(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 899 879**

(51) Int. CI.:

G16B 30/00 (2009.01)
G16B 40/00 (2009.01)
G16B 50/00 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **25.01.2012 PCT/US2012/022513**

(87) Fecha y número de publicación internacional: **02.08.2012 WO12103189**

(96) Fecha de presentación y número de la solicitud europea: **25.01.2012 E 12739779 (2)**

(97) Fecha y número de publicación de la concesión europea: **01.09.2021 EP 2668320**

(54) Título: **Identificación y medición de poblaciones relativas de microorganismos con secuenciación directa de ADN**

(30) Prioridad:

26.01.2011 US 201113014112

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
15.03.2022

(73) Titular/es:

COSMOSID INC. (100.0%)
387 Technology Drive, Suite 3119
College Park, MD 20742, US

(72) Inventor/es:

COLWELL, RITA, R.;
LIVINGSTON, BOYD, THOMAS;
JAKUPCIAK, DAVID;
HASAN, NUR, A.;
JAKUPCIAK, JOHN, P. y
BRENNER, DOUGLAS, M.

(74) Agente/Representante:

SÁEZ MAESO, Ana

DESCRIPCIÓN

Identificación y medición de poblaciones relativas de microorganismos con secuenciación directa de ADN

Antecedentes

Campo de la invención

Esta invención se refiere a métodos para la caracterización de organismos y más particularmente, a la caracterización de las identidades y poblaciones relativas de organismos en una muestra de acuerdo con las reivindicaciones.

Discusión de los antecedentes

Los métodos de ácido nucleico actuales para identificar poblaciones de organismos son específicos solo al nivel de género (ADNr 16S) para bacterias, virus y otros organismos microbianos y no identifican las poblaciones hasta las especies, subespecies y cepas de organismos dentro de la muestra. Las técnicas actuales para detectar e identificar una o una pequeña cantidad de bacterias a nivel de género, especie y/o subespecie se basan en métodos estáticos, tales como la reacción en cadena de la polimerasa (PCR) y matrices de microchip, para detectar firmas de un organismo preespecificado o una pluralidad de organismos preespecificados. Los métodos actuales para detectar e identificar virus requieren pruebas de ácido nucleico específicas y no producen datos sobre la diversidad de la población. Se ha utilizado la secuenciación completa de genomas de virus para demostrar la diversidad de virus. Los métodos convencionales en general emplean métodos de cultivo de laboratorio para bacterias, hongos y parásitos, y son costosos y requieren mucho tiempo.

En ninguno de estos casos, las poblaciones relativas pueden ser determinadas con precisión ni ninguno de estos métodos es capaz de detectar e identificar simultáneamente los organismos presentes en las poblaciones microbianas con respecto a los taxones específicos (género, especie, subespecie y cepa) de bacterias, virus, parásitos, hongos o fragmentos de ácido nucleico, incluidos plásmidos y componentes genómicos móviles. Además, dada la rápida tasa de mutación genómica y la creciente evidencia de transferencia horizontal de genes, los métodos estáticos que se basan en firmas predeterminadas producen falsos resultados negativos si (a) se ha producido una mutación en la secuencia de ácido nucleico de la muestra con respecto a la firma, (b) la firma diana se transfirió horizontalmente, o (c) los vecinos cercanos genómicos están presentes en la muestra.

En el diagnóstico de enfermedades infecciosas, la microbiología convencional todavía se basa en métodos de cultivo laboriosos y que consumen mucho tiempo y pruebas engorrosas para bacterias, virus, parásitos y hongos, y también en pruebas inmunológicas y de ácido nucleico molecular. Los ensayos de ácido nucleico adicionales se utilizan, generalmente, para determinar la presencia de bacterias, virus o bacterias no cultivables específicas. En una fracción significativa de las muestras y hasta el 25 % de todas las muestras, no se identifica ningún agente causal identificable específico de los síntomas del paciente. Además, generalmente se asume que una enfermedad infecciosa siempre es causada por un solo agente microbiano o múltiples agentes, induciendo colectivamente los síntomas observados, cuando están presentes más de una o unas pocas células del agente.

Las poblaciones bacterianas de fondo o microbiomas (bacterias), micobiomas (hongos) y viromas (virus), a los niveles de especies y cepas, no pueden determinarse rápida o fácilmente (es decir, en horas y con un solo método o menos) por los métodos actuales. Sin embargo, para determinar la causa de la enfermedad puede ser necesario normalizar los resultados con respecto a las poblaciones de fondo, pero los métodos actuales carecen de la capacidad para hacerlo. En la ciencia de los alimentos, estas comparaciones relativas con los antecedentes microbianos, hasta el nivel de subespecies y/o cepas, son necesarios para determinar la fuente de contaminación de los alimentos y el grado de patogenicidad.

Por ejemplo, las cepas benignas de Escherichia coli son abundantes en la naturaleza, pero tales cepas pueden mutar, adquirir genes que codifican propiedades patogénicas y/o producción de toxinas y volverse toxigénicas (por ejemplo, E. coli O157: H7). Recientemente, se han identificado seis (6) nuevas cepas patógenas de E. coli no O157 (O111, 0121, O26, O45, 0103 y 0145). Estas nuevas cepas patógenas de E. coli son mucho menos conocidas que E. coli 0157, pero sólo son capaces de causar enfermedades graves, incluida la insuficiencia renal, que suele ser mortal. Estas nuevas cepas patógenas de E. coli son mucho más difíciles de identificar en un microbioma utilizando métodos convencionales porque, aunque estas nuevas cepas se han identificado a partir de sus genomas de ADN completos, no se han desarrollado pruebas que utilicen tecnología de métodos estáticos convencionales y que impliquen una firma genética.

Dada la frecuencia y propensión de la mutación genética en la naturaleza, es probable que cepas patógenas adicionales de E. coli y otras bacterias continúen desarrollándose y evolucionando y, por lo tanto, causen enfermedades. Tal mutación genética continua, un fenómeno que ocurre naturalmente, requiere un método universal para facilitar la identificación incluso cuando se ha producido una mutación. Por lo tanto, existe la necesidad en la técnica tanto de un método universal capaz de identificación microbiana a nivel de especies y cepas como de un método que tenga en cuenta la biodiversidad y la mutación. Los pares de bases de ADN y ARN caracterizan a todos

los organismos vivos, incluidos los microorganismos y los fragmentos de ácido nucleico como los plásmidos, y la secuenciación directa de ADN es el estándar para la identificación de pares de bases de ADN, existe una necesidad en la técnica de la identificación genómica universal a nivel de subespecie y cepa mediante secuenciación directa de ADN y ARN de muestras metagenómicas.

