadora parcial de Illumina PE en un extremo, y por lo tanto no serán amplificados en la PCR de enriquecimiento. La PCR de enriquecimiento (34 ciclos) amplifica el conjunto de moléculas 16S etiquetadas usando unos cebadores que son complementarios de las secuencias adaptadoras parciales de Illumina PE en los extremos de cada molécula de 16S etiquetada (los cebadores de Illumina PE_1 y PE_2, Figura 2).

Las PCR se llevaron a cabo usando el kit Taq PCR core (Qiagen). Las reacciones eran de 50 µl y contenían aproximadamente 500 pg de ADN de molde, 0,25 µM de cebador F, 250 µM de dNTP, 1 x de tampón de PCR, 1 x de solución Q y 1,25 U de polimerasa Taq. Las condiciones del ciclo de la PCR eran a 95 °C durante 1 minuto, a 50 °C durante 2 minutos, después a 72 °C durante 3 minutos. Esto permite la extensión del gen 16S a partir del cebador directo, que introduce unas etiquetas moleculares únicas en cada molécula de 16S en la reacción. Las reacciones de la PCR se sometieron a continuación a una limpieza con microesferas magnéticas usando unas microesferas Agencourt SPRI como sigue. Las reacciones de PCR se mezclaron con 0,6 volúmenes de microesferas mediante un pipeteado, y se incubaron a la temperatura ambiente durante 1 minuto. Los tubos se colocaron en una rejilla magnética durante 3 minutos para permitir que las microesferas se concentraran en el lateral del tubo, y el sobrenadante se eliminó. Las microesferas se lavaron con 200 µl de etanol al 85 % durante 30 segundos, tras lo cual se eliminó el etanol y las microesferas se dejaron secar al aire durante 5 minutos. Una vez secos, los tubos se retiraron de la rejilla magnética y las microesferas se resuspendieron en 35 µl de agua exenta de nucleasa mediante un pipeteado. Después de una incubación de un minuto a la temperatura ambiente, los tubos se colocaron de nuevo en la rejilla magnética durante 3 minutos, seguido de la extracción de la solución que contiene el ADN a un tubo nuevo. La segunda PCR se preparó como se ha descrito anteriormente, excepto por que se usaron 0,25 µM del cebador inverso, y el molde eran 31 µl de la reacción de la PCR de la primera ronda limpiada con las microesferas. El ciclo de PCR aplicado fue a 95 °C durante 1 minuto, a 50 °C durante 2 minutos y a 72 °C durante 3 minutos. Durante esta segunda PCR, los productos de la extensión etiquetados de forma única procedentes de la primera PCR actúan como molde, para producir moléculas de 16S con unas etiquetas moleculares únicas en ambos extremos. La segunda PCR estuvo seguida por otra limpieza con microesferas magnéticas, como se ha descrito anteriormente, y el resultado de esta etapa se usó como molde para la reacción de la PCR final. La reacción de la PCR final se preparó en un volumen de 50 µl y contenía 0,5 µM de ambos cebadores PE_1 y PE_2 (véase la Figura 2), 250 µM de dNTP, 1 x de tampón de PCR, 1 x de solución Q, 31 µl de molde (procedente de la segunda limpieza con microesferas) y 1,25 U de polimerasa Taq. Las condiciones de los ciclos de la PCR eran a 95 °C durante 2 minutos, seguido de 34 ciclos a 95 °C durante un minuto, a 58 °C durante 30 segundos y a 72 °C durante 2 minutos. Esto fue seguido por una extensión final a 72 °C durante 5 minutos. Las PCR se sometieron de nuevo a una limpieza con microesferas como se ha descrito anteriormente, antes de ser analizadas usando un chip de ADN de alta sensibilidad con un Bioanalizador (Agilent).

Ejemplo 4

Tagmentación de los productos etiquetados de longitud completa de la PCR del 16S

Los amplicones etiquetados de forma única de longitud completa de la PCR del 16S fueron sometidos a una tagmentación. El procedimiento de tagmentación utiliza una transposasa para fragmentar el ADN mientras añade simultáneamente una secuencia adaptadora para su uso con la plataforma Illumina. La tagmentación se llevó a cabo usando el kit Nextera-XT según las instrucciones de los fabricantes, con la excepción de la etapa de amplificación mediante la PCR. Aquí llevamos a cabo dos PCR por reacción de tagmentación, cada una con una combinación de uno de los cebadores de PCR de Illumina proporcionados con uno de los cebadores de la anterior PCR de extensión, de forma que se amplificaran únicamente aquellos fragmentos de interés. Aspiramos a producir un conjunto de fragmentos de ADN bien con la secuencia PE_1 (el extremo 5' de la secuencia codificante de los amplicones del 16S) o bien con la PE_2 (el extremo 3' de la secuencia codificante de los amplicones del 16S) en un extremo, y los adaptadores i7 o i5 de Illumina (añadidos durante la reacción de tagmentación) en el otro extremo, respectivamente (Figura 2). Esto proporciono un conjunto de fragmentos de todo el gen 16S, que junto con los amplicones de longitud completa del 16S, pueden ser secuenciados a partir de cualquier extremo con el MiSeq. Las

secuencias procedentes de la misma molécula de molde pueden ser identificadas a través de las etiquetas moleculares únicas en cualquier extremo de la molécula y reensambladas para proporcionar las secuencias del 16S de longitud completa. Los productos de la PCR de la reacción de tagmentación se limpiaron inicialmente usando 1,8 V de microesferas Ampure SPRI según las instrucciones del fabricante, y en las posteriores reacciones de tagmentación, usando 0,6 V de las microesferas para retirar los fragmentos menores de 400 pb.

Ejemplo 5

Secuenciación de los amplicones de longitud completa y tagmentados de 16S con el Illumina MiSeq

La molaridad tanto de los amplicones de longitud completa de 16S etiquetados como la de los productos de la tagmentación se midió a través de un chip de ADN Bioanalizador de alta sensibilidad. Durante el primer análisis de la secuenciación, únicamente se cargaron los productos de la tagmentación (limpiados con 1,8 V de microesferas Ampure SPRI) a una concentración media de 1,5 pM y se secuenciaron con un kit de reactivos MiSeq v2 con 2 x de lecturas de extremos apareados de 150 pb, en una celda de nanoflujo. Para el segundo análisis de secuenciación, los amplicones de longitud completa del 16S etiquetados se combinaron con los productos de la tagmentación (limpiados con 0,6 V de microesferas Ampure SPRI para retirar los fragmentos < 400 pb) en una proporción de 1:9. El conjunto de la muestra se cargó a una molaridad media de 6 pM, y se secuenció con un kit de reactivos MiSeq v2 con 2 x de lecturas de extremos apareados de 250 pb, en una celda de nanoflujo.

Cuando se analizaron los amplicones de longitud completa del 16S etiquetados, se crearon modificaciones en las condiciones del análisis del MiSeq. El archivo Chemistry.xml de la carpeta de recetas del Illumina MiSeq contiene el protocolo usado por el instrumento para la agregación y la secuenciación de los fragmentos de ADN. Ese archivo Chemistry.xml correspondiente a los kits de secuenciación del Illumina Versión 2 fue modificado para aumentar la "WaitDuration" en las etapas de "Amplificación 1", la " Resíntesis " y la "Primera extensión" a 15 segundos. Esto dio como resultado un proceso que permitió la secuenciación de los extremos de los amplicones de longitud completa del 16S etiquetados individuales.

Ejemplo 6

Reconstrucción de las secuencias de longitud completa del 16S de las lecturas etiquetadas del Illumina

La secuenciación produce datos a partir de dos tipos de fragmentos, aquellos que abarcan la totalidad del gen 16S *(fragmentos end+end)* y aquellos que emparejan un extremo del gen 16S con una región en la zona intermedia del gen 16S *(fragmentos end+internal).* Las secuencias de los fragmentos *end+end* codifican el emparejamiento de códigos de barras aleatorios y de códigos de barras de muestras.

Para asignar las secuencias a las muestras, se empareja la reacción del código de barras de la muestra de 8 nt frente a la colección de códigos de barras de muestra conocidos, con una tolerancia de hasta un malapareamiento. Debido a que las regiones internas de la secuencia del 16S podrían emparejarse con un código de barras de una muestra, todas las lecturas de secuencias con un potencial emparejamiento de código de barras de la muestra son cribadas después para evaluar la presencia de la secuencia de hibridación del cebador del 16S proximal o distal secuencia abajo del código de barras de la muestra. Se asume que las lecturas que carecen de un código de barras de muestra conocido o de la secuencia de hibridación del cebador en un extremo, proceden de un fragmento *end+internal*.

Ejemplo 7

Etiquetas consenso moleculares únicas y eliminación de los recombinantes

Debido a errores en la secuenciación, las lecturas derivadas de la misma molécula de molde pueden tener unas secuencias de las etiquetas moleculares únicas ligeramente diferentes en 10 nt. Para estimar las secuencias originales de códigos de barra aleatorios de 10 nt de las moléculas de molde etiquetadas, aplicamos el algoritmo uclust (Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26 (19), 2460-2461; Edgar, R. C. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads, Nature methods) para identificar los agregados con secuencias de códigos de barra aleatorios emparejadas con una identidad > 89 % (por ejemplo, se permite el malapareamiento de 1 de las 10 bases), y se notificaron las secuencias consenso de estos agregados. En primer lugar, identificamos los agregados de los códigos de barras aleatorios en los fragmentos *end+end*. A continuación, identificamos el agregado con la mayor abundancia con cada código de barras aleatorio de 10 nt y desechamos cualquier agregado que contuviera un código de barras aleatorio de 10 nt que se encontrase en un agregado diferente más abundante. Esta etapa aspira a identificar y desechar las combinaciones de los códigos de barras aleatorios que surgieron debido a la recombinación *in vitro*. Es probable que las formas recombinantes estén en una abundancia menor que los moldes parentales (Figura 3). Apreciamos que cuando se secuencian fragmentos arbitrarios de 2 Kpb, no se espera que dicha recombinación *in vitro* se produzca muy frecuentemente debido a la diversidad del conjunto de moléculas de molde. La detección de la recombinación es muy importante para la aplicación a los protocolos de secuenciación del amplicón, tal como para el 16S.

