(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 887 024**

(51) Int. CI.:

G16B 10/00 (2009.01)
G16B 40/10 (2009.01)
G16B 50/10 (2009.01)
C12Q 1/689 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **25.07.2018 E 18185378 (9)**

(97) Fecha y número de publicación de la concesión europea: **25.08.2021 EP 3435264**

(54) Título: **Procedimiento y sistema de identificación y clasificación de unidades taxonómicas operativas en una muestra metagenómica**

(30) Prioridad:

28.07.2017 IN 201721027000

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
21.12.2021

(73) Titular/es:

TATA CONSULTANCY SERVICES LIMITED (100.0%)
Nirmal Building 9th Floor Nariman Point Mumbai 400 021
Maharashtra , IN

(72) Inventor/es:

MANDE, SHARMILA SHEKHAR;
YADAV, DEEPAK y
DUTTA, ANIRBAN

(74) Agente/Representante:

GONZÁLEZ PECES, Gustavo Adolfo

DESCRIPCIÓN

Procedimiento y sistema de identificación y clasificación de unidades taxonómicas operativas en una muestra metagenómica

Referencia cruzada a solicitudes relacionadas y prioridad

Campo técnico

Las realizaciones de la presente memoria se refieren, en general, al campo de la mejora de la precisión de la clasificación taxonómica de la muestra metagenómica y, más particularmente, a un procedimiento y sistema para la identificación y clasificación de unidades taxonómicas operativas en una muestra metagenómica utilizando secuencias de amplicones de lectura corta.

Antecedentes

los estudios metagenómicos emplean la secuenciación del ADN de los genes marcadores filogenéticos para determinar la estructura de la comunidad microbiana perteneciente a un entorno muestreado y para la clasificación taxonómica de los organismos microbianos que la habitan. Sin embargo, la actual generación de tecnologías de secuenciación de ADN de alto rendimiento y rentable, sólo puede generar "lecturas" cortas (fragmentos de secuencias de ADN de~300 - 600 pares de bases), lo que no es suficiente para cubrir toda la longitud de los genes marcadores filogenéticos. Por ejemplo, el marcador filogenético más utilizado para la clasificación taxonómica de las bacterias es el gen 16S rRNA, que tiene una longitud de unos 1500 pb. Dado que sólo una corta región de este gen puede ser objeto de secuenciación de ADN mediante las tecnologías de secuenciación de la generación actual, los experimentos se diseñan para utilizar "regiones hipervariables" (regiones V) específicas del gen 16S rRNA.

Durante la etapa de clasificación taxonómica, estas secuencias cortas se comparan con los catálogos existentes del gen 16S ARNr (a través de búsquedas de similitud de secuencias) para identificar la cepa, la especie, el género, etc., a los que puede atribuirse su origen. Alternativamente, todas las secuencias pertenecientes a una muestra/entorno se agrupan basándose en la similitud de la secuencia, en la que las secuencias que se han agrupado juntas (que tienen una similitud de secuencia significativa) pueden considerarse originarias del mismo grupo de organismos, también conocido como unidad taxonómica operativa (OTU).

Los procedimientos presentes en el estado de la técnica incluyen la clasificación basada en la base de datos de referencia y la recogida *de novo* de UOT. El procedimiento de clasificación basado en la base de datos de referencia funciona bien para un entorno muestreado cuyos microbios residentes ya han sido catalogados mediante estudios anteriores. El procedimiento *de* recogida de OTU *de novo* permite la identificación/detección de grupos taxonómicos presentes en el entorno muestreado aunque no hayan sido caracterizados/clasificados taxonómicamente con anterioridad. Ambos procedimientos tienen algunos inconvenientes.

Los procedimientos actuales para la identificación o clasificación taxonómica de las UOTs basadas en bases de datos de referencia se basan en bases de datos que catalogan genes marcadores de longitud completa (por ejemplo, genes de ARNr 16S) o UOT de referencia identificadas mediante la agrupación de genes marcadores de longitud completa. Dado que las lecturas/secuencias de consulta utilizadas durante la comparación son sólo "lecturas cortas", los resultados de la identificación/clasificación de UOT pueden ser inexactos y subóptimos.

Además, la tasa de evolución (acumulación de mutaciones) no siempre es uniforme a lo largo de un gen marcador elegido en diferentes clados taxonómicos. Es posible que una región corta se mantenga idéntica durante el curso de la evolución, mientras que las regiones flanqueantes son más propensas a las mutaciones. Alternativamente, una fracción importante del gen marcador puede permanecer sin cambios a través de la evolución, salvo un pequeño tramo hipervariable. Por ello, los resultados de la agrupación de OTU pueden variar significativamente en función de la región corta elegida para la secuenciación. Las UOTs identificadas/clasificadas utilizando procedimientos basados en la referencia y procedimientos *de novo* proporcionarán resultados diferentes por las razones anteriores. Una de las solicitudes de patente del estado de la técnica, WO2016172643A3, titulada "Procedimientos y sistemas para la clasificación taxonómica múltiple", divulga procedimientos para identificar una pluralidad de polinucleótidos, así como para detectar la presencia, ausencia o abundancia de una pluralidad de taxones en una muestra. También se proporcionan sistemas para realizar los procedimientos de la divulgación.

Sumario

A continuación, se presenta un sumario simplificado de algunas realizaciones de la divulgación para proporcionar una comprensión básica de las mismas. Este sumario no es una descripción exhaustiva de las realizaciones. No se pretende identificar los elementos clave/críticos de las realizaciones, ni delimitar el ámbito de las mismas. Su único propósito es presentar algunas realizaciones de forma simplificada como preludio a la descripción más detallada que se presenta a continuación.

En vista de lo anterior, una realización de la presente memoria proporciona un sistema para la identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta. El sistema comprende una base de datos convencional de UOT y una base de datos

convencional de secuencias de referencia, un módulo de recogida de muestras metagenómicas, un secuenciador, una memoria y un procesador. La base de datos de OTUs convencionales y la base de datos de secuencias de referencia convencionales tienen una pluralidad de secuencias de nucleótidos agrupadas en una o más de las unidades taxonómicas operativas convencionales (OTUs) y los clados taxonómicos convencionales. El módulo de recogida de muestras metagenómicas recoge la muestra metagenómica una memoria. El secuenciador secuencia la muestra metagenómica recogida. El procesador está configurado para realizar las etapas de: crear una base de datos de OTU personalizada (OTUX) a partir de la muestra metagenómica secuenciada utilizando una pluralidad de segmentos predefinidos de secuencias de nucleótidos de una de las bases de datos de OTUs convencionales o de la base de datos de secuencias de referencia convencionales, donde los segmentos predefinidos de secuencias de nucleótidos se agrupan en OTUs personalizadas utilizando una técnica de agrupación de secuencias; calcular la propensión de una OTU personalizada a partir de la base de datos de OTUs personalizadas (OTUX) utilizando una fórmula predefinida, en la que la propensión se refiere a una probabilidad de que una OTU personalizada esté asociada a uno o más clados taxonómicos convencionales en la base de datos de secuencias de referencia convencional y a las OTUs convencionales en la base de datos de OTUs convencionales crear una matriz de mapeo que enumere todos los valores de las propensiones para cada una de las OTUs personalizadas con respecto a uno o más clados taxonómicos convencionales y OTUs convencionales; utilizar la base de datos de OTUs personalizadas (OTUX) como base de datos de referencia para la selección de OTU de referencia abierta para clasificar las secuencias de amplicones de lectura corta correspondientes a segmentos predefinidos en las OTUs personalizadas adecuadas y construir una tabla de abundancia que represente la proporción de las secuencias de amplicones de lectura corta clasificadas en cada una de las OTUs personalizadas, en la que la tabla de abundancia represente una mayor precisión de la clasificación de las unidades taxonómicas operativas (OTUs) en la muestra metagenómica.

En otro aspecto, un procedimiento para la identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta. Inicialmente, la muestra metagenómica se recoge mediante un módulo de recogida de muestras metagenómicas. A continuación, la muestra metagenómica se secuencia con un secuenciador. En la siguiente etapa, se obtiene una de las bases de datos de unidades taxonómicas operativas (OTUs) convencionales y una base de datos de secuencias de referencia convencionales, en la que la base de datos de OTU convencionales tiene una pluralidad de secuencias de nucleótidos agrupadas en una o más de las unidades taxonómicas operativas (OTUs) convencionales y clados taxonómicos convencionales. En la siguiente etapa, se crea una base de datos de OTU personalizada (OTUX) a partir de la muestra metagenómica secuenciada utilizando una pluralidad de segmentos predefinidos de secuencias de nucleótidos de una de las bases de datos de OTU convencionales o de la base de datos de secuencias de referencia convencionales, en la que los segmentos predefinidos de secuencias de nucleótidos se agrupan en OTU personalizadas utilizando una técnica de agrupación de secuencias. En la siguiente etapa, la propensión de una OTU personalizada de la base de datos de OTUs personalizadas (OTUX) se calcula utilizando una fórmula predefinida, en la que la propensión se refiere a una probabilidad de que una OTU personalizada esté asociada con uno o más clados taxonómicos convencionales en la base de datos de secuencias de referencia convencionales y las OTUs convencionales en la base de datos de OTU convencionales. A continuación, se crea una matriz de mapeo en la que se enumeran todos los valores de las propensiones para cada una de las OTUs personalizadas con respecto a uno o más clados taxonómicos convencionales y OTUs convencionales. En la siguiente etapa, la base de datos de OTU personalizadas (OTUX) se utiliza como base de datos de referencia para la selección de OTU de referencia abierta para clasificar las secuencias de amplicones de lectura corta correspondientes a segmentos predefinidos en OTUs personalizadas adecuadas. Por último, se construye una tabla de abundancia que representa la proporción de las secuencias de amplicones de lectura corta clasificadas en cada una de las OTUs personalizadas, en la que la tabla de abundancia representa una mayor precisión de la clasificación de las unidades taxonómicas operativas (OTUs) en la muestra metagenómica.

En otra realización, se proporciona un medio no transitorio legible por ordenador que tiene incorporado un programa informático para la identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta. Inicialmente, la muestra metagenómica se recoge mediante un módulo de recogida de muestras metagenómicas. A continuación, la muestra metagenómica se secuencia con un secuenciador. En la siguiente etapa, se obtiene una de las bases de datos de unidades taxonómicas operativas (OTUs) convencionales y una base de datos de secuencias de referencia convencionales, en la que la base de datos de OTUs convencionales tiene una pluralidad de secuencias de nucleótidos agrupadas en una o más de las unidades taxonómicas operativas (OTUs) convencionales y clados taxonómicos convencionales. En la siguiente etapa, se crea una base de datos de OTU personalizada (OTUX) a partir de la muestra metagenómica secuenciada utilizando una pluralidad de segmentos predefinidos de secuencias de nucleótidos de una de las bases de datos de OTUs convencionales o de la base de datos de secuencias de referencia convencionales, en la que los segmentos predefinidos de secuencias de nucleótidos se agrupan en OTUs personalizadas utilizando una técnica de agrupación de secuencias. En la siguiente etapa, la propensión de una OTU personalizada de la base de datos de OTU personalizadas (OTUX) se calcula utilizando una fórmula predefinida, en la que la propensión se refiere a una probabilidad de que una OTU personalizada esté asociada con uno o más clados taxonómicos convencionales en la base de datos de secuencias de referencia convencionales y las OTUs convencionales en la base de datos de OTUs convencionales. A continuación, se crea una matriz de mapeo en la que se enumeran todos los valores de las propensiones para cada una de las OTUs personalizadas con respecto a uno o más clados taxonómicos convencionales y OTUs convencionales. En la siguiente etapa, la base de datos de OTUs personalizadas (OTUX) se utiliza como base de datos de referencia para la selección de OTU de referencia abierta para clasificar las secuencias de amplicones de lectura corta correspondientes a segmentos predefinidos en OTU personalizadas adecuadas. Por

último, se construye una tabla de abundancia que representa la proporción de las secuencias de amplicones de lectura corta clasificadas en cada una de las OTUs personalizadas, en la que la tabla de abundancia representa una mayor precisión de la clasificación de las unidades taxonómicas operativas (OTUs) en la muestra metagenómica.