Además, monitorizar poblaciones de microorganismos en el ambiente (por ejemplo, en el suministro de agua) y rastrearlas hasta enfermedades infecciosas en pacientes (por ejemplo, cólera) requiere especificidad de identificación a nivel de subespecie y/o cepa para diagnosticar la enfermedad y su fuente. Además, se requiere el análisis de los microbiomas en la naturaleza para comprender la resistencia a los antibióticos y para monitorizar y prevenir brotes epidémicos o pandemias. Dado que los microorganismos son omnipresentes y muchos, si no la mayoría, existen tanto en formas ambientalmente amigables (comensales no toxigénicos) como también en formas que son una bioamenaza para los humanos (patógeno altamente toxigénico y/o invasivo), ellos no se pueden erradicar por completo; la única forma de minimizar o prevenir la infección es minimizar la exposición a formas patógenas de microbios cuando sus concentraciones son altas, e identificar y rastrear especies y cepas patógenas específicas que infectan a los pacientes.

El análisis de la herramienta básica de búsqueda de alineación local (BLAST) se ha convertido en un método omnipresente para interrogar datos de secuencia. Se han desarrollado muchos métodos de búsqueda de datos que se basan en mejoras de BLAST. Estos incluyen sistemas y métodos para generar índices y búsquedas rápidas de Coincidencias “aproximadas”, “difusas” u “homólogas” (coincidencia perfecta) para una gran cantidad de datos. Los datos se indexan para generar una taxonomía de árbol de búsqueda. Una vez que se genera el índice, se puede proporcionar una consulta para reportar los aciertos dentro de una cierta vecindad de la consulta. En BLAST, se usa una distancia local de un espacio de secuencia local para generar ramas de árbol de búsqueda local.

Sin embargo, existen limitaciones para usar los valores E de salida BLAST, que describen el número de aciertos que uno puede esperar ver por casualidad cuando se busca en una base de datos de un tamaño particular y se usan para medir la importancia de una coincidencia, como criterio para el análisis de datos. Si bien esta medición es posible, el resultado suele estar sesgado tanto por la base de datos utilizada para la comparación como por la longitud de la coincidencia. Las regiones pequeñas de alta similitud pueden generar un valor E artificialmente bajo y anular el nivel global de similitud exhibido por la secuencia. El valor de la puntuación BLAST varía con la longitud del nucleótido consultado y, por lo tanto, no es adecuado solo para el análisis comparativo utilizando puntos de corte universales.

Anteriormente, la aplicación directa de la secuenciación para diagnósticos rápidos y multiplex no había sido posible. El análisis directo de muestras se consideró demasiado complejo para interpretar y se emplean métodos selectivos (por ejemplo, cultivo) para minimizar el número de organismos (principalmente a un tipo) para su análisis. La capacidad de detectar todos los patógenos utilizando una única plataforma no ha sido posible. La biodefensa, la protección de la fuerza, la agricultura y la salud mundial se beneficiarán de la identificación basada en secuencias de todos los patógenos en una muestra y la elaboración de perfiles de patógenos para la toma de decisiones médicas.

Las técnicas de identificación de patógenos que no se basan necesariamente en métodos de cultivo convencionales incluyen técnicas inmunológicas, mediante las cuales se detectan moléculas exclusivas del patógeno (generalmente proteínas) utilizando anticuerpos que se unen específicamente a moléculas únicas, y una variedad de técnicas que se dirigen a secuencias de ADN o de ARN específicas, conocidas colectivamente como técnicas de ácido nucleico (NAT) o técnicas de diagnóstico molecular. Los métodos inmunológicos y NAT actuales son útiles para reconocer una gama limitada de patógenos en condiciones altamente específicas, pero cada uno de estos métodos está sujeto a deficiencias.

Se sabe que las técnicas inmunológicas, o inmunoensayos, adolecen de varias debilidades críticas que limitan su eficacia en el diagnóstico médico. Estas incluyen el volumen y la especificidad de los reactivos, la reactividad cruzada y la pobre inmunogenicidad de algunos organismos, entre las deficiencias. Por ejemplo, a menudo es difícil producir anticuerpos que reaccionen específicamente con el patógeno diana sin reaccionar con otros patógenos (es decir, reactividad cruzada). Hay muchos patógenos que varían las moléculas en sus superficies (por ejemplo, Niesseria gonorrheae), lo que hace imposible detectar a todos los miembros de un grupo diana determinado. Muchos formatos de inmunoensayos, como las pruebas de aglutinación rápida para el estreptococo tipo A, requieren una gran cantidad de organismos para su detección. Esto dificulta la detección temprana de infecciones o requiere el cultivo de los microorganismos antes de la detección inmunológica.

Los métodos de ácido nucleico son mucho más específicos que los métodos inmunológicos porque se dirigen al material genético del patógeno. Casi todos los métodos NAT requieren la amplificación del ácido nucleico diana mediante la reacción en cadena de la polimerasa (PCR). Las limitaciones de la PCR incluyen: (a) la biblioteca de cebadores de ADN para reconocer secuencias en genomas de patógenos es limitada; (b) los mutantes, las cepas y los patógenos modificados no siempre se detectan fácilmente, si es que se detectan; (c) debido a las bibliotecas limitadas de cebadores de ADN, hay poca o ninguna redundancia de reconocimiento para excluir reacciones falsas positivas o negativas; (d) se produce la erosión del cebador/firma; y (e) no se pueden reconocer patógenos desconocidos porque el reconocimiento de cualquier patógeno requiere conocimiento previo de la secuencia de nucleótidos del material genético del patógeno en particular.

Se han utilizado métodos basados en secuenciación para el análisis del genoma completo, pero no para caracterizar e identificar poblaciones de microorganismos ni como herramienta predictiva y forense para la toma de decisiones. Por ejemplo, se han desarrollado métodos para identificar especies y subespecies en una muestra biológica mediante la amplificación selectiva de segmentos de ácido nucleico. Dichos métodos utilizan un cebador o código para una región diana específica (generalmente un gen, genes o fragmentos de genes, incluido el ADN mitocondrial) presente en una fracción diminuta de todas las poblaciones de una muestra. Los métodos implican extracción de ADN de una muestra, amplificación de segmentos divergentes de la diana por PCR o una técnica equivalente, utilizando cebadores de regiones con alta conservación evolutiva entre especies y subespecies, análisis del segmento amplificado mediante comparación de su tamaño en pares de bases con un estándar preestablecido de tamaños y/o análisis del segmento amplificado mediante comparación de la secuencia de ADN de la secuencia resultante con un subconjunto de secuencias específicas de fracciones de un grupo de especies o subespecies consultadas en una base de datos de ordenador.

Estos métodos se han utilizado para el análisis genético de una especie biológica empleando una muestra (material biológico) derivada de aislados únicos o de muestras que contienen mezclas duales o heterogéneas. Amplificación de una región de ADN de la muestra, correspondiente a una(s) posición(es) determinadas y de genoma estrecho, se realiza para determinar el tamaño en pares de bases y/o la secuencia de ADN precisa seguida de mapeo de esa región a través de identificación taxonómica. El mapeo se realiza contra una base de datos de referencia de organismos de regiones amplificadas que contienen tamaños preestablecidos y/o secuencias de ADN de la región correspondiente de una pluralidad de especies y/o subespecies.