Los fragmentos *end+end* pueden no capturar todos los códigos de barras aleatorios presentes en una muestra. El

resto de los códigos de barras aleatorios todavía podría usarse para la reconstrucción de las secuencias del 16S a pesar de que no pueden ser asignados a una muestra sin la información del fragmento *end+end*. Por lo tanto, aplicamos de nuevo el uclust para identificar los agregados de los códigos de barras aleatorios en cada extremo por separado, y añadir cualquier nueva secuencia consenso que no se hubiera encontrado previamente en un fragmento *end+end*.

Finalmente, los códigos de barras aleatorios de la totalidad del conjunto de lecturas son emparejados frente a la colección de secuencias consenso, y las lecturas son agrupadas en agregados para un ensamblaje posterior.

Ejemplo 8

Ensamblaje de los agregados de lectura

Los agregados de lectura contienen lecturas que, con una elevada probabilidad, proceden de la misma molécula de molde. Aplicamos un algoritmo de ensamblaje *de novo* a la lectura del agregado para la reconstrucción de tanta molécula de molde original como fuera posible. Las lecturas son ensambladas usando el desarrollo A5-miseq (Tritt et al (2012). An integrated pipeline for de Novo assembly of Microbial Genomes, PLoS One). El A5-miseq es una revisión del desarrollo original A5, que la extiende para que soporte el ensamblaje de lecturas de hasta 500 nt de longitud y para extraer la secuencia adaptadora de las lecturas, en lugar de desechar las lecturas que contienen la secuencia adaptadora.

En lo sucesivo, este procedimiento se denominará "secuenciación larga" y los datos producidos con este procedimiento, datos "largos".

Ejemplo 9

Análisis de las lecturas del 16S

Se secuenciaron 12 muestras de pie con el protocolo de longitud completa, 6 de las cuales se secuenciaron dos veces con el procedimiento. La totalidad de las 12 muestras también se secuenció usando el procedimiento de Caparoso *et al,* 2012.

Tanto las lecturas de la V4 como las largas fueron analizadas usando el paquete informático QIIME (Caparoso et al (2010), QIME allows analysis of high-throughput community sequence data, Nature Methods 7: 335-335). Se filtró la calidad de las lecturas de la V4 mediante la eliminación de las lecturas menores de 248 o mayores de 253 pb. Para comparar, se extrajo la correspondiente región V4 a partir del conjunto de datos largos, y únicamente aquellas secuencias ensambladas que incluían la región V4 fueron incluidas en el análisis secuencia abajo. Estas secuencias extraídas se denominarán en lo sucesivo "long-V4". Todas las secuencias se agregaron en OTU usando el procedimiento de recolección de referencia cerrada, que asigna secuencias a las OTU pre-agregadas a partir de una base de datos exenta de quimeras (Greengenes). La taxonomía fue evaluada en base a la pertenencia a la base de datos de las OUT pre-agregadas.

Secuenciación corta

Se generó un total de 296.864 secuencias de la V4 con los extremos emparejados a partir de 12 muestras de pie,y un control positivo (únicamente ADN de *Escherichia coli*) y negativo (únicamente el hisopo). De estas secuencias, 11.240 no pudieron ser asignadas a una muestra debido a combinaciones incorrectas de los códigos de barras directo e inverso, lo que indica una tasa de recombinación de al menos el 3,8 %. Se cartografiaron 240.938 secuencias con las 12 muestras de pie, lo que se redujo a 240.426 después del filtrado de calidad (véase la siguiente Tabla 1 para el número de secuencias asignadas a cada muestra). Las OTU agregadas con el procedimiento de referencia cerrada en el QIIME dieron como resultado 1.177 OTU con una similitud del 97 % que contienen 2 o más secuencias. La distribución taxonómica de estas OTU era similar a la que se había notificado previamente para las comunidades cutáneas, dominadas por *Firmicutes* (79,6 % ± 25,7), Actinobacterias (9,3 % ± 12,9) y Proteobacterias (9,9 % ± 22,2).

Tabla 1: número de secuencias analizadas por muestra para los diferentes procedimientos de secuenciación

	Número de secuencias después del filtrado de calidad		
Muestra	V4	Long	Long V4
F1.B1	29.853	69	69
F1.H	10.241	37	37
F2.B2	6.501	30	30
F2.H	5.560	80	80
F3.B2	5.258	4	4

F3.H	38.108	85	85
F4.LB	5.647	32	32
F4.LH	3.266	24	24
F5.LB	13.931	505	505
F5.LH	66.398	836	836
F6.LB	33.714	431	431
F6.LH	21.949	218	218
Total	**240.426**	**2.351**	**2.351**

Secuencias largas

Se ensamblaron 3.914 secuencias del 16S, siendo 2.030 de estas mayores de 1.000 pb (Figura 4). Se asignaron 2.957 secuencias a muestras de pie, mientras que 957 secuencias no pudieron ser asignadas a una muestra debido a combinaciones incorrectas de la etiqueta molecular. únicamente se usaron las lecturas que contenían una región V4 que se correspondían con las secuenciadas con el procedimiento de secuenciación para el análisis secuencia abajo, y se filtró la calidad de estas secuencias en el QIIME mediante la eliminación de las secuencias menores de 700 pb y mayores de 1.500 pb. Esto dio como resultado 2.351 secuencias que se usaron para el análisis (véase la Tabla 1 para los detalles de cuantas secuencias se asignaron a cada muestra).

Las lecturas largas (2.351 usadas para el análisis) se agregaron en 72 OTU, mientras que las secuencias V4-long (correspondientes a la misma región que el conjunto de datos de la V4) se agregaron en 48 OTU. Estas OTU mostraron la misma amplia distribución taxonómica que los datos de la secuencia de la V4 (Figura 5). Aunque había un pequeño aumento en la representación de Actinobacterias (13,6 % ± 21,6) y de Proteobacterias (11,4 ± 26,7), estas diferencias no eran significativas (prueba de la t bilateral, $p > 0,05$).

También se observaron unas asignaciones taxonómicas similares a nivel del genero (Figura 6), con unas comunidades dominadas por *Staphylococcus*, seguidas por los géneros *Corynebacterium, Enhidrobacter* y *Acinetobacter.* El genero *Corynebacterium* tema un aumento en la representación en el conjunto de datos largos en comparación con el procedimiento de secuenciación corto, lo que probablemente representa la diferencia observada en la representación del filo *Actinobacteria*, pero como antes, esta diferencia no era significativa (prueba de la t bilateral, $p > 0,05$). La comparación de las muestras individuales entre los procedimientos de secuenciación corto y largo demostró que las Corynebacterias no estaban coherentemente sobrerrepresentadas en el conjunto de datos ensamblado, y que la media estaba fuertemente influenciada por una muestra en la que *Corynebacterium* representaba únicamente el 0,03 % de las secuencias en la muestra de la V4, pero el 46,67 % de las secuencias en los datos de la secuenciación larga ensamblados (muestra F2_B2).

Velocidades de recombinación

Comparación a nivel de la OTU

Las secuencias del 16S ensambladas (con unas longitudes que varían entre 756 y 1.375) fueron agregadas en OTU usando el procedimiento de referencia cerrada en el QIIME, y de media compartían únicamente el 30,1 % (± 6,8) de las OTU con los datos emparejados de la muestra de la V4 que fueron agregados de la misma forma. Esto puede ser debido a la comparación de conjuntos de datos de diferentes longitudes y a la forma en que las OTU son agregadas en el QIIME. Las secuencias son asignadas a una OTU según la mejor coincidencia frente a las bases de datos de secuencias, que han sido pre-agregadas en OTU con una similitud del 97 %. Supuestamente, se usaron las secuencias de longitud completa de las bases de datos para agregar las OTU, y los agregados con una similitud del 97 % a lo largo de la totalidad del gen 16S pueden no ser similares al 97 % únicamente en la región V4, dado que las diferentes regiones del gen 16S evolucionan a unas tasas diferentes (Schloss PD (2010) The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of rRNA 16S Gene-Based Studies. Plos Computational Biology 6). Por lo tanto, analizamos las OTU agregadas a partir de la región V4 únicamente para las secuencias largas (las secuencias long-V4). En este caso, el 92,2 % (± 12,1) de las OTU estaban compartidas con las OTU de muestra de Caporaso emparejadas (Tabla 3). Aunque se obtuvo una menor cobertura de la secuenciación en el conjunto de datos largos, y posteriormente muchas menos OTU en general, esto demuestra que los datos que se obtuvieron son ampliamente coincidentes con los obtenidos usando las secuencias cortas de la V4. De forma interesante, las secuencias largas se agregaron en un ~ 50 % más de OTU que las secuencias long-V4, lo que demuestra la clasificación más sensible que se puede conseguir con más información de la secuencia por molécula de 16S.

Estos datos indican que este procedimiento recién desarrollado proporciona unos perfiles de comunidad ampliamente coincidentes con respecto a la taxonomía y a la agregación en OTU, y permite una asignación taxonómica más sensible.

Ejemplo 10

Secuenciación de fragmentos largos de *E. coli* K12 MG1655

Se tagmentó el ADN genómico de E. coli K12 MG1655 y los fragmentos de 1,5 - 3 kpb se seleccionaron por tamaños usando una electroforesis en gel de agarosa. Se aplicó un etiquetado molecular a estos fragmentos a través de 2 ciclos de PCR con códigos de barras aleatorios. La secuenciación inicial del conjunto reveló un exceso de diversidad entre las moléculas de molde, de forma que la reconstrucción de los moldes de longitud completa sería inviable. Se usó una serie de diluciones para determinar el grado apropiado hasta el que debería estrecharse la población de moléculas de molde para una secuenciación y reconstrucción con éxito de los moldes de longitud completa (Figura 7). Se secuenciaron ambas diluciones a 50x y a 100x con lecturas rellenadas.