Los expertos en la materia deberían apreciar que cualquier diagrama de bloques aquí presente representa vistas conceptuales de sistemas ilustrativos que encarnan los principios de la presente materia. Del mismo modo, se apreciará que cualquier diagrama de flujo, diagramas de flujo, diagramas de transición de estado, pseudocódigo, y similares representan varios procesos que pueden ser sustancialmente representados en un medio legible por ordenador y así ejecutados por un dispositivo informático o procesador, independientemente de que dicho dispositivo informático o procesador se muestre explícitamente.

Breve descripción de los dibujos

Las realizaciones de la presente memoria se entenderán mejor a partir de la siguiente descripción detallada con referencia a los dibujos, en los que:

La Fig. 1 ilustra un diagrama de bloques para la identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta según una realización de la presente divulgación;

La Fig. 2 muestra la organización de las diferentes regiones hipervariables en el gen 16S rRNA según una realización de la divulgación; y

Las Fig. 3a-3b es un diagrama de flujo que ilustra las etapas implicadas en la identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta según una realización de la presente divulgación.

Descripción detallada de las realizaciones

Las realizaciones de la presente memoria y las diversas características y detalles ventajosos de las mismas se explican más ampliamente con referencia a las realizaciones no limitantes que se ilustran en los dibujos adjuntos y se detallan en la siguiente descripción. Los ejemplos utilizados en la presente memoria están destinados simplemente a facilitar la comprensión de las formas en que las realizaciones de la presente memoria pueden ser practicadas y para permitir a los expertos en la materia practicar las realizaciones de la presente memoria. Por consiguiente, los ejemplos no deben interpretarse como una limitación del alcance de las realizaciones en la presente memoria expuestas.

Glosario - términos utilizados en las realizaciones

La expresión "unidades taxonómicas operativas" u "OTUs" en el contexto de la presente divulgación se refiere a las secuencias que se han agrupado (que tienen una similitud de secuencia significativa) y que pueden considerarse originarias del mismo grupo de organismos. En general, las unidades taxonómicas operativas se definen en función del umbral de similitud. Mientras que una base de datos OTU personalizada se denominará "OTUX"

Con referencia ahora a los dibujos, y más particularmente a la Fig. 1 a la Fig. 3, en las que caracteres de referencia similares denotan características correspondientes de manera consistente a través de las figuras, se muestran realizaciones preferidas y estas realizaciones se describen en el contexto del siguiente sistema y/o procedimiento ejemplar.

Según una realización de la divulgación, en la Fig. 1 se muestra un sistema 100 para la identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta. La divulgación se refiere a un procedimiento y un sistema para mejorar la precisión de la clasificación taxonómica de las secuencias metagenómicas obtenidas mediante la secuenciación de amplicones de lectura corta. La divulgación también proporciona un marco para facilitar la comparación cruzada de las estructuras de la comunidad del microbioma muestreadas a través de diferentes estudios metagenómicos desconectados, en los que se pueden haber utilizado diferentes tecnologías de secuenciación, así como diferentes secuencias de marcadores o amplicones.

Según una realización de la divulgación, el sistema 100 consiste en una interfaz de usuario 102, una base de datos de UOT convencionales 104, una base de datos de secuencias de referencia convencionales 106, una memoria 108 y un procesador 110, como se muestra en la Fig. 1. El procesador 110 está en comunicación con la memoria 108. El procesador 110 está configurado para ejecutar una pluralidad de algoritmos almacenados en la memoria 108. El procesador 110 incluye además una pluralidad de módulos para realizar diversas funciones. El procesador 110 puede incluir el módulo de creación de una base de datos personalizada 112, un módulo de cálculo de la propensión 114, un módulo de creación de una matriz de mapeo 116, un módulo de clasificación 118, un módulo de creación de una tabla de abundancia 120 y un módulo de retro-mapeo 122.

Según una realización de la divulgación, el sistema 100 incluye además un módulo de recogida de muestras metagenómicas 124 y un secuenciador 126. La muestra metagenómica se recoge del intestino de un individuo utilizando el módulo de recogida de muestras metagenómicas 124. Aunque debe apreciarse que la muestra metagenómica también puede recogerse de cualquier otro entorno, como la piel, el mar, el suelo, etc. Los fragmentos

de ADN extraídos de la muestra metagenómica se secuencian con un secuenciador 126. El ADN secuenciado se proporciona entonces al procesador 110 mediante la interfaz de usuario 102. Las muestras de ADN secuenciadas también se denominan secuencias de "consulta". La interfaz de usuario 102 es operada por un usuario. La interfaz de usuario 102 puede incluir una variedad de interfaces de software y hardware, por ejemplo, una interfaz web, una interfaz gráfica de usuario y similares, y puede facilitar múltiples comunicaciones dentro de una amplia variedad de redes N/W y tipos de protocolo, incluyendo redes cableadas, por ejemplo, LAN, cable, etc., y redes inalámbricas, como WLAN, celular o satélite.

Según una realización de la divulgación, el sistema 100 incluye dos bases de datos precalculadas, es decir, la base de datos de UOT convencionales 104 y la base de datos de secuencias de referencia convencionales 106. La base de datos convencional de OTUs 104 y la base de datos convencional de secuencias de referencia 106 tienen una pluralidad de secuencias de nucleótidos agrupadas en una o más unidades taxonómicas operativas convencionales (OTUs), y uno o más clados taxonómicos convencionales respectivamente. Debe apreciarse que las dos bases de datos precalculadas están disponibles en la técnica anterior. El uso de cualquier otra base de datos está dentro del ámbito de la presente divulgación.

Según una realización de la divulgación, el flujo de trabajo tiene dos componentes principales, a saber: (1) un preprocesamiento de una sola vez para crear una base de datos de OTU personalizada llamada bases de datos de referencia OTUX y una "matriz de mapeo" (MAPMAT) para diferentes regiones V y (2) una etapa de selección de OTU de referencia abierta y de asignación/clasificación taxonómica utilizando la(s) base(s) de datos de referencia OTUX. La selección de OTU de referencia abierta implica la selección de OTU y la clasificación taxonómica de secuencias metagenómicas de lectura corta dirigidas a la región V4. Tras el enfoque de selección de OTU de referencia abierta, inicialmente se realiza una asignación de OTU basada en la referencia en el conjunto de secuencias metagenómicas de la consulta utilizando la $OTUX_{V4}$ como base de datos de referencia, en la que cada una de las secuencias de la consulta se clasifica en OTU $OTUX_{V4}$ adecuadas sujetas a un umbral de confianza.

El sistema 100 incluye el módulo de creación de la base de datos personalizada 112 para crear la base de datos personalizada de OTU (OTUX). La base de datos de OTUs personalizadas (OTUX) comprende una pluralidad de OTU personalizadas. La base de datos OTU personalizada (OTUX) se crea utilizando segmentos predefinidos de secuencias de nucleótidos de una de las bases de datos OTUs convencionales 104 o de la base de datos de secuencias de referencia convencionales 106. El segmento predefinido corresponde a una pequeña porción de la secuencia de ADN de longitud completa que puede ser objeto de la secuenciación del amplicón. Además, diferentes secuencias predefinidas corresponden a diferentes porciones de la secuencia completa de ADN que pueden ser extraídas/amplificadas utilizando diferentes cebadores.

Según una realización de la divulgación, el sistema 100 incluye además un módulo de cálculo de la propensión 114. El módulo de cálculo de la propensión 114 está configurado para calcular la propensión de una OTU personalizada a partir de la base de datos de OTUs personalizadas (OTUX) utilizando una fórmula predefinida. La fórmula predefinida es

Fórmula predefinida = (número de segmentos predefinidos de secuencias agrupadas en OTU personalizada correspondiente a la base de datos OTU (OTUX) cuyas contrapartes de longitud completa son asignadas a una OTU convencional o un cado taxonómico convencional presente en una base de datos OTU convencional) / (número total de segmentos predefinidos de secuencias agrupadas en OTU correspondientes a la base de datos OTU (OTUX) personalizada)

La propensión calculada se refiere a una probabilidad de que una UOT personalizada se asocie con uno o más clados taxonómicos convencionales en la base de datos de secuencias de referencia convencionales 106 y con la una o más UOT convencionales en la base de datos de UOTs convencionales 104. Además, el sistema 100 está configurado para crear una matriz de mapeo utilizando el módulo de creación de matrices de mapeo 116. La matriz de mapeo enumera todos los valores de las propensiones para cada una de las OTUs personalizadas presentes en la base de datos de OTU personalizadas (OTUX) con respecto a uno o más clados taxonómicos convencionales y OTUs convencionales.

Según una realización de la divulgación, el sistema 100 incluye además el módulo de clasificación 118. El módulo de clasificación 118 está configurado para utilizar la base de datos de OTU personalizada (OTUX) como base de datos de referencia para la selección de OTU de referencia abierta para clasificar las secuencias de amplicones de lectura corta (secuencias de consulta) correspondientes a los segmentos predefinidos en OTU personalizadas adecuadas. El sistema 100 está configurado además para crear una tabla de abundancia que represente la proporción de las secuencias de amplicones de lectura corta (secuencias de consulta) clasificadas en cada una de las OTUs personalizadas utilizando el módulo 120 de creación de tablas de abundancia.

Según una realización de la divulgación, el sistema 100 puede ampliarse para la selección de OTU y la clasificación taxonómica de metagenomas utilizando cualquier gen marcador / región de secuencias de nucleótidos obtenida de la muestra metagenómica. Sin embargo, a efectos ilustrativos, la presente divulgación ejemplifica el procedimiento y la aplicabilidad utilizando lo siguiente: Gen marcador - gen procariota 16S rRNA (con 9 regiones hipervariables V1-V9); Región hipervariable - V4 (región hipervariable 4); Base de datos de OTU de referencia convencional - Greengenes 13.8 (que contiene secuencias completas de 16S rRNA agrupadas en OTU convencionales). La Fig. 2 muestra la

organización de las diferentes regiones hipervariables en el gen 16S rRNA según una realización de la divulgación. Según otra realización de la divulgación, el gen marcador puede ser cualquier otro gen junto con sus regiones hipervariables, por ejemplo, gen ITS, 23S rRNA, 18S rRNA, etc.