Todos estos métodos tienen un uso limitado cuando una muestra comprende una mezcla de organismos. Solo pueden confirmar la presencia de un organismo previamente conocido o sospechoso, pero no pueden identificar cada uno de los organismos presentes en la muestra y no pueden identificar a nivel de especie, subespecie y/o cepa. Además, si un organismo previamente conocido estuviera presente, pero hubiera sufrido una mutación en la secuencia preespecificada, tales métodos indicarían un falso negativo. Los procesos naturales de mutación, deleciones genéticas y las alteraciones o mutaciones manipuladas forman parte de la creación de biodiversidad que no se puede detectar o incluso abordar con los métodos de la técnica anterior existentes.

El documento US 2009/150084 describe un sistema de identificación del genoma. Rosen et al (2008), Advances in Bioinformatics: 16 de noviembre, páginas 1-12, describe la clasificación de fragmentos de metagenoma usando perfiles de frecuencia N-mero.

Por lo tanto, se necesita un mecanismo para identificar simultáneamente una pluralidad de organismos en una muestra dada con una sola prueba sin tener que utilizar múltiples sondas y sin conocimiento previo de los organismos presentes en la muestra. También es deseable distinguir especies, subespecies y cepas muy similares o interrelacionadas para aplicaciones médicas, agrícolas e industriales.

Existen muchas circunstancias potencialmente mortales en las que sería útil analizar y secuenciar el ADN y/o ARN en una muestra, por ejemplo, en respuesta a un acto de bioterrorismo en el que se hubiera liberado un agente patógeno fatal en el ambiente. En el pasado estos resultados han requerido la participación de muchas personas, lo que exige demasiado tiempo. Como resultado, puede afectarse la rapidez y la precisión.

Un ataque bioterrorista o una epidemia emergente requiere que los primeros en responder, es decir, los médicos en la sala de emergencias (sus opciones o tratamientos al lado de la cama), tomen decisiones inmediatas sobre el tratamiento, y los fabricantes de alimentos, distribuidores, minoristas y personal de salud pública en todo el país identifiquen de manera rápida, precisa y confiable los agentes patógenos y las enfermedades que ellos causan. Los agentes patógenos pueden transmitirse en los alimentos, el aire, el suelo, el agua y los tejidos animales, vegetales y humanos y mediante la presentación clínica en las salas de emergencia. Debido a que los agentes y/o enfermedades potenciales pueden poner en peligro la vida de inmediato y/o ser altamente contagiosas, la identificación debe ser rápida y precisa. Si esto no es posible, representa una debilidad significativa en el control de enfermedades infecciosas, seguridad nacional y respuesta al bioterrorismo.

Se necesita un método y un sistema para identificar de forma rápida y precisa más de un organismo individual (multiplexación) en una muestra e indicar si una especie, cepa y/o subcepa están presentes empleando la comparación del genoma de los ácidos nucleicos presentes en la muestra con los ácidos nucleicos presentes en una base de datos genómica de referencia. El documento US 2009/150084 A1 implica un método de secuenciación de material biológico y emparejamiento probabilístico en tiempo real de cadenas cortas de información de secuenciación para identificar especies presentes en dicho material biológico.

Los rápidos avances en la ingeniería biológica han impactado dramáticamente el diseño y las capacidades de las herramientas de secuenciación de ADN, incluida la secuenciación de alto rendimiento, un método para determinar el orden de las bases en el ADN y el mapeo de la variación genética que revela el sustento genético de la enfermedad en humanos. Este enfoque es útil cuando se secuencian muchas plantillas de ADN diferentes con cualquier número de cebadores. A pesar de estos importantes avances en ingeniería biológica, se ha avanzado poco en la construcción

de dispositivos para identificar rápidamente la información de secuencia y transferir datos de manera más eficiente y efectiva.

Tradicionalmente, la secuenciación de ADN se realizaba mediante un método didesoxi, comúnmente denominado método Sanger [Sanger et al, 1977], que usaba inhibidores de terminación de cadena para detener la extensión de la cadena de ADN durante la síntesis de ADN.

Los métodos para las estrategias de secuenciación continúan desarrollándose. Por ejemplo, es posible construir una matriz de secuencias de ADN (microarreglos) e hibridar secuencias complementarias en un proceso comúnmente conocido como secuenciación por hibridación. Otra técnica considerada estado del arte emplea la extensión del cebador, seguida de la adición cíclica de un solo nucleótido, con cada ciclo seguido de la detección del evento de incorporación. La técnica conocida como secuenciación por síntesis o pirosecuenciación, incluida la secuenciación fluorescente in situ (FISSEQ), es reiterativa en la práctica e implica un proceso en serie de ciclos repetidos de extensión del cebador mientras se secuencia la secuencia de nucleótidos diana. Estos métodos de secuenciación no pueden identificar rápidamente un organismo a partir de los datos de un aislado, y actualmente no existen herramientas para identificar una mezcla de organismos basado en datos metagenómicos creados por estos métodos de secuenciación. Además, los métodos y sistemas convencionales para identificar organismos en muestras metagenómicas basadas en datos de nucleótidos generados por secuenciadores no existen.

A pesar de estos avances, existe la necesidad de métodos y sistemas de identificación rápida del genoma, incluida la comunicación electrónica multidireccional de datos de secuencia de ácido nucleico, datos clínicos, intervención terapéutica y administración personalizada de terapias a poblaciones diana para agilizar las respuestas y acelerar el diagnóstico de enfermedades infecciosas, conservar suministros médicos valiosos y contener el bioterrorismo, la liberación involuntaria y las epidemias patógenas emergentes. Además, se necesita un mecanismo para identificar simultáneamente una pluralidad de organismos en una muestra determinada con una sola prueba sin tener que utilizar múltiples sondas, y es deseable distinguir especies, subespecies y cepas muy similares o interrelacionadas para aplicaciones médicas, agrícolas e industriales.

Resumen

La presente invención se refiere a sistemas y métodos capaces de caracterizar poblaciones de microorganismos dentro de una muestra de acuerdo con las reivindicaciones. La caracterización utiliza emparejamiento probabilístico para identificar genomas microbianos con los que se relacionan los fragmentos metagenómicos extraídos de la muestra. La caracterización incluye además la identificación de la comunidad microbiana de la muestra a nivel de especie y/o subespecie y/o cepa con sus concentraciones o abundancia relativas. Además, los sistemas y métodos permiten la identificación rápida de organismos, incluidos tanto patógenos como comensales en muestras clínicas, de alimentos de agua y ambientales, y la identificación puede lograrse mediante la comparación de muchos (por ejemplo, cientos a millones) fragmentos metagenómicos, que han sido capturados de una muestra y secuenciados, con muchos (por ejemplo, millones o miles de millones) de información de secuencia archivada de los genomas (es decir, bases de datos genómicas de referencia). Lograr la caracterización metagenómica precisa en los extremos superiores de estos espectros, es decir, comparar decenas de millones de fragmentos metagenómicos con bases de datos genómicas archivadas que comprenden miles de millones de nucleótidos, no está documentada previamente en la literatura.