REIVINDICACIONES

1. Un procedimiento para la generación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo individual que comprende:

a) proporcionar al menos una muestra de moléculas de ácidos nucleicos que comprende al menos dos moléculas de ácidos nucleicos de molde objetivo;

b) introducir una primera etiqueta molecular en un extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo y una segunda etiqueta molecular en el otro extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo para proporcionar al menos dos moléculas de ácidos nucleicos de molde etiquetadas, en las que cada molécula de ácido nucleico de molde etiquetada está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular;

c) amplificar las al menos dos moléculas de ácidos nucleicos de molde etiquetadas para proporcionar múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas;

d) secuenciar las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular y la segunda etiqueta molecular; y

e) reconstruir una secuencia consenso para al menos una de las al menos dos moléculas de ácidos nucleicos de molde objetivo en el que la etapa e) comprende

(i) identificar los agregados de las secuencias de las regiones de las múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que es probable que se correspondan con la misma molécula de ácido nucleico de molde objetivo mediante la asignación de unas secuencias que comprenden unas primeras secuencias de etiqueta molecular que son homólogas entre sí y unas segundas secuencias de etiqueta molecular que son homólogas entre sí al mismo agregado;

(ii) seleccionar al menos un agregado de secuencias en el que las secuencias de los agregados seleccionados comprenden una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre si que con una primera etiqueta molecular o una segunda etiqueta molecular diferente;

(iii) reconstruir una secuencia consenso de una primera molécula de ácido nucleico de molde objetivo mediante la alineación de las secuencias de las al menos dos moléculas de ácidos nucleicos de molde en el agregado seleccionado en la etapa (ii) y definir una secuencia consenso a partir de estas secuencias; y

(iv) llevar a cabo las etapas (ii) hasta (iii) con respecto a una segunda molécula de ácido nucleico de molde y/o a una adicional.

2. Un procedimiento para la generación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo individual que es mayor de 1 Kpb en tamaño que comprende:

a) proporcionar al menos una muestra de moléculas de ácidos nucleicos que comprende al menos dos moléculas de ácidos nucleicos de molde objetivo que son mayores de 1 Kpb en tamaño;

b) introducir una primera etiqueta molecular en un extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo y una segunda etiqueta molecular en el otro extremo de cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo para proporcionar al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular;

c) amplificar las al menos dos moléculas de ácidos nucleicos de molde etiquetadas para proporcionar múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas;

d) aislar una fracción de las múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas y fragmentar las moléculas de ácidos nucleicos de molde etiquetadas de la fracción para proporcionar múltiples moléculas de ácidos nucleicos de molde fragmentadas;

e) secuenciar las regiones de las múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular y la segunda etiqueta molecular;

f) secuenciar las múltiples moléculas de ácidos nucleicos de molde fragmentadas; y

g) reconstruir una secuencia consenso para al menos una de las al menos dos moléculas de ácidos nucleicos de molde objetivo de las secuencias que comprenden al menos un subconjunto de las secuencias producidas en

la etapa f), en donde la etapa g) comprende:

(i) identificar los agregados de la secuencia de las regiones de múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que es probable que se correspondan con la misma molécula de ácido nucleico de molde objetivo individual mediante la asignación de unas secuencias que comprenden unas primeras secuencias de etiqueta molecular que son homólogas entre sí y unas segundas secuencias de etiqueta molecular que son homólogas entre sí al mismo agregado;

(ii) analizar las secuencias de las múltiples moléculas de ácidos nucleicos de molde fragmentadas para identificar las secuencias de las múltiples moléculas de ácidos nucleicos de molde fragmentadas que comprenden una primera etiqueta molecular que es homóloga de la primera etiqueta molecular de las secuencias de un primer agregado o una segunda etiqueta molecular que es homóloga de la segunda etiqueta molecular de las secuencias del primer agregado;

(iii) reconstruir la secuencia de una primera molécula de ácido nucleico de molde mediante la alineación de las secuencias que comprenden al menos un subconjunto de las secuencias de las múltiples moléculas de ácidos nucleicos de molde fragmentadas identificadas en la etapa (ii) y definir una secuencia consenso a partir de estas secuencias; y

(iv) llevar a cabo las etapas (i) hasta (iii) con respecto a una segunda molécula de ácido nucleico de molde y/o a una adicional.

3. El procedimiento de la reivindicación 2 en el que:

(A) el procedimiento comprende además una etapa de enriquecer las múltiples moléculas de molde fragmentadas para aumentar la proporción de las múltiples moléculas de ácido nucleico de molde fragmentadas que comprenden la primera etiqueta molecular o la segunda etiqueta molecular y en donde esta etapa es anterior a la etapa f); y/o

(B) la etapa g) (i) comprende además:

determinar una secuencia de consenso para las primeras secuencias de etiqueta molecular y una secuencia de consenso para las segundas secuencias de etiqueta molecular de un primer agregado y la etapa (ii) comprende identificar secuencias de las múltiples moléculas de ácido nucleico de molde fragmentadas que comprenden una primera etiqueta molecular o una segunda etiqueta molecular que es homóloga a la secuencia de consenso para la primera etiqueta molecular o la secuencia de consenso para la segunda etiqueta molecular del primer agregado; y/o

(C) la etapa g) es una etapa del procedimiento implementada por ordenador; y/o

(D) las etapas e) y/o f) se llevan a cabo usando una tecnología de secuenciación que comprende una etapa de PCR de puente, opcionalmente en la que la etapa de la PCR de puente se lleva a cabo usando un tiempo de extensión de más de 15 segundos; y/o

(E) las etapas e) y f) se llevan a cabo en diferentes análisis de secuenciación.

4. Un procedimiento implementado por ordenador para la determinación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo individual que comprende las siguientes etapas:

(a) obtener los datos que comprenden las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprende una primera etiqueta molecular en un extremo y una segunda etiqueta molecular en el otro extremo, en las que cada molécula de ácido nucleico de molde objetivo está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular, y en las que las regiones comprenden la primera etiqueta molecular y la segunda etiqueta molecular;

(b) analizar los datos que comprenden las secuencias de las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular y la segunda etiqueta molecular para identificar los agregados de las secuencias que es probable que se correspondan con la misma molécula de ácido nucleico de molde objetivo individual mediante la asignación de las secuencias que comprenden unas primeras etiquetas moleculares que son homólogas entre sí y unas segundas etiquetas moleculares que son homólogas entre sí al mismo agregado;

(c) obtener los datos que comprenden las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas, en las que cada uno de los fragmentos comprenden bien la primera etiqueta molecular o bien la segunda etiqueta molecular;

(d) analizar las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas para identificar las secuencias de los fragmentos múltiples de las al menos dos moléculas de

ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular que es homóloga de la primera etiqueta molecular de las secuencias de un primer agregado o la segunda etiqueta molecular que es homóloga de la segunda etiqueta molecular de las secuencias del primer agregado;

(e) reconstruir la secuencia de una primera molécula de ácido nucleico de molde objetivo mediante la alineación de las secuencias que comprenden al menos un subconjunto de las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas identificadas en la etapa (d) y definir una secuencia consenso a partir de estas secuencias; y

(f) llevar a cabo las etapas (c) hasta (e) con respecto a una segunda molécula de ácido nucleico de molde objetivo y/o a una adicional, opcionalmente

en el que la etapa (b) comprende adicionalmente la determinación de una secuencia consenso para las primeras secuencias de etiquetas moleculares y una secuencia consenso para las segundas secuencias de etiquetas moleculares de un primer agregado, y la etapa (d) comprende la identificación de las secuencias de las múltiples moléculas de ácidos nucleicos de molde fragmentadas que comprenden una primera etiqueta molecular o una segunda etiqueta molecular que es homóloga de la secuencia consenso para la primera etiqueta molecular o la secuencia consenso para la segunda etiqueta molecular del primer agregado.

5. Un procedimiento implementado por ordenador para la determinación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo que comprende las siguientes etapas:

(a) obtener los datos que comprenden los agregados de las secuencias, en los que:

(i) cada agregado comprende las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas, en las que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprende una primera etiqueta molecular en un extremo y una segunda etiqueta molecular en el otro extremo, en las que cada ácido nucleico de molde objetivo está etiquetado con una única primera etiqueta molecular y una única segunda etiqueta molecular y en las que las regiones comprenden la primera etiqueta molecular y la segunda etiqueta molecular;

(ii) cada agregado comprende las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que cada uno de los fragmentos comprende bien la primera etiqueta molecular o bien la segunda etiqueta molecular;

(iii) las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas en cada agregado comprenden unas primeras etiquetas moleculares y unas segundas etiquetas moleculares que son homólogas entre sí

(iv) las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprenden la primera etiqueta molecular que es homóloga de la primera etiqueta molecular de las secuencias de las regiones de las múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas en ese agregado o la segunda etiqueta molecular que es homóloga de la segunda etiqueta molecular de las secuencias de las regiones de múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en ese agregado;

(b) reconstruir la secuencia de una primera molécula de ácido nucleico de molde objetivo mediante la alineación de las secuencias que comprenden al menos un subconjunto de las secuencias de los fragmentos múltiples de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en un primer agregado y definir una secuencia consenso a partir de estas secuencias; y

(c) llevar a cabo la etapa (b) con respecto a una segunda molécula de ácido nucleico de molde y/o a una adicional.

6. El procedimiento de una cualquiera de las reivindicaciones 2-5 que comprende adicionalmente las etapas de:

(v) identificar los agregados de las secuencias de las regiones de las múltiples copias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que es probable que se correspondan con la misma molécula de ácido nucleico de molde mediante la asignación de las secuencias que comprenden unas primeras secuencias de etiqueta molecular que son homólogas entre sí y unas segundas secuencias de etiqueta molecular que son homólogas entre sí al mismo agregado;

(vi) seleccionar al menos un agregado de secuencias en el que las secuencias de los agregados seleccionados comprenden una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre si que con una primera etiqueta molecular o una segunda etiqueta molecular diferente; en el que la secuencia de la primera molécula de ácido nucleico de molde objetivo es

reconstruida a partir de las secuencias del agregado seleccionado en la etapa (vi), opcionalmente en el que la etapa (vi) consiste en la identificación de los grupos de agregados de las secuencias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas, en los que las secuencias de los agregados de cada grupo tienen unas primeras etiquetas moleculares que son homólogas entre sí y/o la identificación de los grupos de agregados de las secuencias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en los que las secuencias de los agregados de cada grupo tienen unas segundas etiquetas moleculares que son homólogas entre sí y la selección de un agregado del grupo de agregados de las secuencias en el que el agregado que es seleccionado contiene el mayor número de secuencias.

7. El procedimiento de una cualquiera de las reivindicaciones 2-4, en el que:

(A) la etapa e) es una etapa del procedimiento implementada por ordenador; y/o

(B) la etapa d) se lleva a cabo usando una tecnología de secuenciación que comprende una etapa de PCR de puente, opcionalmente en el que la etapa de la PCR de puente se lleva a cabo usando un tiempo de extensión de más de 15 segundos.