Inicialmente, se recuperan todas las secuencias 'prokMSA' no alineadas de la base de datos Greengenes (v13.8 utilizada en esta realización). Para cada una de estas secuencias, también se recupera la clasificación taxonómica para los diferentes niveles jerárquicos taxonómicos, incluyendo el filo, la clase, el orden, la familia, el género y la especie, así como los correspondientes OTU ID de Greengenes (OTU ID convencionales). En la siguiente etapa, se extrae la región V4 de cada secuencia presente en la base de datos. A continuación, las secuencias extraídas se agrupan en función de la similitud de la secuencia, y cada grupo resultante está formado por secuencias que comparten un 99% de identidad de secuencia entre sí. Para la agrupación de secuencias en esta realización se utilizó Cd-hit, cuya referencia se toma del documento de investigación: "Cd-hit: un programa rápido para agrupar y comparar grandes conjuntos de secuencias de proteínas o nucleótidos" por Weizhong Li & Adam Godzik Bioinformatics, (2006) 22:1658-9. En la siguiente etapa, se asigna a cada clúster (OTU) un "OTUX $_{V4}$ ID" único (digamos OTUX $_{V4}$i), y se compilan todos los clústeres para constituir una "base de datos de referencia OTUX $_{V4}$". En el siguiente etapa, se calcula la propensión (MAPMAT $_{V4}$i,j) de que la OTUX $_{V4}$i se asocie a una OTU de Greengenes (GGj) mediante la siguiente fórmula:

MAPMAT$_{v4}$i,j = (número de secuencias agrupadas en OTUX $_{v4}$i cuyas contrapartes de longitud completa son asignadas a GGj / (número total de secuencias agrupadas en OTUX$_{v4}$i)

Además, la matriz de propensión MAPMAT $_{V4}$ se rellena para la base de datos OTUX $_{V4}$ calculando todos los valores para MAPMAT $_{V4}$i,j, en la que,

i = 1 al número total de OTU OTUX $_{V4}$ (digamos N_{OTUX}),

j = 1 al número total de OTU Greengenes (digamos N_{GG}),

y MAPMAT $_{V4}$ es una matriz $N_{GG} \times N_{OTUX}$

$$\text{MAPMAT}_{V4} = \begin{pmatrix} MAPMATv4\ 1{,}1 & MAPMATv4\ 2{,}1 & MAPMATv4\ 3{,}1 & \dots & MAPMATv4\ i{,}1 \\ MAPMATv4\ 1{,}2 & MAPMATv4\ 2{,}2 & MAPMATv4\ 3{,}2 & \dots & MAPMATv4\ i{,}2 \\ MAPMATv4\ 1{,}3 & MAPMATv4\ 2{,}3 & MAPMATv4\ 3{,}3 & \dots & MAPMATv4\ i{,}3 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ MAPMATv4\ 1{,}j & MAPMATv4\ 2{,}j & MAPMATv4\ 3{,}j & \dots & MAPMATv4\ i{,}j \end{pmatrix}$$

En la siguiente etapa, se realiza la selección de OTU y la clasificación taxonómica de las secuencias metagenómicas de lectura corta dirigidas a la región $_{V4}$. Siguiendo el enfoque de selección de OTU de referencia abierta, inicialmente se realiza una asignación de OTU basada en la referencia en el conjunto de consulta de secuencias metagenómicas utilizando la OTUX $_{V4}$ como base de datos de referencia, en la que cada una de las secuencias de consulta se clasifica en OTU OTUX $_{V4}$ apropiadas sujetas a un umbral de confianza. El algoritmo de clasificación utilizado puede ser el clasificador bayesiano ingenuo utilizado por RDP (algoritmo de Wang) con un umbral de confianza bootstrap del 80%, en una realización. En la siguiente etapa, las secuencias que no pueden clasificarse en las OTUX $_{V4}$ existentes se agrupan (por ejemplo, utilizando CD-HIT con un umbral de identidad de secuencia del 99%) en *"* OTU *denovo* ". En la siguiente etapa, se genera una tabla de abundancia de OTU (T_{OTUX}) acumulando el número total de lecturas secuenciadas de una muestra metagenómica que podrían clasificarse / atribuirse a cada una de las OTUs de OTUX $_{V4}$. Los resultados de la clasificación obtenidos en términos de OTUX $_{V4}$ OTU se mapean de nuevo utilizando MAPMAT $_{V4}$ para representar los resultados en términos de la base de datos de secuencias de ARNr 16S de longitud completa utilizada convencionalmente (Greengenes v13.8 en esta realización) OTU ID.

De acuerdo con una realización de la divulgación, el mapeo de vuelta se puede lograr utilizando dos procedimientos alternativos. En el primer procedimiento, para asignar cada una de las secuencias de la consulta a un OTU ID de Greengenes concreto, se sigue el siguiente proceso:

- Para una determinada secuencia de consulta 's' que ha sido asignada a la OTU OTUX $_{V4}$x, recuperar los elementos MAPMAT $_{V4}$ {MAPMAT $_{V4}$x,j} (en la que 'j' = 1 --> N_{GG}, es decir, el número total de OTU de Greengen).
- Se calcula el valor máximo de {MAPMAT $_{V4}$x,j}.
- La secuencia 's' se clasifica en la OTU 'y' de Greengenes (GGy), en la que, MAPMAT $_{V4}$x,y = max{MAPMAT $_{V4}$x,j}
- El proceso se repite para todas las secuencias de consulta y, posteriormente, se genera una tabla de abundancia de OTU (TGG), en términos de OTU ID de Greengenes, acumulando el número total de lecturas secuenciadas de la muestra metagenómica dada que podrían clasificarse/atribuirse a cada una de las OTUs de Greengenes.

En el segundo procedimiento, para representar la estructura de la comunidad microbiana perteneciente a una muestra metagenómica dada en una tabla de abundancia en la que la abundancia de cada microbio (OTU) se representa en términos de valores porcentuales normalizados, se siguen las siguientes etapas:

- Para un conjunto de secuencias de consulta correspondientes a una muestra metagenómica, se genera la tabla de abundancia / perfil T_{OTUX} en la que se representa el número total de secuencias asignadas a cada una de las OTUs $OTUX_{V4}$.

$$T_{OTUX} = \begin{pmatrix} a \\ b \\ c \\ \vdots \\ z \end{pmatrix} \ldots \ldots \; OTUX_{v4}1,\; OTUX_{v4}2,\; OTUX_{v4}3,\; \ldots\; OTUX_{v}i$$

Por ejemplo, T_{OTUX} puede representarse en forma de una matriz de columnas (de tamaño $N_{OTUX} \times 1$) como la representada anteriormente, en la que "i" varía de 1 a N_{OTUX}, es decir, el número total de $OTUX_{V4}$, y en la que "a" es el número de secuencias asignadas a la OTU $OTUX_{V4}1$, "b" es el número de secuencias asignadas a $OTUX_{V4}2$, "c" es el número de secuencias asignadas a $OTUX_{V4}3$, y así sucesivamente.

- Obtener una tabla/perfil de abundancia de OTU (T_{GGraw}) para el conjunto de secuencias de consulta, en términos de OTU ID de Greeengenes multiplicando la matriz $MAPMAT_{V4}$ con la matriz T_{OTUX}. Cabe señalar que, dada la naturaleza de la matriz MAPMAT, los valores de abundancia de cada una de las OTUs de Greengenes en T_{GGraw} pueden ser un valor fraccionario.

$$T_{GGraw} = MAPMAT_{V4} * T_{OTUX}$$

en la que, T_{GGraw} es una matriz de columnas de tamaño ($N_{GG} \times 1$), y N_{GG} es el número total de OTUs de Greengenes.

- Obtener una tabla/perfil de abundancia de OTU normalizada en porcentaje ($T_{GG\%}$) realizando la siguiente transformación en cada elemento de T_{GGraw}

$$\mathrm{TGG\%j} = \frac{TGGraw\ j}{\sum_{j=1}^{NGG} TGGraw\ j} * 100$$

en la que, $T_{GG\%}$ es una matriz de columnas de tamaño ($N_{GG \times 1}$), y N_{GG} es el número total de OTU de Greengenes.

En la última etapa, la abundancia de grupos taxonómicos presentes en la muestra metagenómica, obtenida en forma de cualquiera de las tres matrices de columnas, a saber, T_{OTUX}, T_{GG} y $T_{GG\%,}$ se representan además en cualquier nivel taxonómico deseado utilizando la información de la jerarquía taxonómica asociada a las Greengenes OTU. Así, las anotaciones/categorizaciones precisas permiten identificar eficazmente, en una muestra metagenómica, la presencia de grupos taxonómicos específicos. Además, se pueden analizar los grupos taxonómicos específicos, que pueden incluir cepas microbianas infecciosas, microbios de importancia industrial, etc. La categorización precisa proporciona además un marco para facilitar la comparación cruzada de las estructuras comunitarias del microbioma muestreadas en diferentes estudios metagenómicos desconectados.

En funcionamiento, en la Fig. 3 se muestra un diagrama de flujo 200 que ilustra las etapas implicadas para la identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta según una realización de la divulgación. Inicialmente, en la etapa 202, la muestra metagenómica se recoge utilizando el módulo de recogida de muestras metagenómicas 124. La muestra metagenómica puede recogerse del intestino, la piel, el mar, el suelo, etc. En la siguiente etapa 204, la muestra metagenómica recogida se secuenciará con el secuenciador 126. En la siguiente etapa 206, se obtiene una de las bases de datos de unidades taxonómicas operativas convencionales (OTUs) 104 y una base de datos de secuencias de referencia convencional 106. La base de datos OTU convencional 104 tiene una pluralidad de segmentos predefinidos de secuencias de nucleótidos agrupados en una o más de las unidades taxonómicas operativas convencionales (OTUs) y los clados taxonómicos convencionales.

En la etapa 208, se crea una base de datos de OTU personalizada (OTUX) utilizando un segmento predefinido de secuencias de nucleótidos de una de las bases de datos de OTUs convencionales o de la base de datos de secuencias

de referencia convencionales. Los segmentos predefinidos de secuencias de nucleótidos se agrupan en OTUs personalizadas mediante una técnica de agrupación de secuencias. Además, en la etapa 210, la propensión de una OTU personalizada de la base de datos de OTUs personalizadas (OTUX) se calcula utilizando una fórmula predefinida. La propensión se refiere a una probabilidad de que la UOT personalizada esté asociada a uno o más clados taxonómicos convencionales en la base de datos de secuencias de referencia convencionales 106, o a la una o más UOT convencionales en la base de datos de UOTs convencionales 104

En la siguiente etapa 212, se crea una matriz de mapeo. La matriz de mapeo enumera todos los valores de las propensiones para cada una de las OTUs personalizadas con respecto a uno o más clados taxonómicos convencionales y OTUs convencionales. En la etapa 214, la base de datos de OTU personalizadas (OTUX) se utiliza como base de datos de referencia para la selección de OTU de referencia abierta para clasificar las secuencias de amplicones de lectura corta (secuencias de consulta) correspondientes a segmentos predefinidos en OTUs personalizadas adecuadas. Y finalmente, en la etapa 216, se construye una tabla de abundancia que representa la proporción de las secuencias de amplicones de lectura corta clasificadas en cada una de las OTUs personalizadas, en la que la tabla de abundancia representa una mayor precisión de la clasificación de las unidades taxonómicas operativas (OTU) en la muestra metagenómica.