En un aspecto, la presente invención proporciona un método de acuerdo con la reivindicación 1.

En una realización de la invención, los resultados probabilísticos pueden tener la forma de un mapa de probabilidad de probabilidades de que especies y/o subespecies y/o cepas de organismos contenidas en la base de datos de referencia estén presentes dentro de la muestra. El mapa de probabilidad puede permitir la correlación de las probabilidades del mapa de probabilidad con las poblaciones relativas y/o concentraciones de organismos contenidos en la muestra. El método puede compensar aún más el error de la máquina mediante el uso de una cantidad de lecturas de fragmentos metagenómicos estadísticamente significativos lo suficientemente grandes como para que los errores se normalicen. El error de la máquina para el que la normalización compensa puede comprender el error de máquina de un secuenciador utilizado para generar la pluralidad de lecturas de fragmentos metagenómicos, y la compensación puede comprender el uso de suficientes lecturas de fragmentos metagenómicos para que el error de máquina del secuenciador se normalice a un valor cercano a cero.

El método puede comprender además extraer los fragmentos metagenómicos de la muestra.

En otra realización más de la invención, el método comprende la contabilidad de la biodiversidad. La contabilidad de la biodiversidad puede comprender identificar: (a) elementos genéticos móviles mediante transferencia lateral de genes, recombinación o plásmido u otra inserción de mobiloma; (b) inserciones y deleciones, y (c) identificación y detección de cepas parientes cercanas (por ejemplo, primo) relacionadas por mutación, inserción y/o deleción.

La base de datos de referencia contiene las identidades genómicas de uno o más de la pluralidad de organismos contenidos en la muestra. Cada una de la pluralidad de lecturas de fragmentos metagenómicos puede tener una

longitud de lectura mayor o igual a 12 pares de bases y menor o igual a 100 pares de bases. Sin embargo, las lecturas de fragmentos metagenómicos que tienen una longitud superior a 100 pares de bases se pueden utilizar adicional o alternativamente. Los métodos probabilísticos comprenden, para cada una de la pluralidad de lecturas de fragmentos metagenómicos, detectar y retener las correlaciones causales entre la lectura de fragmento metagenómico y lecturas de genoma de la base de datos de referencia que contiene identidades genómicas de organismos; e integración de las correlaciones causales retenidas por cepa genómica y especie para identificar un conjunto de genomas de microorganismos contenidos en la muestra. Los métodos probabilísticos comprenden además la creación de conjuntos de patrones independientes de inclusión de subconjuntos y exclusión de subconjuntos del conjunto de genomas y, para cada conjunto de patrón independiente, emparejando el conjunto con lecturas diana. Cada uno de los emparejamientos da como resultado una estimación independiente de la concentración del genoma en el conjunto. Las estimaciones independientes pueden dar una estimación detallada de las concentraciones de cepas genómicas incluso para comunidades microbianas estrechamente relacionadas.

Los métodos probabilísticos comprenden el emparejamiento probabilístico. Los métodos probabilísticos comprenden: filtrado primario para determinar qué especies y cepas de la base de datos de referencia pueden estar en la muestra metagenómica; y filtrado secundario y terciario para eliminar tanto los falsos negativos como los falsos positivos y para identificar a nivel de cepa, qué organismos están contenidos en la muestra.

En una realización de la invención, los organismos identificados incluyen genomas de bacterias, virus, parásitos, hongos y/o fragmentos de ácido nucleico que incluyen plásmidos y componentes genómicos móviles. Los organismos identificadores contenidos en la muestra pueden ser capaces de identificar bacterias, virus, parásitos, hongos y fragmentos de ácido nucleico, incluidos plásmidos y componentes genómicos móviles contenidos en la muestra. La pluralidad de lecturas de fragmentos metagenómicos son lecturas de secuencias de fragmentos metagenómicos extraídos de la muestra. Los fragmentos metagenómicos extraídos de la muestra son de ácido nucleico genómico, proteína y/o una combinación con metabolitos extraídos de la muestra. Cada uno de los fragmentos metagenómicos extraídos de la muestra puede ser un fragmento de una secuencia de ácido nucleico. Cada uno de los fragmentos metagenómicos extraídos de la muestra puede ser un fragmento de una secuencia de ácido desoxirribonucleico (ADN). Cada uno de los fragmentos metagenómicos extraídos de la muestra puede ser una secuencia de ácido ribonucleico (ARN). Cada uno de los fragmentos metagenómicos extraídos de la muestra es un fragmento de un plásmido u otra unidad de secuencia de ácido nucleico.

En otra realización más de la invención, la pluralidad de lecturas de fragmentos metagenómicos se obtiene de la muestra: recolectando la muestra, extrayendo fragmentos metagenómicos de la muestra y secuenciando los fragmentos metagenómicos.

En otra realización de la invención, la pluralidad de lecturas de fragmentos metagenómicos que se pueden obtener de la muestra se incluye en un archivo metagenómico. El método comprende además la creación de una lista de palabras de referencia para cada una de la pluralidad de lecturas del genoma de la base de datos de referencia que contiene identidades genómicas de organismos y creación de un catálogo de listas de palabras de referencia. El método comprende la creación de una lista de palabras de referencia para cada una de la pluralidad de lecturas del genoma de la base de datos de referencia que contiene las identidades genómicas de los organismos, y la creación de un catálogo de listas de palabras de referencia. Cada lista de palabras de referencia es un género, especie o cepa. El método puede comprender además la creación de una lista de palabras de secuencia de muestra para cada una de la pluralidad de lecturas de fragmentos metagenómicos obtenidas de la muestra. La comparación de la muestra de lecturas de fragmentos metagenómicos obtenidas de la muestra con la pluralidad de lecturas del genoma de la base de datos de referencia que contiene las identidades genómicas de organismos pueden comprender: para cada una de las palabras de secuencia de muestra de la lista de palabras de secuencia de muestra, comparar la palabra de secuencia de muestra con las palabras de referencia de cada una de las listas de palabras de referencia, e identificar coincidencias entre la palabra de secuencia de muestra y una o más de las palabras de referencia. Las coincidencias identificadas pueden ser coincidencias exactas. Las coincidencias identificadas pueden comprender coincidencias inexactas. El método puede comprender, además: para cada una de la pluralidad de lecturas del genoma de la base de datos de referencia, sumar el número de coincidencias para la lectura del genoma, y comparar la suma del número de coincidencias para cada una de la pluralidad de lecturas del genoma con las sumas de la cantidad de coincidencias para cada una de las otras de la pluralidad de lecturas del genoma. El método puede comprender, además: para cada una de la pluralidad de lecturas del genoma de la base de datos de referencia, sumar el número de coincidencias únicas para la lectura del genoma y comparar la suma de la cantidad de coincidencias únicas para cada uno de la pluralidad de lecturas del genoma con las sumas de los números de coincidencias únicas para cada una de las otras de la pluralidad de lecturas del genoma. Una coincidencia única puede ser una coincidencia de una palabra de secuencia de muestra con una palabra de referencia contenida en solo una de las listas de palabras de referencia. La creación de la lista de palabras de referencia comprende dividir un genoma leído de la base de datos de referencia en palabras en un carácter de límite de palabra. La creación de la lista de palabras de referencia puede comprender guardar solo palabras que tengan una longitud mayor o igual a una longitud mínima de palabra. la longitud mínima de palabra puede ser igual a diecinueve letras. El método puede comprender además rellenar una tabla hash con las palabras de referencia de cada una de las listas creadas de palabras de referencia.