8. Un procedimiento implementado por ordenador para la determinación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo individual que comprende las siguientes etapas:

(a) obtener los datos que comprenden las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprende una primera etiqueta molecular en un extremo y una segunda etiqueta molecular en el otro extremo, en las que cada molécula de ácido nucleico de molde objetivo está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular y en las que las regiones comprenden la primera etiqueta molecular y la 5 segunda etiqueta molecular;

(b) analizar los datos que comprenden las secuencias de las regiones de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas que comprenden la primera etiqueta molecular y la segunda etiqueta molecular para identificar los agregados de las secuencias que es probable que se correspondan con la misma molécula de ácido nucleico de molde mediante la asignación de unas secuencias que comprenden unas primeras etiquetas moleculares que son homólogas entre sí y unas segundas etiquetas moleculares que son homólogas entre sí al mismo agregado;

(c) seleccionar al menos un agregado de secuencias en el que las secuencias de los agregados seleccionados comprenden una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre si que con una primera etiqueta molecular o una segunda etiqueta molecular diferente;

(d) reconstruir una secuencia consenso de una primera molécula de ácido nucleico de molde objetivo mediante la alineación de al menos un subconjunto de las moléculas de las secuencias del agregado seleccionado en la etapa (c) y definir una secuencia consenso a partir de estas secuencias; y

(e) llevar a cabo las etapas (c) hasta (d) con respecto a una segunda molécula de ácido nucleico de molde objetivo y/o a una adicional.

9. La etapa del procedimiento de la reivindicación 1 (iv) o la etapa del procedimiento de la reivindicación 8 (c) que consiste en la identificación de los grupos de agregados de las secuencias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que las secuencias de los agregados de cada grupo tienen unas etiquetas moleculares en 5' que son homólogas entre sí y/o la identificación de los grupos de agregados de las secuencias de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que las secuencias de los agregados de cada grupo tienen unas etiquetas moleculares en 3' que son homólogas entre sí y la selección de un agregado de un grupo de agregados de las secuencias en el que el agregado que es seleccionado contiene el mayor número de secuencias.

10. Un procedimiento implementado por ordenador para la determinación de las secuencias de al menos una molécula de ácido nucleico de molde objetivo que comprende

(a) obtener los datos que comprenden un agregado de secuencias;

(b) reconstruir una secuencia consenso de una primera molécula de ácido nucleico de molde mediante la alineación de las secuencias de al menos un subconjunto de las secuencias del agregado seleccionado;

en el que las secuencias del agregado seleccionado comprenden las secuencias de las regiones de múltiples copias de al menos dos moléculas de ácidos nucleicos de molde etiquetadas en las que cada una de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas comprende una primera etiqueta molecular

en un extremo y una segunda etiqueta molecular en el otro extremo, en las que cada una de las al menos dos moléculas de ácidos nucleicos de molde objetivo está etiquetada con una única primera etiqueta molecular y una única segunda etiqueta molecular y en las que las regiones comprenden la primera etiqueta molecular y la segunda etiqueta molecular; y cada secuencia del agregado seleccionado

(i) comprende una primera etiqueta molecular que es homóloga de la primera etiqueta molecular de las otras secuencias de ese agregado, y una segunda etiqueta molecular que es homóloga de la segunda etiqueta molecular de las otras secuencias de ese agregado;

(ii) comprende una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre si que con una primera etiqueta molecular o una segunda etiqueta molecular diferente.

11. El procedimiento de una cualquiera de las reivindicaciones 3 (B)-10 en el que:

(A) las primeras etiquetas moleculares de las secuencias del mismo agregado tienen una identidad de al menos el 90 % de la secuencia entre sí y/o

(B) las segundas etiquetas moleculares de las secuencias del mismo agregado tienen una identidad de al menos el 90 % de la secuencia entre sí.

12. El procedimiento de la reivindicación 3 (B), 6 o 7 que es un procedimiento implementado por ordenador.

13. El procedimiento de una cualquiera de las reivindicaciones anteriores en el que:

(A) las regiones comprenden más de 25 pares de bases que comprenden la primera etiqueta molecular o la segunda etiqueta molecular; y/o

(B) se secuencian las regiones que comprenden la longitud total de las al menos dos moléculas de ácidos nucleicos de molde etiquetadas; y/o

(C) la primera etiqueta molecular y la segunda etiqueta molecular son introducidas en las al menos dos moléculas de ácidos nucleicos de molde usando un procedimiento seleccionado entre el grupo que consiste en una PCR, una tagmentación y un cizallamiento físico o una digestión de restricción de la al menos una molécula de ácido nucleico de molde, seguido de la ligación de los ácidos nucleicos que comprenden la etiqueta molecular en 5' o la etiqueta molecular en 3', opcionalmente

en el que la primera etiqueta molecular y la segunda etiqueta molecular son introducidas en las al menos dos moléculas de ácidos nucleicos de molde mediante una PCR usando unos cebadores que comprenden una porción que comprende la primera etiqueta molecular o la segunda etiqueta molecular y una porción que tiene una secuencia que es capaz de hibridar con las al menos dos moléculas de ácidos nucleicos de molde; y/o

(D) las al menos dos moléculas de ácidos nucleicos de molde codifican el 16S ribosómico microbiano; y/o

(E) la menos una de las al menos dos moléculas de ácidos nucleicos de molde es menor de 20 10 Kpb en tamaño.

14. Un programa informático adaptado para llevar a cabo el procedimiento de la reivindicación 4, 5, 8 o 10, la etapa g) del procedimiento de la reivindicación 2 o la etapa e) del procedimiento de la reivindicación 1, cuando dicho programa es ejecutado en un dispositivo electrónico.

15. Un medio legible por ordenador que almacena el programa informático de la reivindicación 14.

Figura 1

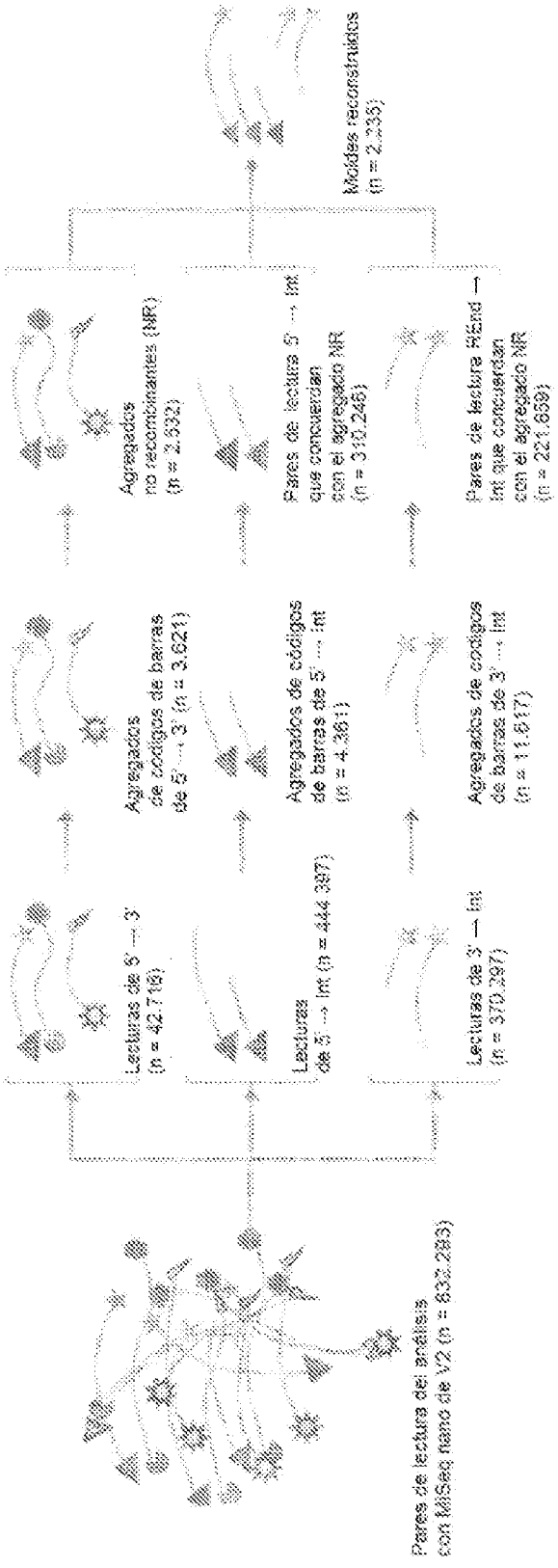

Pares de lectura del análisis
con MiSeq nano de V2 (n = 832.283)
Lecturas de 5' → 3'
(n = 42.716)
Lecturas
de 5' → Int (n = 444.387)
Lecturas de 3' → Int
(n = 370.297)
Agregados
de códigos de barras
de 5' → 3' (n = 3.621)
Agregados de códigos
de barras de 5' → Int
(n = 4.361)
Agregados de códigos
de barras de 3' → Int
(n = 11.617)
Agregados
no reconocimientos (NR)
(n = 2.632)
Pares de lectura 5' → Int
que concuerdan
con el agregado NR
(n = 310.246)
Pares de lectura REnd →
Int que concuerdan
con el agregado NR
(n = 221.659)
Moldes reconstruidos
(n = 2.235)

Figura 2a

Tabla 1: cebadores usados durante la amplificación y la secuenciación del gen 16S