Según una realización de la divulgación, el sistema 100 también puede ser validado como sigue: Para validar la utilidad de la innovación presentada, se utiliza el MAPMAT preprocesado para la región V4 del gen 16S rRNA, que se creó utilizando el procedimiento descrito anteriormente. Para obtener conjuntos de lecturas metagenómicas cortas para clasificarlas en OTUs / otros grupos taxonómicos utilizando el procedimiento presentado, se crearon múltiples metagenomas simulados pertenecientes a cuatro entornos diferentes, a saber, el intestino de niños sanos (GUT), la piel humana sana (SKIN), el mar Mediterráneo (SEA) y el suelo (SOIL) utilizando el siguiente procedimiento. Se recuperaron conjuntos de datos disponibles públicamente relativos a muestras metagenómicas de los entornos mencionados. Se obtuvieron las proporciones globales de los diferentes géneros presentes en cada uno de los ambientes. Posteriormente, se creó un metagenoma simulado perteneciente a un entorno concreto extrayendo aleatoriamente genes de ARNr 16S de longitud completa de la base de datos RDP (v10.3), en la que las proporciones de los diferentes géneros en el subconjunto de secuencias extraídas aleatoriamente reflejaban de forma justa las proporciones observadas en los conjuntos de datos disponibles públicamente considerados. Se crearon 100 conjuntos de datos metagenómicos simulados (cada uno con 10000 secuencias) para cada uno de los 4 entornos ($D_{GUT/F,}$ $D_{SKIN/F,}$ $D_{SEA/F,}$ $D_{SOIL/F}$). Para imitar los conjuntos de datos metagenómicos obtenidos a través de la secuenciación de lecturas cortas, sólo se recortaron las regiones V4 de cada una de las secuencias completas que constituían estos metagenomas simulados y se construyó un conjunto correspondiente de metagenomas simulados de "lectura corta" ($D_{GUT/ V4,}$ $D_{SKIN/ V4}$, $D_{SEA/ V4}$, $D_{SOIL/ V4}$) que sólo contenían las regiones V4.

Inicialmente, las secuencias de longitud completa pertenecientes a cada uno de los conjuntos de datos metagenómicos simulados ($D_{GUT/F}$,$D_{SKIN/F}$, $D_{SEA/F}$, $D_{SOIL/F}$) fueron sometidas a un "selección OTU" (clasificación taxonómica a nivel de OTU) frente a la base de datos Greengenes utilizando el clasificador bayesiano ingenuo utilizado por RDP (algoritmo de Wang con un umbral de confianza bootstrap del 80%). Dado que las secuencias del gen del ARNr 16S de longitud completa se compararon con una base de datos de secuencias del ARNr 16S de longitud completa, los resultados obtenidos reflejaron la mejor clasificación de UOT que se podía conseguir utilizando la secuenciación del amplicón del ARNr 16S (utilizando el mismo algoritmo) y se consideró como la "línea de base" o el "patrón oro" (GS). Los conjuntos de datos metagenómicos simulados de "lectura corta" se sometieron posteriormente a una clasificación taxonómica mediante los dos procedimientos siguientes:

(a) Enfoque convencional (CA): Cada uno de los metagenomas pertenecientes a los conjuntos $D_{OUT/ V4}$, $D_{SKIN/ V4}$, $D_{SEA/ V4}$ y $D_{SOIL/ V4}$ se clasificaron utilizando el clasificador bayesiano ingenuo utilizado por RDP (algoritmo de Wang con un umbral de confianza bootstrap del 80%), y con la base de datos Greengenes OTU como referencia. Estos resultados representan la clasificación taxonómica que puede obtenerse utilizando el enfoque convencional de selección de OTU/clasificación taxonómica, en el que se utilizan secuencias de lectura corta (que cubren una determinada región de un gen marcador) como consulta frente a una base de datos de OTU constituida por genes marcadores de longitud completa. Para facilitar la comparación, se generaron tablas de abundancia que representan la proporción de OTUs (y otros taxones), tanto en términos de recuentos de secuencias en bruto como de abundancia normalizada en porcentaje.

(b) Enfoque OTUX (OTUX): Cada uno de los metagenomas pertenecientes a los conjuntos $D_{OUT/ V4}$, $D_{SKIN/ V4}$, $D_{SEA/ V4}$ y $D_{SOIL/ V4}$ se clasificaron utilizando el clasificador bayesiano ingenuo que utiliza el RDP (algoritmo de Wang con un umbral de confianza bootstrap del 80%), y con la base de datos $OTUX_{V4}$ como referencia. Estos resultados representan la clasificación taxonómica que puede obtenerse utilizando el novedoso enfoque OTUX de selección de OTU/clasificación taxonómica, en el que las secuencias de lectura corta (que cubren una determinada región de un gen marcador) se utilizan como consulta contra una base de datos de OTU precalculada correspondiente a una región hipervariable específica (V4 en este caso). Cabe señalar que la tabla de abundancia de OTU obtenida (T_{OTUX}) informa de los resultados en términos de $OTUX_{V4}$ OTU ID y los resultados pueden considerarse equivalentes a los obtenidos mediante la *"* selección de OTU *de novo* ". Para facilitar la comparación, estos resultados se han vuelto a asignar en términos de OTU ID de Greengenes y se han proporcionado en la tabla de abundancia de OTU T_{GG}, en la que se representan los recuentos brutos de secuencias asignadas a OTU individuales de Greengenes. Además, también se genera una tabla de

abundancia normalizada en porcentaje $T_{GG\%}$, en la que la abundancia/proporción de OTUs (y/o otros taxones) se representa en términos de porcentaje normalizado.

Los resultados de ambos enfoques, el convencional (CA) y el OTUX, obtenidos con los metagenomas simulados de "lectura corta", fueron comparados en base a los tres parámetros: (1) Exactitud de las asignaciones taxonómicas a nivel de OTU, género y familia, evaluada en términos de número correcto de asignaciones (según la GS/línea de base) mediante el enfoque convencional (CA) y el enfoque OTUX; (2) distancia de Unifrac y Bray-Curtis entre la tabla de abundancia normalizada en porcentaje de la GS/línea de base y las generadas por los enfoques convencional (CA) y OTUX; y (3) tiempo de cálculo y memoria utilizados por los enfoques convencional (CA) y OTUX.

Los parámetros primero y segundo mencionados anteriormente pueden explicarse con los siguientes resultados. Las siguientes tablas muestran la mejora del rendimiento de la asignación de OTU basada en OTUX propuesta en esta innovación en comparación con los enfoques convencionales. Se crearon 100 metagenomas simulados para cada uno de los 4 entornos seleccionados, a saber, intestino, piel, mar y suelo. Cada uno de los metagenomas constituía 10000 secuencias que abarcaban la región variable V4. Los conjuntos de datos se sometieron a la asignación de OTU utilizando el enfoque convencional (CA), es decir, utilizando amplicones de la región V como consulta contra la base de datos de referencia Greengenes, así como el enfoque OTUX (OTUX), es decir, utilizando amplicones de la región V como consulta contra las bases de datos de referencia OTUX correspondientes a una región V apropiada. Las asignaciones de OTUX obtenidas para las secuencias individuales, así como la tabla de abundancia obtenida con el enfoque de OTUX (T_{OTUX}), se mapearon de nuevo en términos de OTU IDs de Greengenes (TGG) para comparar los resultados de los dos enfoques. Los resultados de estas asignaciones taxonómicas se evaluaron en cuanto a su exactitud comparándolos con una línea de base/"Gold-Standard" (GS) que se refiere a las asignaciones de OTU obtenidas utilizando las correspondientes secuencias del gen 16S rRNA de longitud completa con la base de datos Greengenes.

Se representa el número medio de asignaciones correctas para 100 metagenomas simulados pertenecientes a cada uno de los entornos. Se ha realizado una prueba T para evaluar si los resultados utilizando OTUX superan significativamente al procedimiento CA. Además, las tablas de abundancia taxonómica normalizada obtenidas por CA y OTUX se han comparado con el GS (Gold-Standard) utilizando distancias Unifrac (tanto ponderadas como no ponderadas), y distancias Bray-Curtis. Los resultados indican que el procedimiento OTUX es superior al procedimiento CA. Los resultados obtenidos con diferentes regiones V convencionales (o combinaciones de ellas) son los siguientes. Los resultados se han representado para diferentes niveles taxonómicos, a saber, OTU, género y familia.

(i) Para el nivel de OTU

Región V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	5803,21	3796,71	4941,84	4156,57
Predicciones correctas de CA	2318,88	320,92	2182,47	2287,78
Puntuación de la prueba T	647,99	721,05	641,73	435,75
Valor P	0,00E+00	2,24E-234	7,85E-275	8,78E-271
Unifrac dist no ponderado (GS vs OTUX)	0,369	0,471	0,446	0,462
Dist. Unifrac no bautizada (GS vs CA)	0,503	0,647	0,601	0,628
Dist. Unifrac ponderada (GS vs OTUX)	0,289	0,235	0,242	0,225
Dist. Unifrac ponderada (GS vs CA)	0,331	0,466	0,406	0,337
Disimilitud de Bray Curtis (GS vs OTUX)	0,471	0,375	0,357	0,333
Disimilitud de Bray Curtis (GS vs CA)	0,577	0,676	0,570	0,464

Región V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Región V2V3				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**PIEL**	**SEA**	**SOIL**
Predicciones correctas de OTUX	5448,5	5371,45	6357,35	5547,01
Predicciones correctas de CA	3808,17	1800,91	3183,1	4706,66
Puntuación de la prueba T	281,45	623,68	668,47	159,73
Valor P	2,04E-258	1,75E-306	4,04E-290	1,21E-210
Unifrac dist no ponderado (GS vs OTUX)	0,302	0,324	0,330	0,329
Dist. Unifrac no ponderada (GS vs CA)	0,311	0,356	0,370	0,300
Dist. Unifrac ponderada (GS vs OTUX)	0,237	0,101	0,127	0,077
Dist. Unifrac ponderada (GS vs CA)	0,281	0,349	0,335	0,152
Disimilitud de Bray Curtis (GS vs OTUX)	0,378	0,221	0,202	0,161
Disimilitud de Bray Curtis (GS vs CA)	0,428	0,528	0,469	0,215

Región V3V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	6262,35	5437,89	6296,22	4626,71
Predicciones correctas de CA	3077,94	1477,09	3292,01	4548,06
Puntuación de la prueba T	470,76	718,45	602,28	15,97
Valor P	1,84E-295	0,00E+00	6,89E-291	2,16E-37
Dist. de Unifrac sin pareja (GS vs OTUX)	0,323	0,363	0,367	0,392
Dist. Unifrac no ponderada (GS vs CA)	0,382	0,406	0,377	0,369
Dist. Unifrac ponderada (GS vs OTUX)	0,241	0,124	0,128	0,141
Dist. Unifrac ponderada (GS vs CA)	0,315	0,348	0,324	0,149
Disimilitud de Bray Curtis (GS vs OTUX)	0,412	0,264	0,200	0,262
Disimilitud de Bray Curtis (GS vs CA)	0,501	0,560	0,458	0,230

Región V3V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Región V5V6				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	5856,33	5471,54	6255,22	5226,09
Predicciones correctas de CA	3079,6	1492,12	2871,11	3710,12
Puntuación de la prueba T	418,18	744,17	645,27	325,53
Valor P	8,47E-287	0,00E+00	2,60E-303	1,77E-259
Unifrac dist no ponderado (GS vs OTUX)	0,338	0,359	0,368	0,410
Dist. Unifrac no ponderada (GS vs CA)	0,386	0,431	0,430	0,472
Dist. Unifrac ponderada (GS vs OTUX)	0,235	0,104	0,153	0,162
Dist. Unifrac ponderada (GS vs CA)	0,310	0,331	0,346	0,194
Disimilitud de Bray Curtis (GS vs OTUX)	0,398	0,245	0,269	0,286
Disimilitud de Bray Curtis (GS vs CA)	0,501	0,558	0,500	0,317