En otro aspecto, la presente invención proporciona

Un aparato para caracterizar material biológico en una muestra que contiene material genético de una pluralidad de organismos, comprendiendo el aparato un procesador y una memoria, en el que el procesador y la memoria están configurados para realizar el método de la reivindicación 1.

En una realización de la invención, el fragmento de nucleótidos se compara con las secuencias de ácido nucleico en una base de datos mediante emparejamiento probabilístico, que incluye, pero no se limita a, el enfoque bayesiano, el enfoque bayesiano recursivo o el enfoque bayesiano neófito.

Los enfoques probabilísticos pueden usar probabilidades bayesianas para considerar dos factores importantes para llegar a una conclusión precisa: (i) P(t i/R) es la probabilidad de que un organismo que exhibe el patrón de prueba R pertenezca al taxón t i, y (ii) P(R/t i) es la probabilidad de que los miembros del taxón t i presenten el patrón de prueba R. El patrón mínimo dentro de una ventana deslizante integrada en las herramientas ayudará a los investigadores a determinar “si” y “cómo” los organismos han sido modificados genéticamente.

Las variaciones adicionales comprendidas dentro de los sistemas y métodos se describen en la descripción detallada de la invención a continuación.

Breve descripción de los dibujos

Los dibujos adjuntos, que se incorporan aquí y forman parte de la especificación, ilustran diversas realizaciones de la presente invención. En los dibujos, los mismos números de referencia indican elementos idénticos o funcionalmente similares.

La Figura 1 es una ilustración esquemática de un sistema divulgado, que puede usarse para la ID genómica de muestras metagenómicas a nivel de especie y cepa.

La Figura 2 es una ilustración esquemática más detallada del sistema de la figura 1.

La Figura 3 es una ilustración esquemática de la interacción funcional entre el casete intercambiable y otros componentes en una realización del sistema de la figura 1.

La Figura 4 es una vista en perspectiva frontal de una realización de un dispositivo de secuenciación electrónico de mano, que puede usarse para la ID genómica de muestras metagenómicas a nivel de especie y cepa.

La Figura 5 es un diagrama de flujo que ilustra un proceso de funcionamiento del sistema de la figura 1.

La Figura 6 es una ilustración esquemática de la interacción del sistema de la figura 1 con diversas entidades potencialmente implicadas en el sistema.

La Figura 7 es una ilustración esquemática de la interacción funcional entre un dispositivo de secuenciación electrónico de mano con el centro de análisis remoto.

La Figura 8 es una ilustración esquemática de la arquitectura general del módulo de software probabilístico.

La Figura 9 muestra el porcentaje de secuencias únicas en función de la longitud de lectura.

La Figura 10 es un resumen de las principales etapas de secuenciación.

La Figura 11 es una ilustración esquemática de un instrumento capaz de caracterizar poblaciones de microorganismos en una muestra de acuerdo con una realización de la presente invención.

La Figura 12 es un diagrama de flujo de nivel superior que ilustra un proceso que se puede realizar para caracterizar poblaciones de microorganismos en una muestra.

La Figura 13 es un diagrama de flujo que ilustra un proceso que se puede realizar para identificar genomas de microorganismos contenidos dentro de una muestra.

La Figura 14 es un diagrama de flujo que ilustra una realización de un proceso de catalogación de sustancias.

La Figura 15 es un diagrama de flujo que ilustra una realización de un proceso de identificación y análisis de muestras desconocidas.

La Figura 16 es un diagrama de flujo que ilustra una realización de un proceso de búsqueda de palabras.

Las Figuras 17A-17E ilustran las medidas de población relativa de 16S en comparación con la secuenciación directa de ADN con identificación genómica de la presente invención.

La Figura 18 ilustra una comparación de la concentración relativa observada y la concentración real en una muestra con el número relativo de lecturas.

La Figura 19 ilustra un ejemplo del sistema y método de la presente invención aplicado a la medición de poblaciones del microbioma a nivel de especie para un paciente con enfermedad de Crohn.

Descripción detallada de realizaciones preferidas

Las realizaciones de los sistemas y métodos para la caracterización de poblaciones de microorganismos en una muestra se describen aquí con referencia a las figuras.

Los métodos y sistemas descritos en la presente invención pueden usar la información de secuencia única más corta, que en una mezcla de ácidos nucleicos en una muestra no caracterizada tiene la longitud única mínima (n) con respecto a la información de secuencia completa generada o recopilada. Además de las secuencias de longitud única, también se pueden comparar las no únicas. La probabilidad de identificación de un genoma aumenta con múltiples coincidencias. Algunos genomas tendrán secuencias únicas mínimas más largas que otros genomas. El método de coincidencia de secuencias de longitud corta (n) puede continuar en paralelo con la generación o recopilación de información de secuencia. Las comparaciones ocurren tan rápido como (en tiempo real) se generan o recopilan secuencias posteriores más largas. Esto da como resultado una reducción considerable del espacio de decisión porque los cálculos se realizan temprano en términos de generación/recopilación de información de secuencia. El emparejamiento probabilístico puede incluir, pero no se limita a, emparejamiento perfecto, unicidad de subsecuencia, emparejamiento patrón, múltiples emparejamientos de subsecuencias dentro de la longitud n, emparejamiento inexacto, semilla y extensión, mediciones de distancia y mapeo de árboles filogenéticos. Puede proporcionar una canalización automatizada para hacer coincidir la información de secuencia tan rápido como se genera o en tiempo real. El instrumento de secuenciación puede continuar recolectando más cadenas de información de secuencia en paralelo con la comparación. La información de secuencia subsecuente también se puede comparar y puede aumentar la confianza de un genoma o identificación de especies en la muestra. El método no necesita esperar el montaje de información de secuencia de las lecturas cortas en cóntigos más grandes.