Nombre del cebador	Secuencia
Long_forward_1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNNN%N%GTTGGCCGCGAGAGTTTGATCMTG GCTCAG
Long_forward_2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNNN%N%TATTAACTNCGAGAGTTTGATCMT GGCTCAG
Long_forward_3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNNN%N%CTAATKGCBNCGAGAGTTTGATCM TGGCTCAG
Long_forward_4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNNN%NNNAACCAGTCNNNCGAGAGTTTGATC MTGGCTCAG
Long_forward_5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNNN%NNNGAACGGAKCGAGAGTTTGATCMTG GCTCAG
Long_forward_6	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNNN%N%ACTGAAGTNCGAGAGTTTGATCMT GGCTCAG
Long_forward_7	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNNN%NNTTGGCTAYNNCGAGAGTTTGATCM TGGCTCAG
Long_forward_8	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNN%NNNNTGGCGATT%NNCGAGAGTTTGATC MTGGCTCAG
Long_forward_9	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNNN%NNNCCTCTGATCGAGAGTTTGATCMTG GCTCAG
Long_forward_10	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNN%NNN%NNNNCTCATCCGNCGAGAGTTTGATCMT GGCTCAG
Long_forward_11	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NNN%NNNNTTCAGCGANNCGAGAGTTTGATCM TGGCTCAG
Long_forward_12	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NN%NN%NNGGATGCCA%NNCGAGAGTTTGATC MTGGCTCAG
Long_forward_13	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NN%NN%NNNCGGTCGAGCGAGAGTTTGATCMTG GCTCAG
Long_forward_14	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NN%NN%NNNAAGACTACNCGAGAGTTTGATCMT GGCTCAG
Long_forward_15	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NN%NN%NNNAACGCTAANNCGAGAGTTTGATCM TGGCTCAG
Long_forward_16	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NN%NN%NNNGCCTACGCNN%CGAGAGTTTGATC MTGGCTCAG
Long_forward_17	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NN%NN%NNNTGACTGCTCGAGAGTTTGATCMTG GCTCAG
Long_forward_18	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NN%NN%NNNCAACCTTAN%CGAGAGTTTGATCM TGGCTCAG
Long_forward_20	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NNN%NN%NNGGAGGCTCN%NCGAGAGTTTGATC MTGGCTCAG
Long_forward_21	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NN%NN%NNNAATCGATACGAGAGTTTGATCMTG GCTCAG
Long_forward_22	ACACTCTTTCCCTACACGACGCTCTTCCGATCTN%NN%NN%NNNACCAATTGBNCGAGAGTTTGATCMT GGCTCAG

Figura 2b

Long_forward_23	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNCCTAATAANNCGAGAGTTTGATCM TGGCTCAG
Long_forward_24	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGGATTAGGNNNCGAGAGTTTGATC MTGGCTCAG
Long_forward_25	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNGCGTTACCCGAGAGTTTGATCMTG GCTCAG
Long_reverse_1	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGTTGCCCGTAGACGGGCGGTGTGT RCA
Long_reverse_2	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTATTAACTNNNTAGACGGCCGGTG TGTRCA
Long_reverse_3	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCTAATCGCTAGACGGCCGTGTGT RCA
Long_reverse_4	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNAACCAGTCNNNTAGACGGCCGGTG TGTRCA
Long_reverse_5	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGAACGGAGTAGACGGCCGGTGTGT RCA
Long_reverse_6	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNACTGAAGTNNTAGACGGCCGGTGT GTRCA
Long_reverse_7	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTTGCTATNNNTAGACGGCCGGTG TGTRCA
Long_reverse_8	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTGCCGATTAGACGGGCGGTGTGT RCA
Long_reverse_9	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCCTCTGATNTAGACGGGCGGTGTG TRCA
Long_reverse_10	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCTCATGCGNNTAGACGGCCGGTGT GTRCA
Long_reverse_11	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTTCAGCGANTAGACGGGCGGTGTG TRCA
Long_reverse_12	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGATGCCANNTAGACGGGCGGTGT GTRCA
Long_reverse_13	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCGGTCGAGNTAGACGGGCGGTGTG TRCA
Long_reverse_14	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNAAGACTACNNNTAGACGGGCGGTG TGTRCA
Long_reverse_15	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNAACGCTAATAGACGGGCGGTGTGT RCA
Long_reverse_16	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGCCTACGCNTAGACGGGCGGTGTG TRCA
Long_reverse_17	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNTGACTGCTNNTAGACGGGCGGTGT GTRCA
Long_reverse_18	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNATTGCCGCNTAGACGGGCGGTGTG TRCA
Long_reverse_19	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCAACCTTANNTAGACGGGCGGTGT GTRCA

Figura 2c

Long_reverse_20	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGAGGCTGNTAGACGGGCGGTGTG TRCA
Long_reverse_21	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNAATCGATANNTAGACGGGCGGTGT GTRCA
Long_reverse_22	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNACCAATTGNTAGACGGGCGGTGTG TRCA
Long_reverse_23	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNCCTAATAANTAGACGGGCGGTGTG TRCA
Long_reverse_24	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGATTAGGNNTAGACGGGCGGTGT GTRCA
Long_reverse_25	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGCGTTACCGNNTAGACGGGCGGTG TGTRCA
PE_1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG
PE_2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCG
Caponso_forward_1	AATGATACGGCGACCACCGAGATCTACACAACCAGTCTATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caponso_forward_2	AATGATACGGCGACCACCGAGATCTACACAACGCTAATATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caponso_forward_3	AATGATACGGCGACCACCGAGATCTACACAAGACTACTATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caponso_forward_4	AATGATACGGCGACCACCGAGATCTACACAATCGATATATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caponso_forward_5	AATGATACGGCGACCACCGAGATCTACACACCAATTGTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caponso_forward_6	AATGATACGGCGACCACCGAGATCTACACACTGAAGTTATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caponso_forward_7	AATGATACGGCGACCACCGAGATCTACACATTGCCGCTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caponso_forward_8	AATGATACGGCGACCACCGAGATCTACACCAACCTTATATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caponso_forward_9	AATGATACGGCGACCACCGAGATCTACACCCTAATAATATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caponso_forward_10	AATGATACGGCGACCACCGAGATCTACACCCTCTGATTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caponso_forward_14	AATGATACGGCGACCACCGAGATCTACACGAACCGAGTATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caponso_forward_15	AATGATACGGCGACCACCGAGATCTACACGCGTTACCTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caponso_forward_18	AATGATACGGCGACCACCGAGATCTACACGGATGCCATATGGTAATTGTGTGCCAGCMGCCGCGGT AA
Caponso_forward_31	AATGATACGGCGACCACCGAGATCTACACGTTGGCCGTATGGTAATTGTGTGCCAGCMGCCGCGGTA A
Caponso_forward_32	AATGATACGGCGACCACCGAGATCTACACTGACTGCTTATGGTAATTGTGTGCCAGCMGCCGCGGTA A

Figura 2d

Caporaso_forward_24	AATGATACGGCGACCACCGAGATCTACACTCAGCGATATGGTAATTGTGTGCCAGCMGCCGCGGTAA
Caporaso_reverse_1	CAAGCAGAAGACGGCATACGAGATAACCAGTCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_7	CAAGCAGAAGACGGCATACGAGATATTGCCGCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_8	CAAGCAGAAGACGGCATACGAGATCAACCTTAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_9	CAAGCAGAAGACGGCATACGAGATCCTAATAAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_15	CAAGCAGAAGACGGCATACGAGATGCCTACGCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_16	CAAGCAGAAGACGGCATACGAGATGCGTTACCAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_17	CAAGCAGAAGACGGCATACGAGATGGAGGCTGAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_23	CAAGCAGAAGACGGCATACGAGATTGGCGATTAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_24	CAAGCAGAAGACGGCATACGAGATTTCAGCGAAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_reverse_25	CAAGCAGAAGACGGCATACGAGATTTGGCTATAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Illumina_E_1	AATGATACGGCGACCACCGA
Illumina_E_2	CAAGCAGAAGACGGCATACGA
Caporaso_read_1	TATGGTAATTGTGTGCCAGCMGCCGCGGTAA
Caporaso_read_2	AGTCAGTCAGCCGGACTACHVGGGTWTCTAAT
Caporaso_index_read	ATTAGAWACCCBDGTAGTCCGGCTGACTGACT

Figura 3

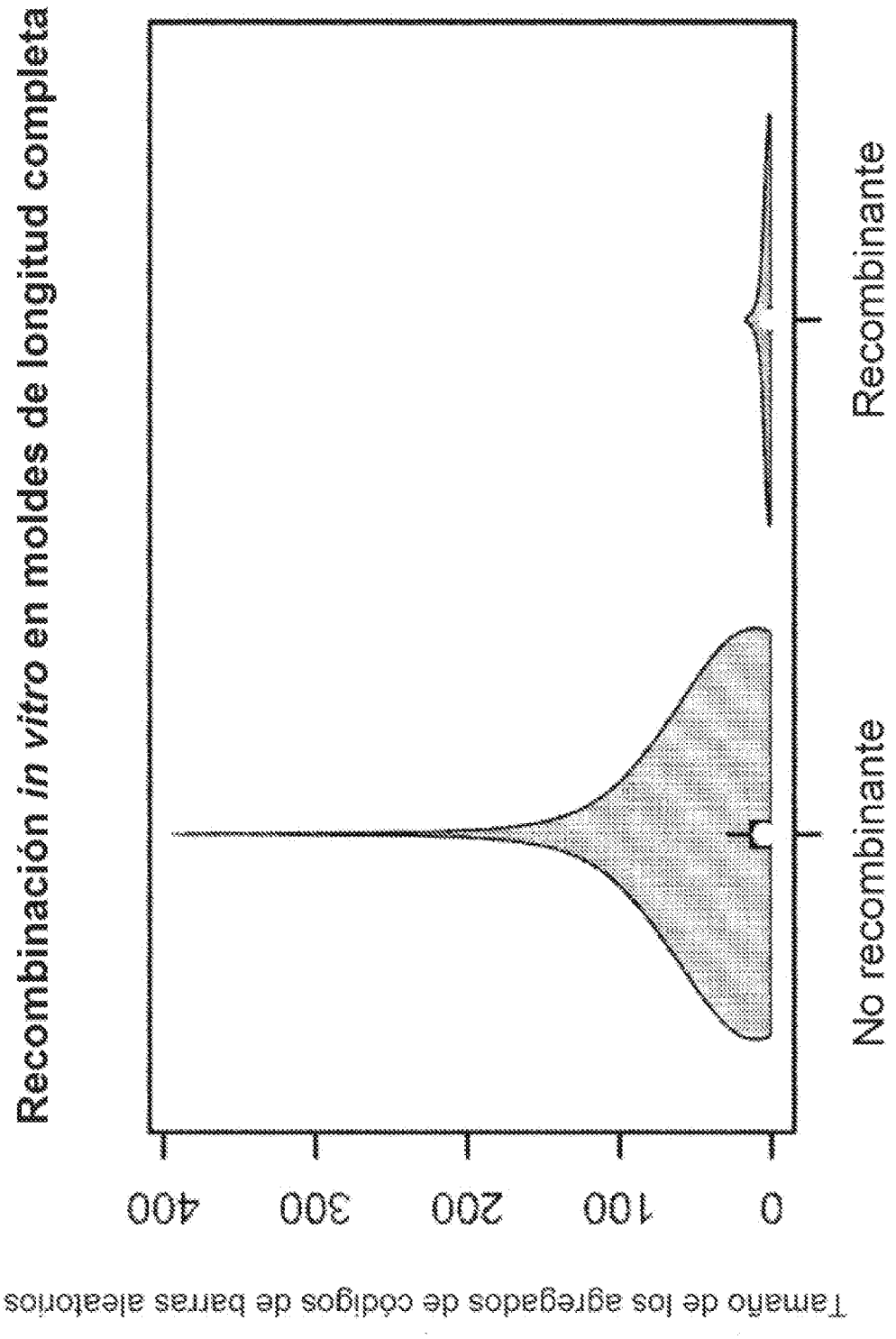

Recombinación *in vitro* en moldes de longitud completa
Recombinante
No recombinante
0
100
200
300
400
Tamaño de los agregados de códigos de barras aleatorios