(ii) Para el nivel de género

Región V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	8555,4	8882,85	6999,15	7910,04
Predicciones correctas de CA	7690,71	6453,09	6729,89	7019,75
Puntuación de la prueba T	229,95	579,01	84,31	245,69
Valor P	2,46E-232	0,00E+00	1,75E-155	1,96E-235
Unifrac dist no ponderado (GS vs OTUX)	0,089	0,226	0,151	0,109
Dist. Unifrac no ponderada (GS vs CA)	0,077	0,172	0,155	0,128
Dist. Unifrac ponderada (GS vs OTUX)	0,059	0,060	0,105	0,070

Región V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Dist. Unifrac ponderada (GS vs CA)	0,107	0,246	0,123	0,128
Disimilitud de Bray Curtis (GS vs OTUX)	0,099	0,118	0,148	0,093
Disimilitud de Bray Curtis (GS vs CA)	0,140	0,312	0,158	0,171
Región V2V3				
Promedio de 100 metagenomas simulados	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
(cada uno con 10000 secuencias)				
Predicciones correctas de OTUX	8572,26	9324,14	7766,74	8264,82
Predicciones correctas de CA	8635,28	9092,03	6579,84	6760,23
Puntuación de la prueba T	-15,15	80,60	284,14	376,72
Valor P	6,31E-35	2,60E-152	1,65E-260	4,81 E-265
Unifrac dist no ponderado (GS vs OTUX)	0,078	0,105	0,096	0,069
Dist. Unifrac no ponderada (GS vs CA)	0,046	0,053	0,045	0,070
Dist. Unifrac ponderada (GS vs OTUX)	0,053	0,031	0,072	0,059
Dist. Unifrac ponderada (GS vs CA)	0,035	0,037	0,133	0,159
Disimilitud de Bray Curtis (GS vs OTUX)	0,092	0,077	0,107	0,088
Disimilitud de Bray Curtis (GS vs CA)	0,045	0,049	0,177	0,204
Región V3V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	8053,54	9371,94	8011,17	7946,09
Predicciones correctas de CA	8401,74	8353,05	7123,73	8309,01
Puntuación de la prueba T	-77,00	292,51	222,56	-90,05
Valor P	1,51E-145	2,03E-238	5,23E-236	4,50E-158
Unifrac dist no ponderado (GS vs OTUX)	0,043	0,106	0,096	0,066
Dist. Unifrac no ponderada (GS vs CA)	0,057	0,055	0,041	0,055

Región V3V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Dist. Unifrac ponderada (GS vs OTUX)	0,087	0,027	0,053	0,067
Dist. Unifrac ponderada (GS vs CA)	0,052	0,094	0,094	0,031
Disimilitud de Bray Curtis (GS vs OTUX)	0,123	0,069	0,083	0,095
Disimilitud de Bray Curtis (GS vs CA)	0,069	0,123	0,123	0,042
Región V5V6				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	8799,65	9243,32	7503,73	7103,06
Predicciones correctas de CA	8109,87	8178,85	5482,86	7507,01
Puntuación de la prueba T	194,77	285,25	445,08	-86,34
Valor P	1,18E-227	6,34E-257	1,30E-292	1,61E-158
Unifrac dist no ponderado (GS vs OTUX)	0,151	0,180	0,154	0,110
Dist. Unifrac no ponderada (GS vs CA)	0,048	0,068	0,066	0,049
Dist. Unifrac ponderada (GS vs OTUX)	0,042	0,035	0,095	0,156
Dist. Unifrac ponderada (GS vs CA)	0,076	0,106	0,209	0,094
Disimilitud de Bray Curtis (GS vs OTUX)	0,082	0,079	0,134	0,214
Disimilitud de Bray Curtis (GS vs CA)	0,098	0,138	0,271	0,123

(iii) Para el nivel familiar

Región V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	9921,11	9697,28	9330,66	9299,72
Predicciones correctas de CA	9798,95	8697,26	9333,41	9128,97
Puntuación de la prueba T	84,68	414,59	-1,55	82,57

Región V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Valor P	3,05E-141	4,83E-291	1,22E-01	1,28E-145
Unifrac dist no ponderado (GS vs OTUX)	0,088	0,071	0,070	0,023
Dist. Unifrac no ponderada (GS vs CA)	0,026	0,013	0,029	0,026
Dist. Unifrac ponderada (GS vs OTUX)	0,013	0,028	0,027	0,015
Dist. Unifrac ponderada (GS vs CA)	0,010	0,097	0,017	0,020
Disimilitud de Bray Curtis (GS vs OTUX)	0,046	0,076	0,060	0,023
Disimilitud de Bray Curtis (GS vs CA)	0,013	0,126	0,022	0,027
Región V2V3				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	9801,26	9822,84	9499,73	9333,05
Predicciones correctas de CA	9721,21	9689,8	9462,58	9312,76
Puntuación de la prueba T	39,77	69,30	25,08	10,78
Valor P	5,04E-95	1,95E-135	5,03E-62	1,28E-21
Dist. de Unifrac sin pareja (GS vs OTUX)	0,035	0,067	0,061	0,049
Dist. Unifrac no ponderada (GS vs CA)	0,009	0,026	0,010	0,016
Dist. Unifrac ponderada (GS vs OTUX)	0,023	0,022	0,025	0,026
Dist. Unifrac ponderada (GS vs CA)	0,018	0,020	0,007	0,006
Disimilitud de Bray Curtis (GS vs OTUX)	0,066	0,068	0,055	0,036
Disimilitud de Bray Curtis (GS vs CA)	0,023	0,028	0,009	0,008
Región V3V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	9932,18	9926,97	9535,05	9383,12
Predicciones correctas de CA	9848,85	9846,64	9525,05	9382,02
Puntuación de la prueba T	66,98	60,62	8,76	0,80

Región V3V4				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Valor P	2,07E-122	1,47E-118	8,93E-16	4,25E-01
Unifrac dist no ponderado (GS vs OTUX)	0,024	0,023	0,014	0,012
Dist. Unifrac no ponderada (GS vs CA)	0,001	0,003	0,005	0,012
Dist. Unifrac ponderada (GS vs OTUX)	0,016	0,011	0,01 2	0,012
Dist. Unifrac ponderada (GS vs CA)	0,008	0,009	0,002	0,001
Disimilitud de Bray Curtis (GS vs OTUX)	0,047	0,047	0,029	0,019
Disimilitud de Bray Curtis (GS vs CA)	0,011	0,012	0,003	0,001
Región V5V6				
Promedio de 100 metagenomas simulados (cada uno con 10000 secuencias)	**Entornos**			
	GUT	**SKIN**	**SEA**	**SOIL**
Predicciones correctas de OTUX	9948,13	9834,1	9532,83	9316,6
Predicciones correctas de CA	9401,43	9702,02	9483,68	9372,81
Puntuación de la prueba T	228,65	69,12	34,35	-38,16
Valor P	8,47E-157	2,40E-133	4,98E-82	4,02E-92
Unifrac dist no ponderado (GS vs OTUX)	0,109	0,115	0,081	0,065
Dist. Unifrac no ponderada (GS vs CA)	0,018	0,003	0,003	0,003
Dist. Unifrac ponderada (GS vs OTUX)	0,017	0,019	0,020	0,029
Dist. Unifrac ponderada (GS vs CA)	0,042	0,019	0,005	0,001
Disimilitud de Bray Curtis (GS vs OTUX)	0,066	0,061	0,045	0,041
Disimilitud de Bray Curtis (GS vs CA)	0,055	0,025	0,007	0,002

El tercer parámetro de "tiempo de cálculo y memoria utilizada" mencionado anteriormente puede explicarse con los siguientes resultados. La tabla siguiente muestra el tiempo de cálculo medio requerido por los enfoques convencional (CA) y OTUX para clasificar cada secuencia. También se han indicado los picos de utilización de la memoria por parte de estos enfoques. La prueba de validación se realizó en un servidor basado en Intel Xeon con 40 núcleos de procesamiento (2,0 GHz) y una memoria RAM total de 128 GB. Los valores de tiempo y uso de memoria indicados en la tabla se han normalizado para un solo núcleo de procesamiento.

Región V dirigida	**Tiempo medio (en segundos) necesario para clasificar una sola lectura**	
	Enfoque OTUX	**CA**
V1	0,217	0,507
V2	0,854	1,703
V3	0,640	1,156
V4	1,064	2,078
V5	0,427	1,003
V6	0,482	1,248
V7	0,552	1,206
V8	0,779	1,568
V9	0,202	1,031
V1V2	3,214	5,135
V1V3	7,265	12,857
V2V3	5,308	9,225
V3V4	3,681	7,035
V3V5	7,951	14,235
V3V6	12,794	19,902
V4V5	4,616	8,41 7
V4V6	8,007	14,345
V5V6	3,723	6,501
V6V8	4,711	9,281
Pico de uso de la memoria	1,078 GB	1,261 GB

Los resultados indicaron un rendimiento superior del procedimiento OTUX sobre el enfoque convencional en todos los aspectos comparados. Además, la función de retro-mapeo implementada en el procedimiento OTUX permite una comparación cruzada realista entre los resultados metagenómicos generados con la secuenciación de lectura corta dirigida a cualquiera de las regiones hipervariables.

La descripción anterior describe la materia objeto de la presente memoria para permitir a cualquier persona experta en la materia hacer y utilizar las realizaciones. El alcance de las realizaciones de la materia está definido por las reivindicaciones y puede incluir otras modificaciones que se le ocurran a los expertos en la materia. Estas otras modificaciones se consideran dentro del alcance de las reivindicaciones si tienen elementos similares que no difieren del lenguaje literal de las reivindicaciones o si incluyen elementos equivalentes con diferencias insustanciales del lenguaje literal de las reivindicaciones.

Las realizaciones de la presente divulgación proporcionan un sistema y procedimiento para la identificación y clasificación de unidades taxonómicas operativas (OTU) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta.

Sin embargo, debe entenderse que el alcance de la protección se extiende a dicho programa y además a un medio legible por ordenador que tenga un mensaje en el mismo; dicho medio de almacenamiento legible por ordenador contiene medios de código de programa para la implementación de una o más etapas del procedimiento, cuando el programa se ejecuta en un servidor o dispositivo móvil o cualquier dispositivo programable adecuado. El dispositivo de hardware puede ser cualquier tipo de dispositivo que pueda ser programado, incluyendo, por ejemplo, cualquier tipo de ordenador como un servidor o un ordenador personal, o similar, o cualquier combinación de los mismos. El dispositivo también puede incluir medios que podrían ser, por ejemplo, medios de hardware como, por ejemplo, un circuito integrado de aplicación específica (ASIC), una matriz de puertas programables en campo (FPGA), o una combinación de medios de hardware y software, por ejemplo, un ASIC y una FPGA, o al menos un microprocesador y al menos una memoria con módulos de software ubicados en los mismos. Así, los medios pueden incluir tanto medios de hardware como de software. Las realizaciones del procedimiento descrito en la presente memoria podrían implementarse en hardware y software. El dispositivo también puede incluir medios de software. Alternativamente, las realizaciones pueden ser implementadas en diferentes dispositivos de hardware, por ejemplo, utilizando una pluralidad de CPUs.