En algunas realizaciones, el sistema y los métodos pueden proporcionar toma, aislamiento y separación de ácidos nucleicos, secuenciación de ADN, redes de bases de datos, procesamiento de información, almacenamiento de datos, visualización de datos y comunicación electrónica para acelerar la entrega de datos relevantes para permitir el diagnóstico o la identificación de organismos con aplicaciones para brotes patógenos y respuestas apropiadas. En estas realizaciones, el sistema puede incluir un dispositivo de secuenciación portátil que transmite electrónicamente datos a una base de datos para la identificación de organismos relacionados con la determinación de la secuencia de ácidos nucleicos y otras moléculas poliméricas o de tipo cadena y emparejamiento probabilístico de datos.

Las Figuras 1 y 2 ilustran una realización de un sistema 100 que incluye un dispositivo 105 de secuenciación, que puede ser un dispositivo de secuenciación electrónico portátil de mano. El dispositivo 105 de secuenciación puede configurarse para ser fácilmente sostenido y utilizado por un usuario (U), y puede ser capaz de comunicarse a través de una red 110 de comunicación con muchas otras entidades potencialmente relevantes.

El dispositivo puede configurarse para recibir una muestra del sujeto (SS) y una muestra ambiental (ES), respectivamente. La muestra del sujeto (como sangre, saliva, etc.) puede incluir el ADN del sujeto, así como el ADN de cualquier organismo (patógenos o de otro tipo) en el sujeto. La muestra ambiental (ES) puede incluir, pero no se limita a, organismos en su estado natural en el ambiente (incluidos alimentos, aire, agua, suelo, tejido). Ambas muestras (SS, ES) pueden verse afectadas por una infección natural, un acto de bioterrorismo o una epidemia emergente. Ambas muestras (SS, ES) pueden recolectarse simultáneamente a través de un tubo o hisopo y pueden recibirse en una solución o en un sólido (como una perla) en una membrana o portaobjetos, placa, capilar o canal. Las muestras (SS, ES) pueden secuenciarse simultáneamente. Situaciones específicas de circunstancias pueden requerir el análisis de una muestra compuesta por una mezcla de las muestras (SS, ES). Un primer respondedor se puede contactar una vez que se identifica un emparejamiento probabilístico y/o durante la recolección de datos en tiempo real y la interpretación de datos. A medida que pasa el tiempo, se puede identificar un porcentaje cada vez mayor de la secuencia.

El dispositivo 105 de secuenciación puede incluir los siguientes componentes funcionales, como se ilustra en la Figura 3, que permiten que el dispositivo 105 analice una muestra del sujeto (SS) y una muestra ambiental (ES), comunique el análisis resultante a una red 110 de comunicación.

Los receptores 120 y 122 de muestras se pueden acoplar a un bloque 130 de extracción y aislamiento de ADN, que luego entrega las muestras al bloque 130 a través de un sistema de flujo. El bloque 130 extrae el ADN de las muestras y lo aísla para que pueda procesarse y analizarse más. Esto se puede lograr mediante el uso de una plantilla de reactivo (es decir, una cadena de ADN que sirve como patrón para la síntesis de una cadena complementaria de ácido

nucleico), que se puede administrar combinada con las muestras 120, 122 utilizando tecnología de transporte de fluidos conocida. Los ácidos nucleicos en las muestras 120, 122 se separan mediante el bloque 130 de extracción y aislamiento, produciendo una corriente de fragmentos de nucleótidos o moléculas individuales no amplificadas. Una realización podría incluir el uso de métodos de amplificación.

Un casete 140 intercambiable puede acoplarse de manera removible al dispositivo 105 de secuenciación y al bloque 130. El casete 140 puede recibir la corriente de moléculas del bloque 130 y puede secuenciar el ADN y producir datos de secuencia de ADN.

El casete 140 intercambiable se puede acoplar y proporcionar los datos de la secuencia de ADN al procesador 160, donde se logra el emparejamiento probabilístico. Una realización podría incluir el rendimiento de 16 GB de datos transferidos a una velocidad de 1 Mb/s. Se prefiere un casete 140 de secuenciación para obtener la información de secuencia. Se pueden intercambiar diferentes casetes que representan diferentes métodos de secuenciación. La información de secuencia se puede comparar mediante comparación probabilística. Los algoritmos de emparejamiento ultrarrápido y las bases de datos de firmas ponderadas pregeneradas pueden comparar los datos de la secuencia de novo con los datos de la secuencia almacenada.

El procesador 160 puede ser, por ejemplo, un circuito integrado específico de la aplicación diseñado para lograr una o más funciones específicas o habilitar uno o más dispositivos o aplicaciones específicas. El procesador 160 puede controlar todos los demás elementos funcionales del dispositivo 105 de secuenciación. Por ejemplo, el procesador 160 puede enviar/recibir los datos de la secuencia de ADN para ser almacenados en un almacén 170 de datos (memoria). El almacén 170 de datos también puede incluir cualquier tipo o forma de memoria adecuada para almacenar datos en una forma recuperable por el procesador 160.

El dispositivo 105 de secuenciación puede incluir además un componente 180 de comunicación al que el procesador 160 puede enviar datos recuperados del almacén 170 de datos. El componente 180 de comunicación puede incluir cualquier tecnología adecuada para comunicarse con la red 110 de comunicación, tal como cableada, inalámbrica, satelital, etc.

El dispositivo 105 de secuenciación puede incluir un módulo 150 de entrada de usuario, que el usuario (U) puede proporcionar entrada al dispositivo 105. Esto puede incluir cualquier tecnología de entrada adecuada como botones, panel táctil, etc. Finalmente, el dispositivo 105 de secuenciación puede incluir un módulo 152 de salida de usuario que puede incluir una pantalla para salida visual y/o un dispositivo de salida de audio.

El dispositivo 105 de secuenciación también puede incluir un receptor 102 de sistema de posicionamiento global (GPS), que puede recibir datos de posicionamiento y enviar los datos al procesador 160, y una fuente 104 de alimentación (es decir, batería, adaptador enchufable) para suministrar energía eléctrica o de otro tipo a una carga de salida o grupo de cargas del dispositivo 105 de secuenciación.

El casete 140 intercambiable se ilustra esquemáticamente con más detalle en la Figura 3. El casete 140 se puede acoplar de forma extraíble al dispositivo 105 de secuenciación y al bloque 130 e incluye un método de secuenciación de última generación (es decir, secuenciación de alto rendimiento). El sistema basado en química húmeda o estado sólido se puede construir en la plataforma a través de un casete intercambiable “conecte & encienda”. El casete 140 puede recibir el flujo de moléculas del bloque 130 y puede secuenciar el ADN a través del método de secuenciación y puede producir datos de secuencia de ADN. Las realizaciones incluyen métodos basados en, pero no limitados a, secuenciación por síntesis, secuenciación por ligación, secuenciación de molécula única y pirosecuenciación. Otra realización más incluye una fuente para el campo 142 eléctrico y aplica el campo 142 eléctrico a la corriente de moléculas para efectuar la electroforesis del ADN dentro de la corriente. El casete incluye una fuente 144 de luz para emitir una luz 144 fluorescente a través de la corriente de ADN. El casete incluye, además, un sensor biomédico (detector) 146 para detectar la emisión de luz fluorescente y para detectar/determinar la secuencia de ADN de la corriente de muestra. Además de la luz fluorescente, el sensor biomédico es capaz de detectar luz en todas las longitudes de onda apropiadas para los restos etiquetados para secuenciar.