Figura 4

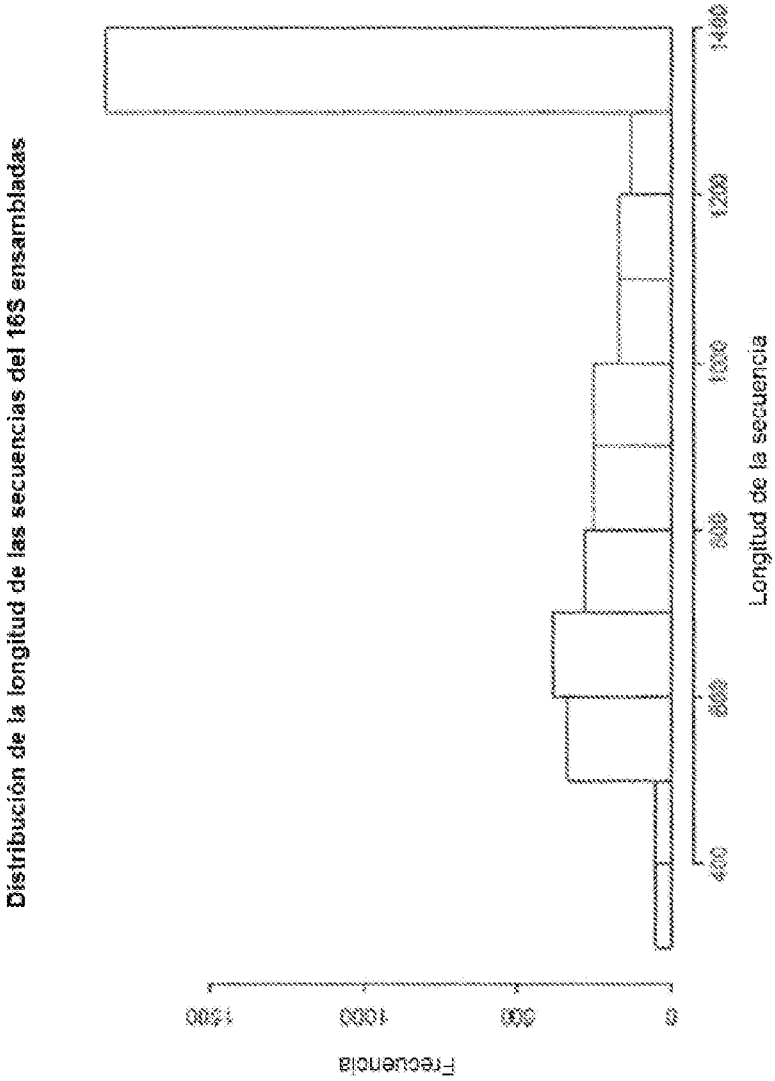

Distribución de la longitud de las secuencias del 16S ensambladas
Frecuencia
0
500
1000
1500
Longitud de la secuencia
400
600
800
1000
1200
1400

Figura 5

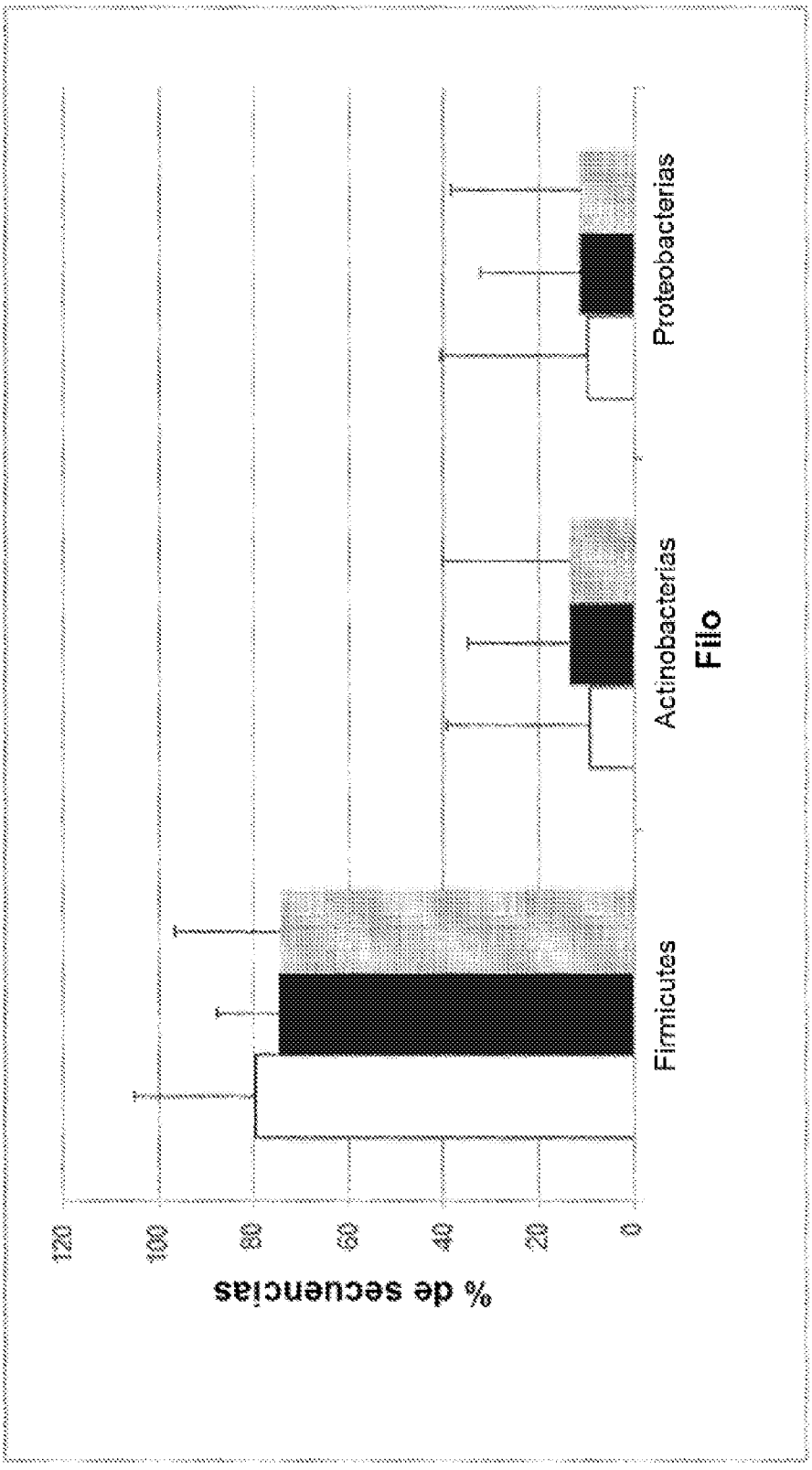

Proteobacterias
Actinobacterias
Firmicutes
Filo
120
100
80
60
40
20
0
% de secuencias