Las realizaciones de la presente memoria pueden comprender elementos de hardware y software. Las realizaciones que se implementan en software incluyen, pero no se limitan a, firmware, software residente, microcódigo, etc. Las funciones realizadas por diversos módulos descritos en la presente memoria pueden implementarse en otros módulos o en combinaciones de otros módulos. A los efectos de la presente descripción, un medio utilizable por ordenador o legible por ordenador puede ser cualquier aparato que pueda comprender, almacenar, comunicar, propagar o transportar el programa para su uso por el sistema de ejecución de instrucciones, el aparato o el dispositivo, o en relación con los mismos.

El medio puede ser un sistema (o aparato o dispositivo) electrónico, magnético, óptico, electromagnético, infrarrojo o semiconductor o un medio de propagación. Ejemplos de un medio legible por ordenador incluyen una memoria de semiconductor o de estado sólido, una cinta magnética, un disquete informático extraíble, una memoria de acceso aleatorio (RAM), una memoria de sólo lectura (ROM), un disco magnético rígido y un disco óptico. Algunos ejemplos actuales de discos ópticos son el disco compacto de sólo lectura (CD-ROM), el disco compacto de lectura/escritura (CD-R/W) y el DVD.

Un sistema de procesamiento de datos adecuado para almacenar y/o ejecutar código de programa incluirá al menos un procesador acoplado directa o indirectamente a elementos de memoria a través de un bus de sistema. Los elementos de memoria pueden incluir memoria local empleada durante la ejecución real del código de programa, almacenamiento masivo y memorias caché que proporcionan almacenamiento temporal de al menos parte del código de programa con el fin de reducir el número de veces que el código debe ser recuperado del almacenamiento masivo durante la ejecución.

Los dispositivos de entrada/salida (E/S) (incluyendo pero sin limtación, teclados, pantallas, dispositivos señaladores, etc.) pueden ser acoplados al sistema directamente o a través de controladores de E/S intervinientes. Los adaptadores de red también pueden acoplarse al sistema para permitir que el sistema de procesamiento de datos se acople a otros sistemas de procesamiento de datos o a impresoras o dispositivos de almacenamiento remotos a través de redes privadas o públicas intermedias. Los módems, los módems por cable y las tarjetas Ethernet son sólo algunos de los tipos de adaptadores de red disponibles actualmente.

Un entorno de hardware representativo para la puesta en práctica de las realizaciones puede incluir una configuración de hardware de un sistema de tratamiento de la información/ordenador de acuerdo con las realizaciones de la presente memoria. El sistema en cuestión comprende al menos un procesador o unidad central de procesamiento (CPU). Las CPUs están interconectadas a través del bus del sistema a varios dispositivos, como una memoria de acceso aleatorio (RAM), una memoria de sólo lectura (ROM) y un adaptador de entrada/salida (E/S). El adaptador de E/S puede conectarse a dispositivos periféricos, como unidades de disco y unidades de cinta, o a otros dispositivos de almacenamiento de programas que puedan ser leídos por el sistema. El sistema puede leer las instrucciones inventivas en los dispositivos de almacenamiento de programas y seguir estas instrucciones para ejecutar la metodología de las realizaciones de la presente memoria.

El sistema incluye además un adaptador de interfaz de usuario que conecta un teclado, un ratón, un altavoz, un micrófono y/u otros dispositivos de interfaz de usuario, como un dispositivo de pantalla táctil (no mostrado), al bus para recoger la entrada del usuario. Además, un adaptador de comunicación conecta el bus a una red de procesamiento de datos, y un adaptador de visualización conecta el bus a un dispositivo de visualización que puede ser incorporado como un dispositivo de salida como un monitor, una impresora o un transmisor, por ejemplo.

La descripción anterior se ha presentado con referencia a varias realizaciones.

REIVINDICACIONES

1. Un procedimiento de identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta, el procedimiento comprende:

recibir la muestra metagenómica mediante un módulo de recogida de muestras metagenómicas (124);

secuenciar la muestra metagenómica mediante un secuenciador (126);

obtener al menos una de las bases de datos de unidades taxonómicas operativas convencionales (OTUs) (104) y una base de datos de secuencias de referencia convencionales (106), en el que la base de datos de OTUs convencionales tiene una pluralidad de secuencias de nucleótidos agrupadas en una o más de las unidades taxonómicas operativas convencionales (OTUs) y los clados taxonómicos convencionales;

crear, mediante un procesador (110), una base de datos de OTU personalizada (OTUX) utilizando una pluralidad de segmentos predefinidos de secuencias de nucleótidos de la al menos una de las bases de datos de OTU convencionales y de la base de datos de secuencias de referencia convencionales, en la que los segmentos predefinidos de secuencias de nucleótidos se agrupan en OTUs personalizadas utilizando una técnica de agrupación de secuencias;

calcular, mediante el procesador (110), la propensión de una OTU personalizada a partir de la base de datos de OTUs personalizadas (OTUX) utilizando una fórmula predefinida, en la que la fórmula predefinida comprende: (un número de segmentos predefinidos de secuencias agrupados en la OTU personalizada correspondiente a la base de datos de OTU personalizada (OTUX) cuyos homólogos de longitud completa están asignados a una OTU convencional o a un clado taxonómico convencional presente en una base de datos de OTU convencional) / (número total de segmentos predefinidos de secuencias agrupados en la OTU personalizada correspondiente a la base de datos de OTU personalizada (OTUX)), en el que la propensión se refiere a una probabilidad de que la OTU personalizada esté asociada a uno o más clados taxonómicos convencionales en la base de datos de secuencias de referencia convencional y a las OTUs convencionales en la base de datos de OTUs convencionales;

crear, mediante el procesador (110), una matriz de mapeo que enumere todos los valores de las propensiones para cada una de las OTUs personalizadas con respecto a uno o más clados taxonómicos convencionales y a las OTUs convencionales;

utilizar, por el procesador (110), la base de datos de OTUs personalizadas (OTUX) como base de datos de referencia para la selección de OTU de referencia abierta para clasificar las secuencias de amplicones de lectura corta obtenidas de la muestra metagenómica secuenciada correspondientes a segmentos predefinidos en OTUs personalizadas adecuadas;

construir, mediante el procesador (110), una tabla de abundancia que represente la proporción de las secuencias de amplicones de lectura corta clasificadas en cada una de las OTUs personalizadas, ya sea acumulando el número total de lecturas secuenciadas de la muestra metagenómica que se clasifica en las OTUs personalizadas o representando la estructura de la comunidad microbiana perteneciente a la muestra metagenómica, en la que la abundancia de cada OTU microbiana se representa en términos de valores de abundancia normalizados en porcentaje; y

mapear de nuevo, por el procesador (110), las OTUs personalizadas utilizando la matriz de mapeo para representar la proporción de las secuencias de amplicones de lectura corta clasificadas en las OTUs personalizadas en términos de una o más de las OTUs convencionales y de los clados taxonómicos convencionales, en el que la tabla de abundancia representa una mayor precisión de la clasificación de las unidades taxonómicas operativas (OTUs) en la muestra metagenómica.

2. El procedimiento según la reivindicación 1 comprende además representar la abundancia de los grupos taxonómicos presentes en la tabla de abundancia en forma de un nivel taxonómico utilizando la información de la jerarquía taxonómica asociada con al menos uno de las OTUs personalizadas y las OTUs convencionales mapeadas, la representación de los grupos taxonómicos en forma de niveles taxonómicos permite la identificación de la presencia de grupos taxonómicos específicos en la muestra metagenómica, lo que facilita la comparación cruzada de las estructuras comunitarias del microbioma muestreadas a través de diferentes estudios metagenómicos desconectados, en los que se utilizan diferentes tecnologías de secuenciación, así como diferentes secuencias de marcadores o amplicones en los diferentes estudios metagenómicos desconectados.

3. El procedimiento según la reivindicación 1, en el que el etapa de crear una base de datos de clústeres de OTUs personalizada comprende:

recuperar todas las secuencias no alineadas de la base de datos convencional de OTU y de la base de datos convencional de secuencias de referencia;

extraer el segmento predefinido de un gen marcador para cada una de las secuencias no alineadas presentes en la base de datos convencional, en el que el gen marcador son las regiones de secuencias de nucleótidos obtenidas de la muestra metagenómica;

agrupar las secuencias extraídas basándose en un umbral de similitud predefinido utilizando la técnica de agrupación de secuencias; y

compilar las secuencias agrupadas para constituir la base de datos de clústeres de OTU personalizada.

4. El procedimiento según la reivindicación 1, en el que la base de datos de referencia convencional es una de las bases de datos Greengenes, SILVA o RDP que contiene secuencias del gen 16S rRNA de longitud completa.

5. El procedimiento según la reivindicación 1, en el que el gen marcador es el gen 16S rRNA procariota que tiene nueve regiones V1 a V9.

6. El procedimiento según la reivindicación 1, en el que la secuencia de amplicón de lectura corta se utiliza como una consulta contra la base de datos de OTU personalizada correspondiente a una región hipervariable, en la que la región hipervariable es la región V4 del gen 16S rRNA.

7. Un sistema de identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta, el sistema comprende:

al menos una de las bases de datos convencionales OTUs (104) y una base de datos convencional de secuencias de referencia (106) que tiene una pluralidad de secuencias de nucleótidos agrupadas en una o más de las unidades taxonómicas operativas convencionales (OTUs) y los clados taxonómicos convencionales;

un módulo (124) de recogida de muestras metagenómicas para recibir la muestra metagenómica;

un secuenciador (126) para secuenciar la muestra metagenómica recibida;

una memoria (108); y

un procesador (110) en comunicación con la memoria (108), en el que el procesador está configurado para realizar las etapas de:

crear una base de datos de OTU personalizada (OTUX) utilizando una pluralidad de segmentos predefinidos de secuencias de nucleótidos de la al menos una de las bases de datos de OTU convencionales y de la base de datos de secuencias de referencia convencionales, en la que los segmentos predefinidos de secuencias de nucleótidos se agrupan en OTUs personalizadas utilizando una técnica de agrupación de secuencias;

calcular la propensión de una OTU personalizada de la base de datos de OTU personalizada (OTUX) utilizando una fórmula predefinida, en el que la fórmula predefinida comprende: (un número de segmentos predefinidos de secuencias agrupados en la OTU personalizada correspondiente a la base de datos de OTU personalizada (OTUX) cuyos homólogos de longitud completa están asignados a una OTU convencional o a un clado taxonómico convencional presente en una base de datos de OTU convencional) / (número total de segmentos predefinidos de secuencias agrupados en la OTU personalizada correspondiente a la base de datos de OTU personalizada (OTUX)), en el que la propensión se refiere a una probabilidad de que la OTU personalizada esté asociada a uno o más clados taxonómicos convencionales en la base de datos de secuencias de referencia convencional y a las OTUs convencionales en la base de datos de OTUs convencionales;

crear una matriz de mapeo que enumere todos los valores de las propensiones para cada una de las OTUs personalizadas con respecto a uno o más clados taxonómicos convencionales y a las OTUs convencionales;

utilizar la base de datos de OTU personalizadas (OTUX) como base de datos de referencia para la selección de OTU de referencia abierta para clasificar las secuencias de amplicones de lectura corta obtenidas de la muestra metagenómica secuenciada correspondientes a segmentos predefinidos en OTU personalizadas adecuadas;

construir una tabla de abundancia que represente la proporción de las secuencias de amplicones de lectura corta clasificadas en cada una de las OTUs personalizadas, ya sea acumulando el número total de lecturas secuenciadas de la muestra metagenómica que se clasifica en las OTUs personalizadas o representando la estructura de la comunidad microbiana perteneciente a la muestra metagenómica, en la que la abundancia de cada OTU microbiana se representa en términos de valores de abundancia normalizados en porcentaje; y

mapear de nuevo las OTUs personalizadas utilizando la matriz de mapeo para representar la proporción de las secuencias de amplicones de lectura corta clasificadas en las OTUs personalizadas en términos de una o más OTUs convencionales y clados taxonómicos convencionales, en el que la tabla de abundancia representa una mayor precisión de la clasificación de las unidades taxonómicas operativas (OTUs) en la muestra metagenómica.