La detección fluorescente comprende la medición de la señal de un resto marcado de al menos uno de los uno o más nucleótidos o análogos de nucleótidos. La secuenciación usando nucleótidos fluorescentes típicamente implica fotoblanqueo del marcador fluorescente después de detectar un nucleótido añadido. Las realizaciones pueden incluir métodos fluorescentes basados en perlas, FRET, marcadores infrarrojos, pirofosfatasa, ligasa que incluyen nucleótidos marcados o polimerasa o el uso de terminadores cíclicos reversibles. Las realizaciones pueden incluir métodos directos de nanoporos o guías de ondas ópticas que incluyen moléculas individuales inmovilizadas o en solución. Los métodos de fotoblanqueo incluyen una intensidad de señal reducida, que se construye con cada adición de un nucleótido marcado con fluorescencia a la cadena del cebador. Al reducir la intensidad de la señal, se secuencian opcionalmente plantillas de ADN más largas.

El fotoblanqueo incluye aplicar un pulso de luz al cebador de ácido nucleico en el que se ha incorporado un nucleótido fluorescente. El pulso de luz típicamente comprende una longitud de onda igual a la longitud de onda de la luz absorbida por el nucleótido fluorescente de interés. El pulso se aplica durante aproximadamente 50 segundos o

menos, aproximadamente 20 segundos o menos, aproximadamente 10 segundos o menos, aproximadamente 5 segundos o menos, aproximadamente 2 segundos o menos, aproximadamente 1 segundo o menos, o aproximadamente 0. El pulso destruye la fluorescencia de los nucleótidos marcados con fluorescencia y/o el cebador o ácido nucleico marcado con fluorescencia, o lo reduce a un nivel aceptable, por ejemplo, un nivel de fondo, o un nivel suficientemente bajo para evitar la acumulación de señal durante varios ciclos.

El sensor (detector) 146 monitoriza opcionalmente al menos una señal de la plantilla de ácido nucleico. El sensor (detector) 146 incluye opcionalmente o está conectado operativamente a un ordenador que incluye software para convertir la información de la señal del detector en información de resultado de secuenciación, por ejemplo, concentración de un nucleótido, identidad de un nucleótido, secuencia del nucleótido plantilla, etc. Además, las señales de muestra se calibran opcionalmente, por ejemplo, calibrando el sistema de microfluidos mediante el seguimiento de una señal de una fuente conocida.

Como se muestra en la Figura 2, el dispositivo 105 de secuenciación puede comunicarse a través de una red 110 de comunicación con una variedad de entidades que pueden ser relevantes para notificar en el caso de un acto bioterrorista o un brote epidémico. Estas entidades pueden incluir un primer respondedor (es decir, Red de respuesta de laboratorio (es decir, laboratorios de referencia, laboratorios seminales, laboratorios nacionales), GenBank®, Centro para el Control de Enfermedades (CDC), médicos, personal de salud pública, registros médicos, datos del censo, aplicación de la ley, fabricantes de alimentos, distribuidores de alimentos, y minoristas de alimentos.

Ahora se describe una realización de ejemplo del dispositivo 105 de secuenciación discutido anteriormente con referencia a la figura 4 que ilustra una vista anterior del dispositivo. El dispositivo es un dispositivo de secuenciación portátil de mano y se ilustra en comparación con el tamaño de las monedas C. El dispositivo 105 tiene aproximadamente 11 pulgadas de largo y es fácilmente transportable. (En la Figura 4, las monedas se muestran para escala). Dos puertos 153, 154 están ubicados en un lado del dispositivo y representan los receptores 120, 122 de muestra. El puerto 153 es para recibir una muestra del sujeto (SS) o una muestra ambiental (ES) para ser analizada y secuenciada. El puerto 154 es para el control de secuenciación (SC). Los dos puertos diferentes están diseñados para determinar si una muestra del sujeto (SS) o una muestra ambiental (ES) contiene materiales que dan como resultado una falla de secuenciación, en caso de que ocurra una falla de secuenciación, o funcionan en una capacidad CLIA. El dispositivo 105 incluye un módulo 150 de entrada de usuario, que el usuario (U) puede proporcionar entrada al dispositivo 105. En esta realización particular, el módulo 150 de entrada de usuario tiene la forma de un panel táctil, sin embargo, se puede utilizar cualquier tecnología adecuada. El panel táctil incluye botones 150 a para visualización, 150 b, 150 c para registrar datos, 150 d para transmisión de datos en tiempo real y recepción, y 150 e para control de energía para activar o desactivar el dispositivo. Alternativamente, el teclado puede incorporarse en la pantalla de visualización y todas las funciones pueden controlarse mediante una interfaz de cristal líquido. Las técnicas adecuadas se describen en la Publicación de Patente de los Estados Unidos No. solicitud 2007/0263163, cuya descripción completa se incorpora aquí como referencia. Esto puede ser mediante emparejamiento de dispositivos con Bluetooth o enfoques similares. Las funciones incluyen teclas de dígitos, etiquetadas con letras del alfabeto, como el lugar común en los teclados telefónicos, como una tecla de borrar, tecla de espacio, tecla de escape, tecla de impresión, tecla de entrada, arriba/abajo, izquierda/derecha, caracteres adicionales y cualquier otro que desee el usuario. El dispositivo incluye además un módulo 152 de salida de usuario, en forma de una pantalla visual, para mostrar información para el usuario (U). También se puede proporcionar un dispositivo de salida de audio si se desea como se ilustra en 157 a y 157 b. Finalmente, el dispositivo 105 de secuenciación incluye diodos 155 y 156 emisores de luz para indicar la transmisión o recepción de datos. La función de las teclas/botones es controlar todos los aspectos de la secuenciación de muestras, la transmisión de datos y el emparejamiento probabilístico y los controles de interfaz, que incluyen, pero no se limitan a, encendido/apagado, envío, tecla de navegación, teclas de función, borrar y funciones de pantalla LCD y herramientas de visualización con rango genómico calculado por algoritmos para listar la confianza de coincidencias. Una realización incluye un sistema basado en Internet donde múltiples usuarios pueden transmitir/recibir datos simultáneamente hacia/desde un motor de búsqueda de red jerárquica.