Figura 6

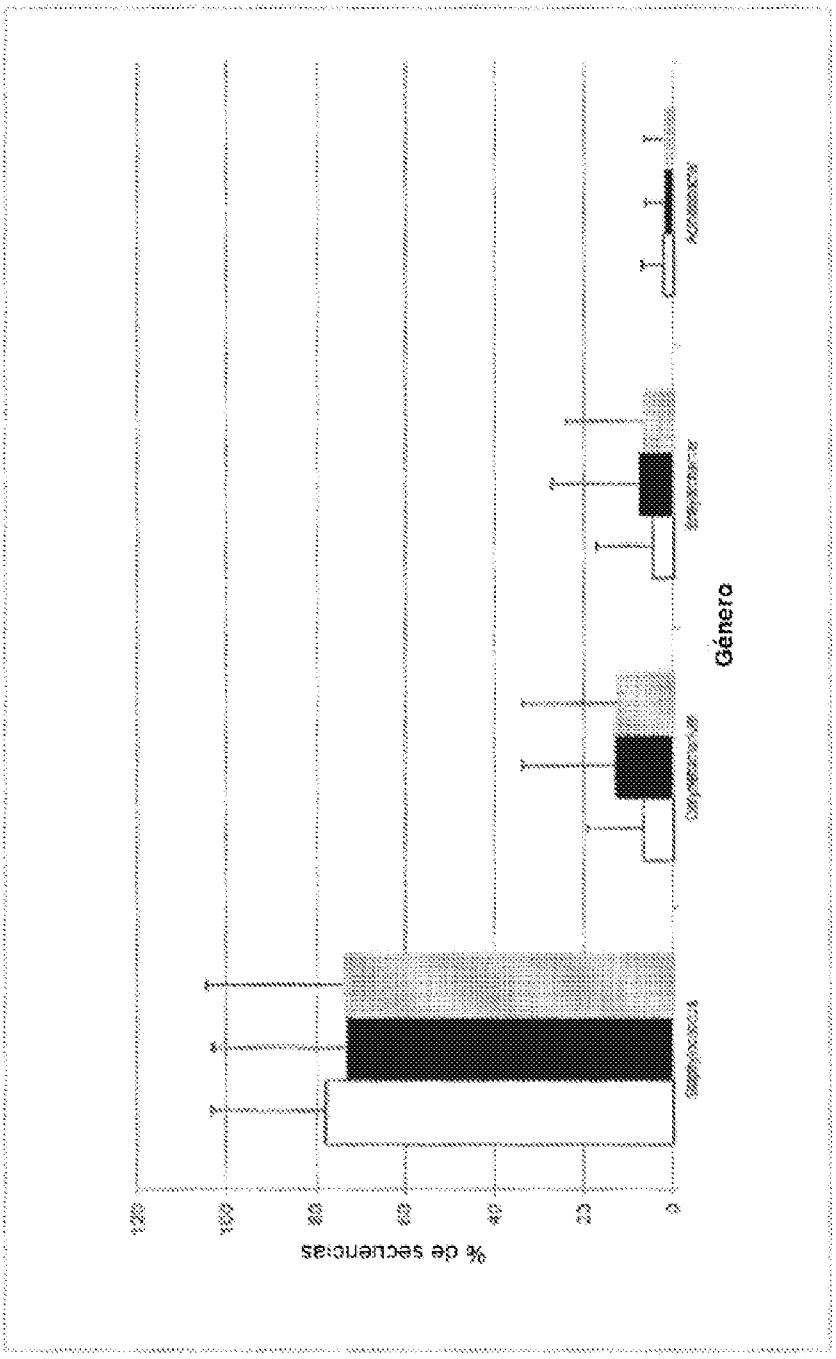


Figura 7

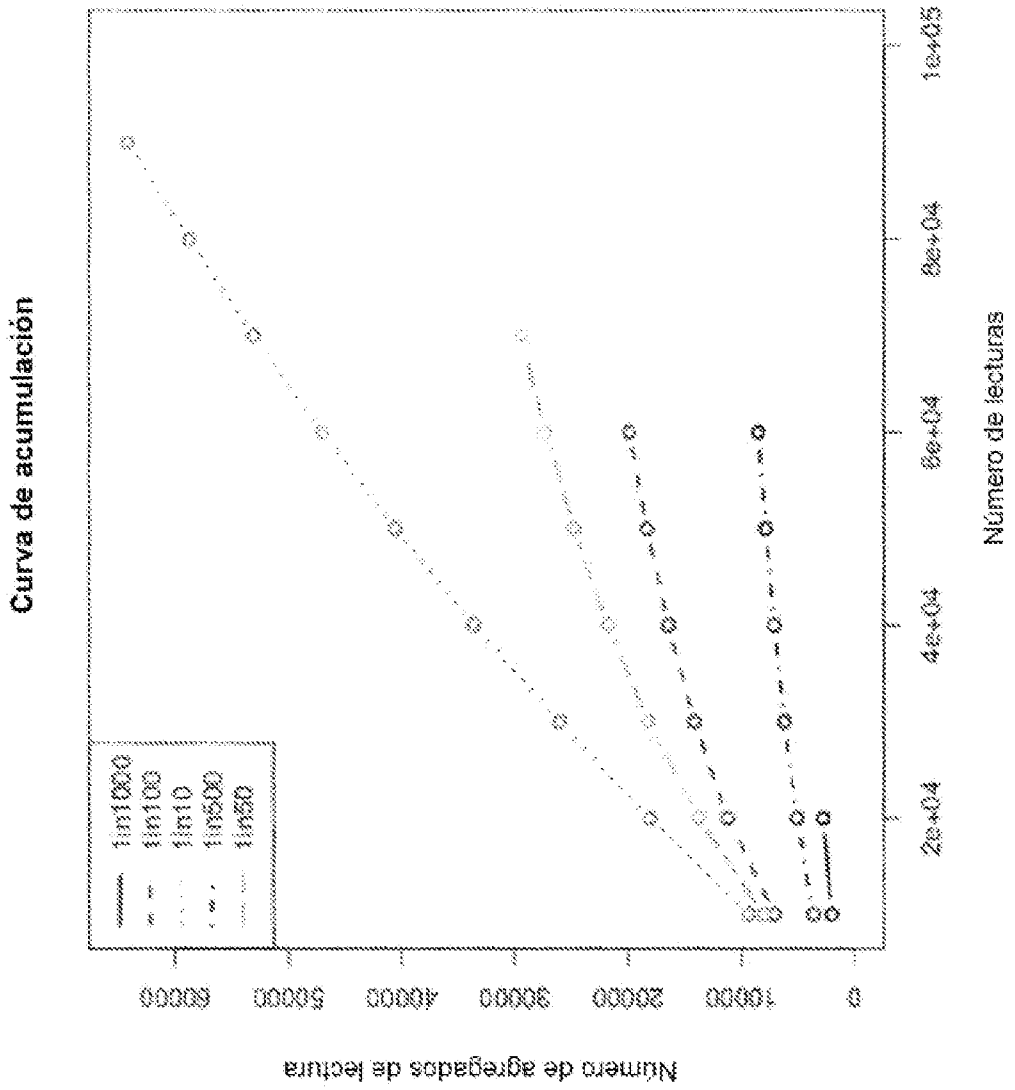

Curva de acumulación
Número de agregados de lectura
Número de lecturas
0
10000
20000
30000
40000
50000
60000
2e+04
4e+04
6e+04
8e+04
1e+05
1/m1000
1/m100
1/m10
1/m500
1/m50

Figura 8

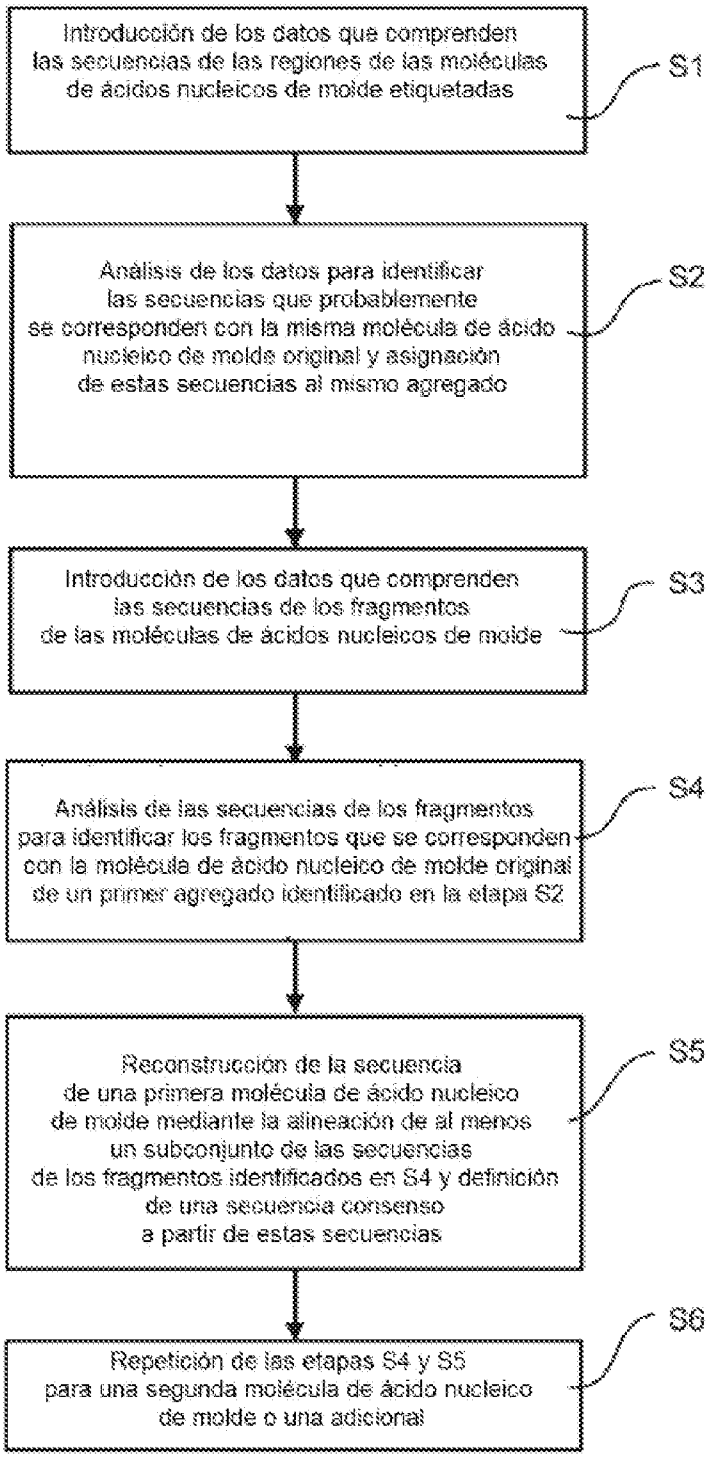

Introducción de los datos que comprenden las secuencias de las regiones de las moléculas de ácidos nucleicos de molde etiquetadas
S1
Análisis de los datos para identificar las secuencias que probablemente se corresponden con la misma molécula de ácido nucleico de molde original y asignación de estas secuencias al mismo agregado
S2
Introducción de los datos que comprenden las secuencias de los fragmentos de las moléculas de ácidos nucleicos de molde
S3
Análisis de las secuencias de los fragmentos para identificar los fragmentos que se corresponden con la molécula de ácido nucleico de molde original de un primer agregado identificado en la etapa S2
S4
Reconstrucción de la secuencia de una primera molécula de ácido nucleico de molde mediante la alineación de al menos un subconjunto de las secuencias de los fragmentos identificados en S4 y definición de una secuencia consenso a partir de estas secuencias
S5
Repetición de las etapas S4 y S5 para una segunda molécula de ácido nucleico de molde o una adicional
S6

Figura 9

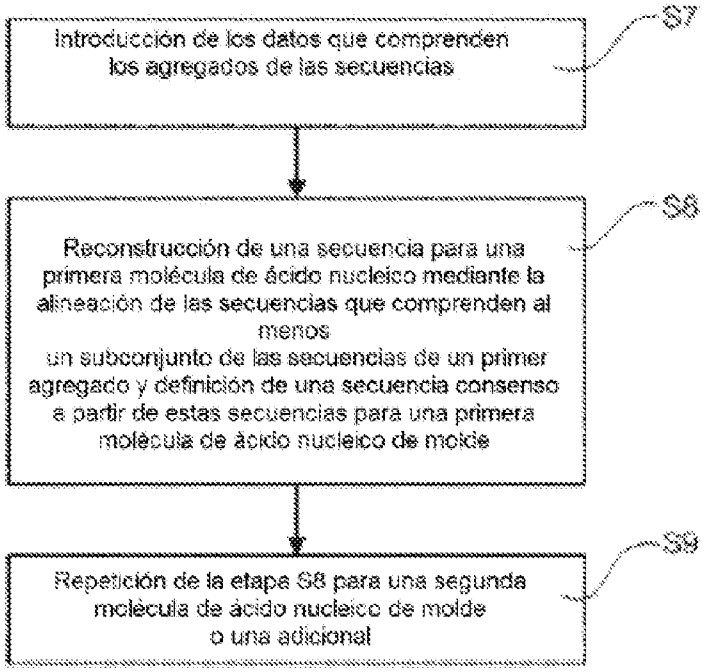

Introducción de los datos que comprenden los agregados de las secuencias
S7
Reconstrucción de una secuencia para una primera molécula de ácido nucleico mediante la alineación de las secuencias que comprenden al menos un subconjunto de las secuencias de un primer agregado y definición de una secuencia consenso a partir de estas secuencias para una primera molécula de ácido nucleico de molde
S8
Repetición de la etapa S8 para una segunda molécula de ácido nucleico de molde o una adicional
S9