8. El sistema según la reivindicación 7, en el que el procesador está configurado para representar la abundancia de los grupos taxonómicos presentes en la tabla de abundancia en forma de un nivel taxonómico utilizando la información de la jerarquía taxonómica asociada con al menos una de las OTUs personalizadas y las OTUs convencionales mapeadas, la representación de los grupos taxonómicos en forma de niveles taxonómicos permite la identificación de la presencia de grupos taxonómicos específicos en la muestra metagenómica, lo que facilita la comparación cruzada de las estructuras comunitarias del microbioma muestreadas a través de diferentes estudios metagenómicos desconectados, en los que se utilizan diferentes tecnologías de secuenciación, así como diferentes secuencias de marcadores o amplicones en los diferentes estudios metagenómicos desconectados.

9. El sistema según la reivindicación 7, en el que el procesador está configurado para utilizar la secuencia de amplicón de lectura corta como una consulta contra la base de datos OTU personalizada correspondiente a una región hipervariable, en la que la región hipervariable es la región V4 del gen 16S rRNA.

10. Un medio no transitorio legible por ordenador que tiene incorporado en el mismo un programa de ordenador que comprende instrucciones que, cuando el programa es ejecutado por un ordenador, hacen que el ordenador lleve a cabo un procedimiento para la identificación y clasificación de unidades taxonómicas operativas (OTUs) en una muestra metagenómica utilizando secuencias de amplicones de lectura corta, el procedimiento comprende:

obtener al menos una de las bases de datos de unidades taxonómicas operativas convencionales (OTU) (104) y una base de datos de secuencias de referencia convencionales (106), en el que la base de datos de OTUs convencionales tiene una pluralidad de secuencias de nucleótidos agrupadas en una o más de las unidades taxonómicas operativas convencionales (OTUs) y los clados taxonómicos convencionales;

crear, mediante un procesador (110), una base de datos de OTU personalizada (OTUX) utilizando una pluralidad de segmentos predefinidos de secuencias de nucleótidos de la al menos una de las bases de datos de OTUs convencionales y de la base de datos de secuencias de referencia convencionales, en la que los segmentos predefinidos de secuencias de nucleótidos se agrupan en OTUs personalizadas utilizando una técnica de agrupación de secuencias;

calcular, mediante el procesador (110), la propensión de una OTU personalizada a partir de la base de datos de OTU personalizadas (OTUX) utilizando una fórmula predefinida, en la que la fórmula predefinida comprende: (número de segmentos predefinidos de secuencias agrupados en la OTU personalizada correspondiente a la base de datos de OTU personalizada (OTUX) cuyos homólogos de longitud completa están asignados a una OTU convencional o a un clado taxonómico convencional presente en una base de datos de OTU convencional) / (número total de segmentos predefinidos de secuencias agrupados en la OTU personalizada correspondiente a la base de datos de OTU personalizada (OTUX)), en el que la propensión se refiere a una probabilidad de que una OTU personalizada esté asociada a uno o más clados taxonómicos convencionales en la base de datos de secuencias de referencia convencional y a las OTUs convencionales en la base de datos de OTU convencionales;

crear, mediante el procesador (110), una matriz de mapeo que enumere todos los valores de las propensiones para cada una de las OTUs personalizadas con respecto a uno o más clados taxonómicos convencionales y a las OTUs convencionales;

utilizando, por el procesador (110), la base de datos OTUs personalizadas (OTUX) como base de datos de referencia para la selección de OTU de referencia abierta para clasificar las secuencias de amplicones de lectura corta obtenidas de la muestra metagenómica secuenciada correspondientes a segmentos predefinidos en OTUs personalizadas adecuadas;

la construcción, por parte del procesador (110), de una tabla de abundancia que represente la proporción de las secuencias de amplicones de lecturas cortas clasificadas en cada una de las OTUs personalizadas, ya sea acumulando el número total de lecturas secuenciadas de la muestra metagenómica que se clasifica en las OTUs personalizadas o representando la estructura de la comunidad microbiana perteneciente a la muestra metagenómica, en la que la abundancia de cada OTU microbiana se representa en términos de valores de abundancia normalizados en porcentaje;

mapear de nuevo las OTUs personalizadas utilizando la matriz de mapeo para representar la proporción de las secuencias de amplicones de lectura corta clasificadas en las OTUs personalizadas en términos de una o más de las OTUs convencionales y los clados taxonómicos convencionales, en el que la tabla de abundancia representa una mayor precisión de la clasificación de las unidades taxonómicas operativas (OTUs) en la muestra metagenómica.

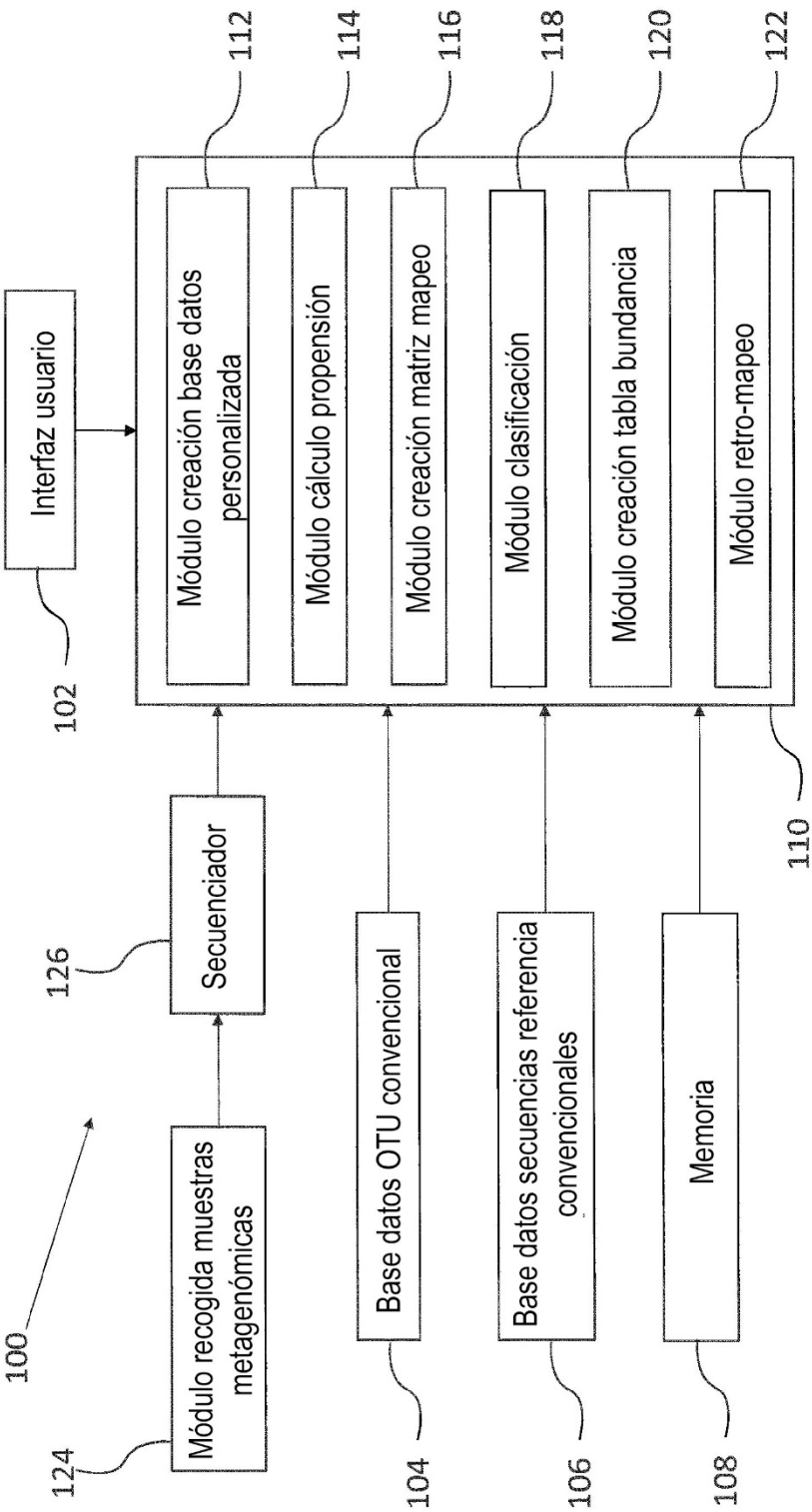
100
102
Interfaz usuario
110
112
Módulo creación base datos personalizada
114
Módulo cálculo propensión
116
Módulo creación matriz mapeo
118
Módulo clasificación
120
Módulo creación tabla bundancia
122
Módulo retro-mapeo
124
Módulo recogida muestras metagenómicas
126
Secuenciador
104
Base datos OTU convencional
106
Base datos secuencias referencia convencionales
108
Memoria