Figura 10

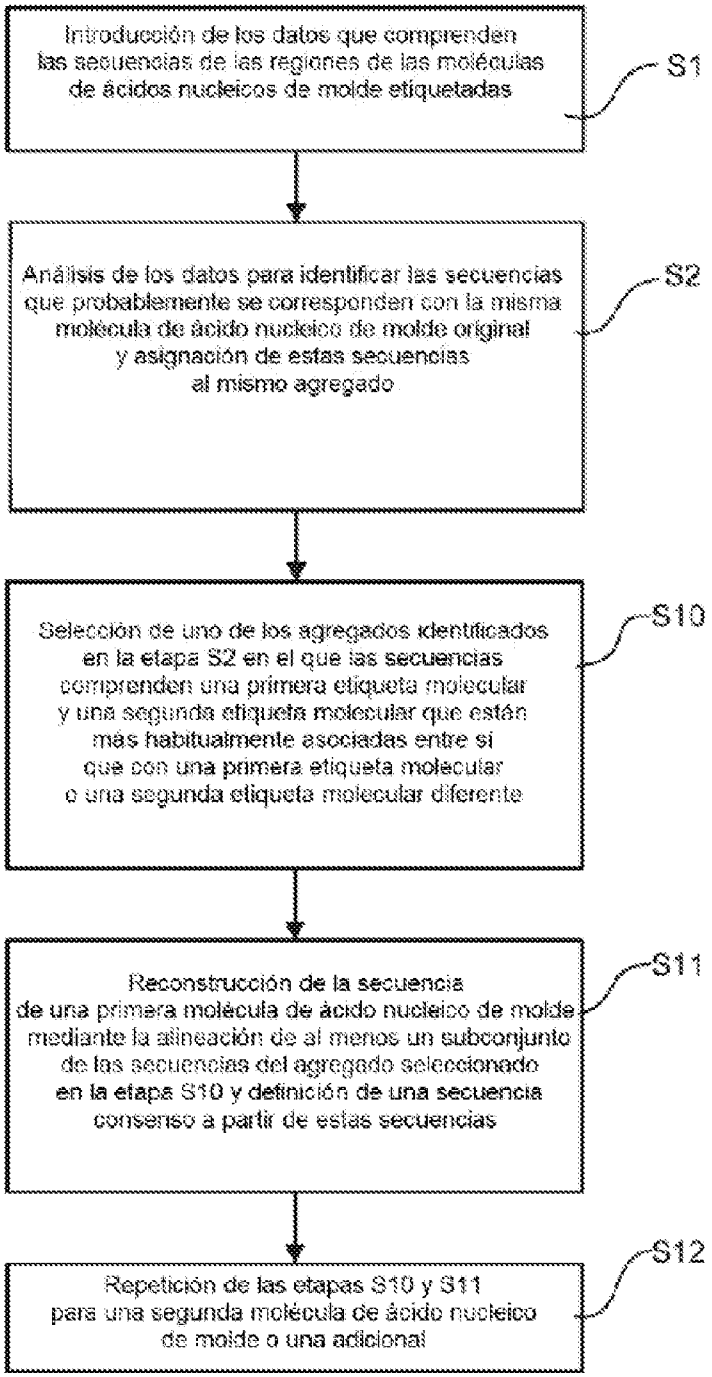
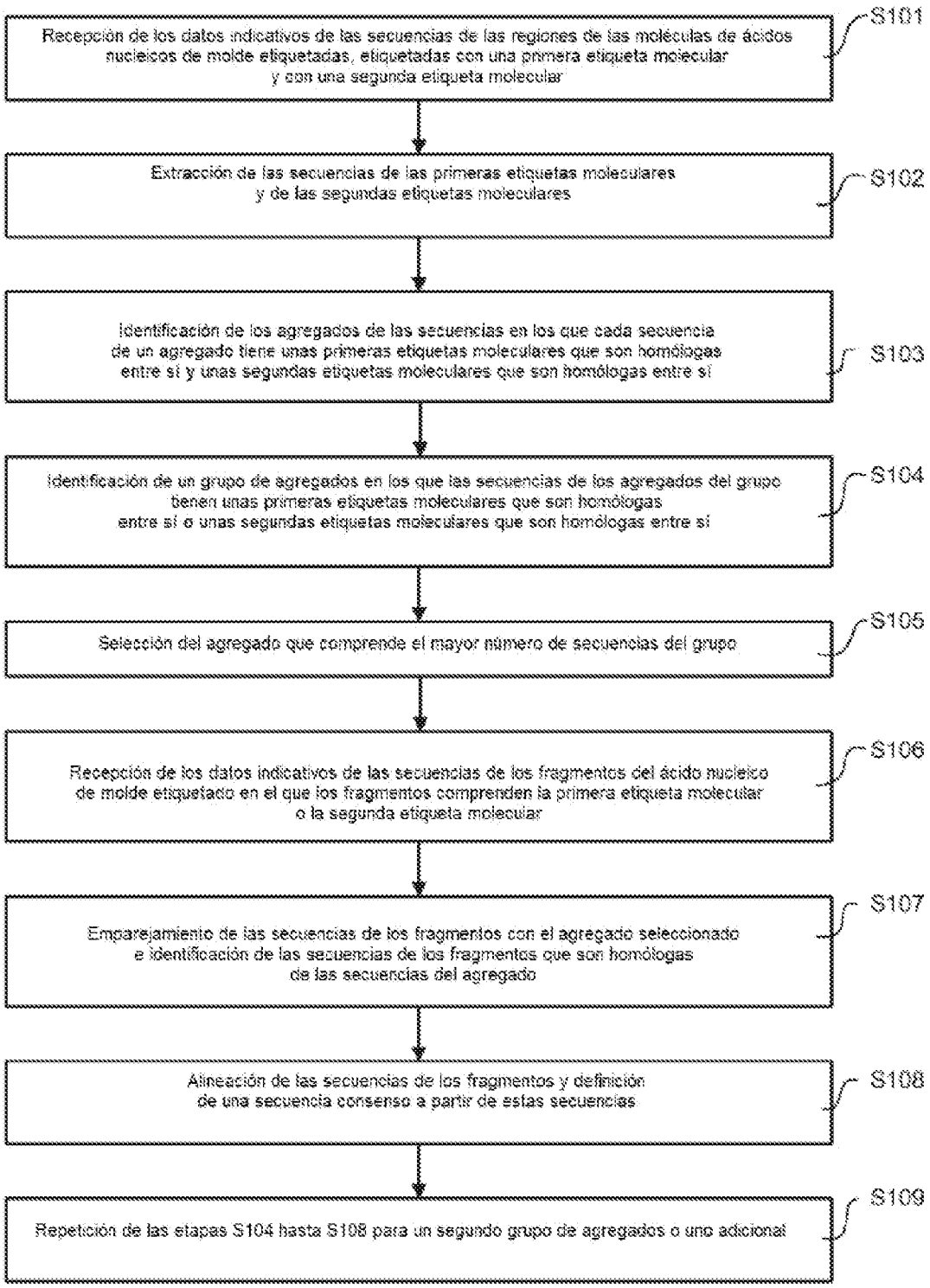

Introducción de los datos que comprenden las secuencias de las regiones de las moléculas de ácidos nucleicos de molde etiquetadas
S1
Análisis de los datos para identificar las secuencias que probablemente se corresponden con la misma molécula de ácido nucleico de molde original y asignación de estas secuencias al mismo agregado
S2
Selección de uno de los agregados identificados en la etapa S2 en el que las secuencias comprenden una primera etiqueta molecular y una segunda etiqueta molecular que están más habitualmente asociadas entre sí que con una primera etiqueta molecular o una segunda etiqueta molecular diferente
S10
Reconstrucción de la secuencia de una primera molécula de ácido nucleico de molde mediante la alineación de al menos un subconjunto de las secuencias del agregado seleccionado en la etapa S10 y definición de una secuencia consenso a partir de estas secuencias
S11
Repetición de las etapas S10 y S11 para una segunda molécula de ácido nucleico de molde o una adicional
S12

Figura 11


Recepción de los datos indicativos de las secuencias de las regiones de las moléculas de ácidos nucleicos de molde etiquetadas, etiquetadas con una primera etiqueta molecular y con una segunda etiqueta molecular
S101
Extracción de las secuencias de las primeras etiquetas moleculares y de las segundas etiquetas moleculares
S102
Identificación de los agregados de las secuencias en los que cada secuencia de un agregado tiene unas primeras etiquetas moleculares que son homólogas entre sí y unas segundas etiquetas moleculares que son homólogas entre sí
S103
Identificación de un grupo de agregados en los que las secuencias de los agregados del grupo tienen unas primeras etiquetas moleculares que son homólogas entre sí o unas segundas etiquetas moleculares que son homólogas entre sí
S104
Selección del agregado que comprende el mayor número de secuencias del grupo
S105
Recepción de los datos indicativos de las secuencias de los fragmentos del ácido nucleico de molde etiquetado en el que los fragmentos comprenden la primera etiqueta molecular o la segunda etiqueta molecular
S106
Emparejamiento de las secuencias de los fragmentos con el agregado seleccionado e identificación de las secuencias de los fragmentos que son homólogas de las secuencias del agregado
S107
Alineación de las secuencias de los fragmentos y definición de una secuencia consenso a partir de estas secuencias
S108
Repetición de las etapas S104 hasta S108 para un segundo grupo de agregados o uno adicional
S109