FIG. 1

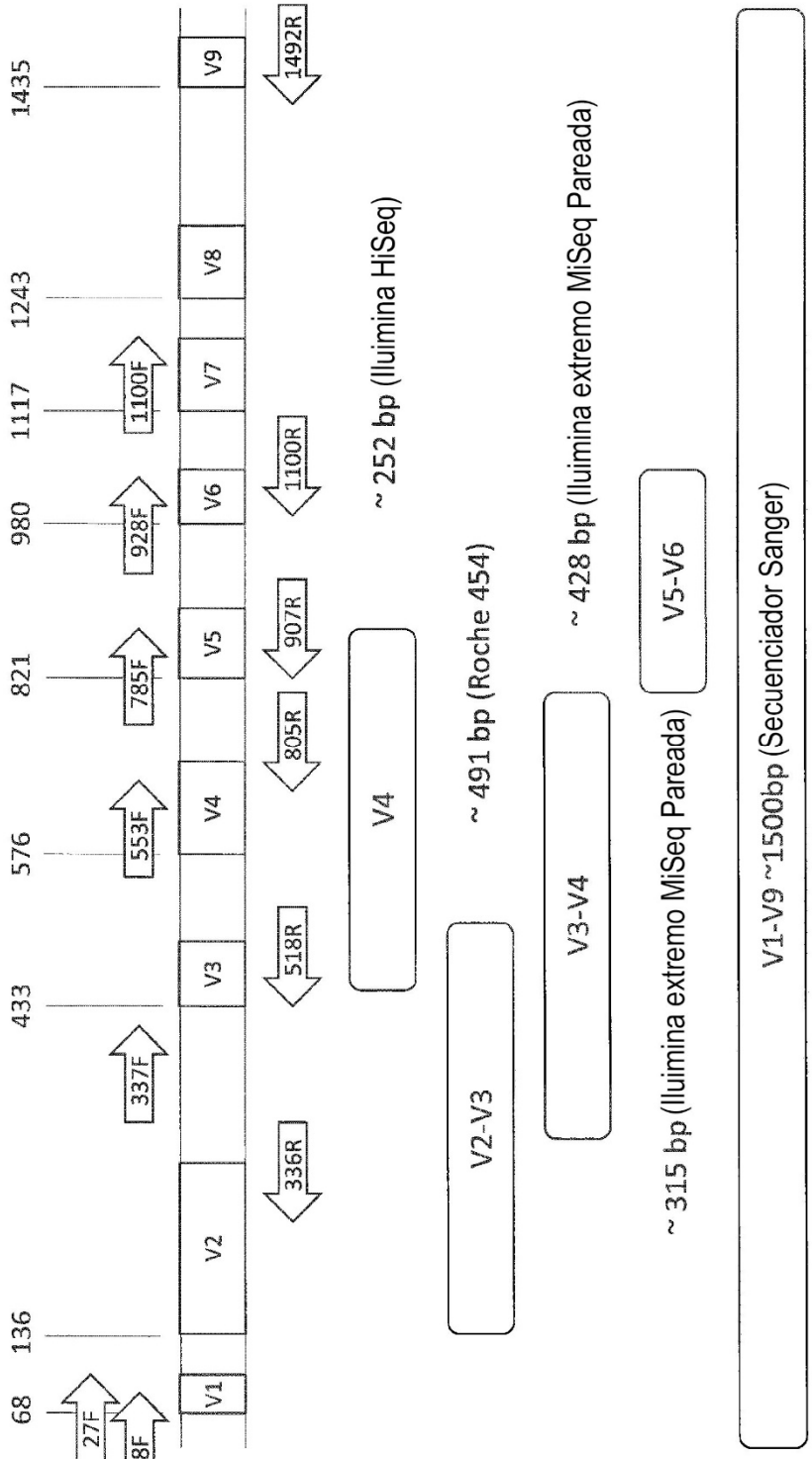
68
136
433
576
821
980
1117
1243
1435
27F
8F
337F
553F
785F
928F
1100F
V1
V2
V3
V4
V5
V6
V7
V8
V9
336R
518R
805R
907R
1100R
1492R
V4
~ 252 bp (Iluimina HiSeq)
V2-V3
~ 491 bp (Roche 454)
V3-V4
~ 428 bp (Iluimina extremo MiSeq Pareada)
V5-V6
~ 315 bp (Iluimina extremo MiSeq Pareada)
V1-V9 ~1500bp (Secuenciador Sanger)

FIG. 2

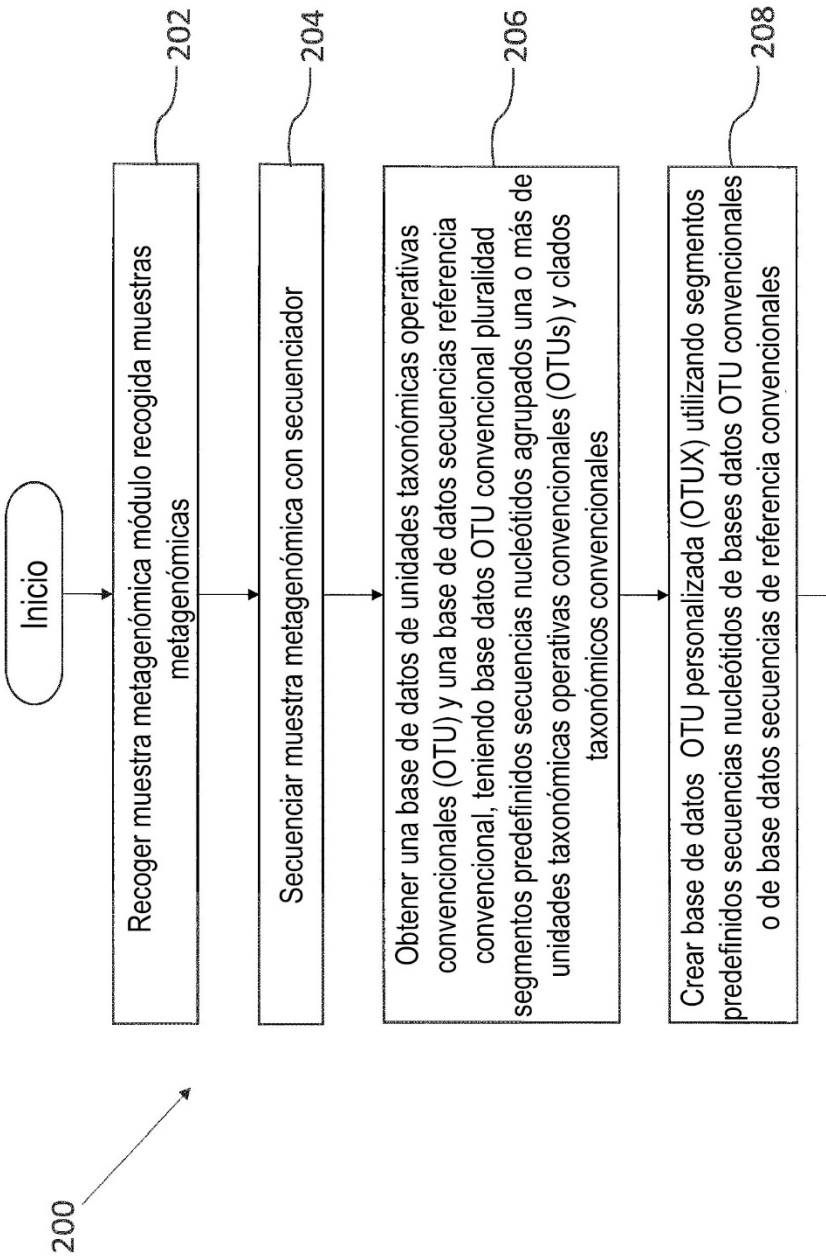
200
Inicio
Recoger muestra metagenómica módulo recogida muestras metagenómicas
202
Secuenciar muestra metagenómica con secuenciador
204
Obtener una base de datos de unidades taxonómicas operativas convencionales (OTU) y una base de datos secuencias referencia convencional, teniendo base datos OTU convencional pluralidad segmentos predefinidos secuencias nucleótidos agrupados una o más de unidades taxonómicas operativas convencionales (OTUs) y clados taxonómicos convencionales
206
Crear base de datos OTU personalizada (OTUX) utilizando segmentos predefinidos secuencias nucleótidos de bases datos OTU convencionales o de base datos secuencias de referencia convencionales
208
A

FIG. 3a

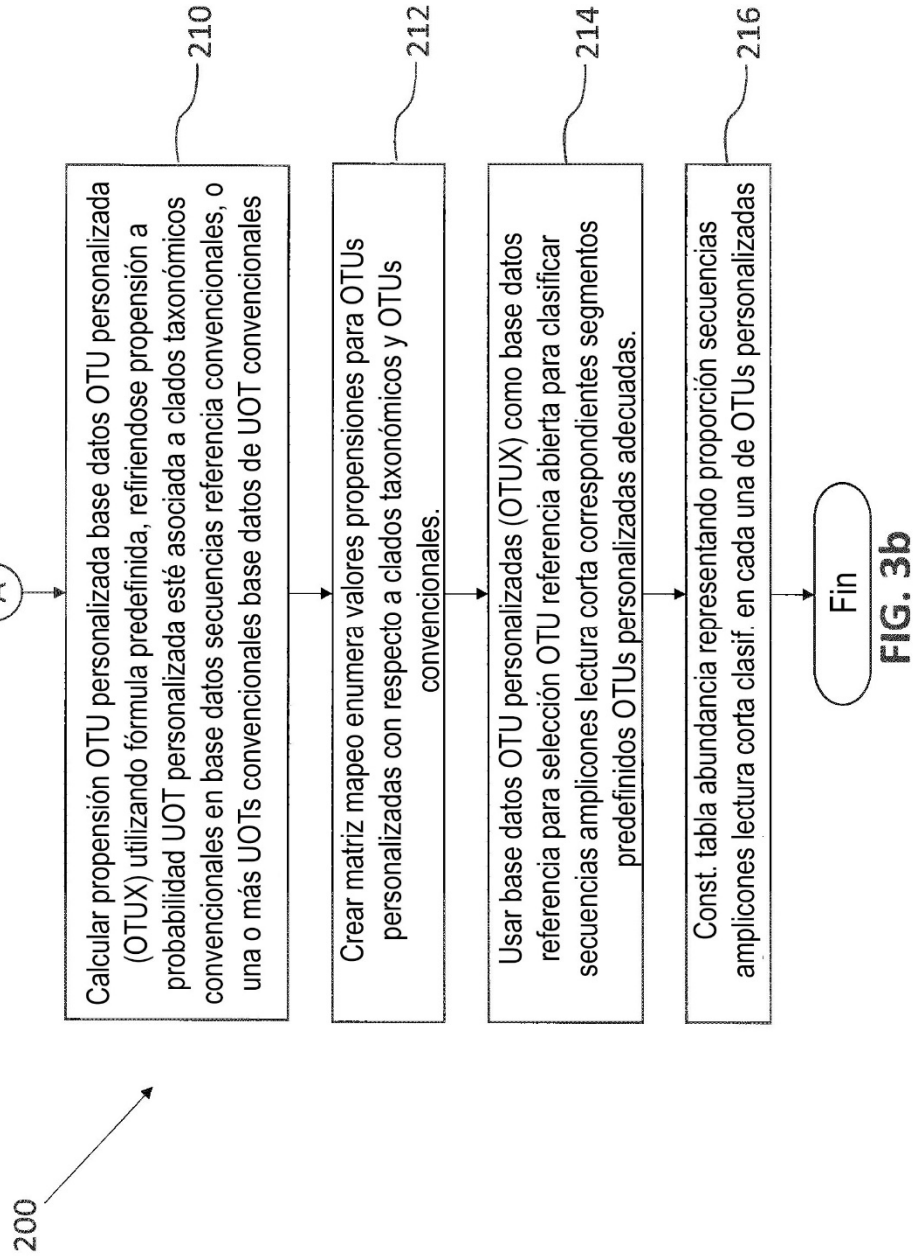
200
A
Calcular propensión OTU personalizada base datos OTU personalizada (OTUX) utilizando fórmula predefinida, refiriendose propensión a probabilidad UOT personalizada esté asociada a clados taxonómicos convencionales en base datos secuencias referencia convencionales, o una o más UOTs convencionales base datos de UOT convencionales
210
Crear matriz mapeo enumera valores propensiones para OTUs personalizadas con respecto a clados taxonómicos y OTUs convencionales.
212
Usar base datos OTU personalizadas (OTUX) como base datos referencia para selección OTU referencia abierta para clasificar secuencias amplicones lectura corta correspondientes segmentos predefinidos OTUs personalizadas adecuadas.
214
Const. tabla abundancia representando proporción secuencias amplicones lectura corta clasif. en cada una de OTUs personalizadas
216
Fin

FIG. 3b