La Figura 5 es un diagrama de flujo que ilustra un proceso de operación del sistema 100 de una realización del sistema 100 como se describe arriba. Como se muestra en la figura 5, un proceso de operación del dispositivo incluye en 200 recibir muestras de sujetos recolectadas (SS) y muestra ambiental (ES) en los receptores 120, 122 de muestras. En 202, las muestras pasan al Bloque 130 de Extracción y Aislamiento de ADN donde se analiza la muestra y se extrae el ADN de las muestras y se aísla. En 203, el casete 140 intercambiable recibe el ADN aislado del bloque 130 y secuencia el ADN. Dependiendo del casete y si es necesario, con la aplicación de un campo 142 eléctrico y de una luz 144 fluorescente, un sensor 146 biomédico dentro del casete 140 detecta/determina la secuencia de ADN de la corriente de la muestra. En 204, los datos secuenciados se procesan y almacenan en un almacén 170 de datos. En 205, los datos secuenciados se comparan mediante emparejamiento probabilístico y se logra la identificación del genoma. El proceso es de naturaleza reiterativa. La información resultante se puede transmitir a través de una red 110 de comunicación. Los datos de GPS (sistema de posicionamiento global) también se pueden transmitir opcionalmente en la etapa 205. En 206, el dispositivo recibe electrónicamente datos del emparejamiento. En 207, el dispositivo muestra visualmente los datos recibidos electrónicamente del emparejamiento a través de un módulo 152 de salida de usuario. Si se requiere un análisis adicional, en 208, los datos secuenciados se transmiten electrónicamente a entidades de interpretación de datos (es decir, personal de salud pública, registros médicos, etc.) a través de la red de comunicación.

Un enfoque de investigación multimétodo puede mejorar la respuesta rápida a un incidente e integrar la atención primaria con la detección de organismos. Se puede utilizar una respuesta triangular, que involucra datos de instrumentos cuantitativos de la secuenciación del ADN para converger con cuidados críticos cualitativos. Una infraestructura de listas de verificación de observación y auditorías de los datos de secuenciación de ADN recopilados en el campo en múltiples ubicaciones pueden usarse para comparar la apariencia de un organismo, por ejemplo, bioamenaza entre ubicaciones. El análisis estadístico inferencial de los datos genómicos puede combinarse con observaciones médicas para desarrollar categorías de prioridades. La información recopilada y compartida entre las bases de datos de los centros médicos y los centros genómicos puede permitir la triangulación de un incidente, la magnitud del incidente y la entrega de la intervención correcta a las personas afectadas en el momento adecuado.

La Figura 6 ilustra la interacción entre el sistema 100 y diversas entidades de recursos potenciales. El dispositivo 105 está configurado para interactuar con estas entidades de recursos a través de una red de comunicación inalámbrica o cableada. El dispositivo 105 puede transmitir información (310) de datos secuenciados triangulados ilustrando los “Datos de muestra”, los “Datos del paciente” y la “Intervención del tratamiento”. El dispositivo 105 puede transmitir y recibir datos de secuencias de ADN hacia y desde los recursos 320 de emparejamiento de secuencias, que incluyen GenBank® y una red de respuesta de laboratorio que incluye Sentinel Labs, Reference Labs y Laboratorios Nacionales.

Cada uno de los laboratorios tiene roles específicos. Los laboratorios centinela (hospitales y otros laboratorios clínicos comunitarios) son los encargados de descartar o derivar los agentes críticos que encuentren a los laboratorios de referencia de la LRN cercanos. Los laboratorios de referencia (laboratorios estatales y de salud pública local donde se observan prácticas de nivel 3 de seguridad biológica (BSL-3)) realizan pruebas confirmatorias (decisorias). Los laboratorios nacionales (BSL-4) mantienen una capacidad capaz de manejar agentes virales como Ébola y variola major y realizar caracterización definitiva.

El sistema 100 puede transmitir y recibir datos hacia y desde los recursos 330 de interpretación de datos, incluidas las entidades que hacen cumplir la ley, el personal de salud pública, los registros médicos y los datos del censo. Finalmente, el dispositivo 105 puede transmitir y recibir datos hacia y desde un primer respondedor 320 que incluyen doctores o médicos en una sala de emergencias El sistema 100 en general está configurado para comunicarse con el Centro para el Control de Enfermedades (CDC) 340 para proporcionar información pertinente al personal adecuado.

La Figura 7 es una ilustración esquemática de la interacción funcional entre un dispositivo de secuenciación electrónico portátil con el centro de análisis remoto. El dispositivo 105 puede incluir una unidad 103 de llamada base para procesar la secuencia recibida por el casete 140 intercambiable. Tales secuencias y sitios SNP se ponderan individualmente de acuerdo con su probabilidad encontrada en cada especie. Estas ponderaciones pueden calcularse teóricamente (mediante simulación) o experimentalmente. El dispositivo también incluye un procesador 109 de emparejamiento probabilístico acoplado a la unidad 103 de llamada base. El emparejamiento probabilístico se puede realizar en tiempo real o tan rápido como la llamada de la base de la secuencia o la recopilación de datos de la secuencia. El procesador 109 de emparejamiento probabilístico, utilizando un enfoque bayesiano, puede recibir la secuencia resultante y los datos de calidad, y puede calcular las probabilidades para cada lectura de secuenciación mientras considera las puntuaciones de calidad de la secuencia generado por la unidad 103 de llamada base. El procesador 109 de emparejamiento probabilístico puede utilizar una base de datos generada y optimizada antes de su uso para la identificación de patógenos. Un sistema 107 de alerta está acoplado al procesador 109 de emparejamiento probabilístico y puede recopilar información del procesador 109 de emparejamiento probabilístico (en el sitio) y mostrar el(los) organismo(s) mejor emparejado(s) en tiempo real.

El sistema 107 de alerta está configurado para acceder a los datos del paciente, es decir, el diagnóstico médico o la evaluación de riesgos para un paciente, en particular los datos de pruebas o ensayos de diagnóstico en el punto de atención, incluidos inmunoensayos, electrocardiogramas, rayos X y otras pruebas similares, y proporcionar una indicación de una condición médica o riesgo o ausencia del mismo. El sistema de alerta puede incluir software y tecnologías para leer o evaluar los datos de la prueba y para convertir los datos en información de diagnóstico o evaluación de riesgos. Dependiendo de la identidad del genoma del bioagente y los datos médicos del paciente, se puede administrar una “Intervención de tratamiento” eficaz. El tratamiento puede basarse en la mitigación o neutralización eficaz del bioagente y/o sus efectos secundarios y basarse en el historial del paciente si existen contraindicaciones. El sistema de alerta puede basarse en el grado y número de ocurrencias. El número de ocurrencias puede basarse en la identificación genómica del bioagente. Un valor se puede declarar cuando el resultado está dentro o supera un umbral determinado por agencias gubernamentales, como el CDC o el Departamento de Defensa o Seguridad Nacional. El sistema de alerta está configurado para permitir que los médicos utilicen la funcionalidad de los datos de identificación genómica con el paciente. La comunicación permite un flujo rápido de información y una toma de decisiones precisa para las acciones de los primeros respondedores u otros sistemas clínicos.

El dispositivo 105 incluye además un compresor 106 de datos acoplado a la unidad 103 de llamada base, configurado para recibir la secuencia resultante y los datos de calidad para la compresión. El almacén 170 de datos está acoplado al compresor 106 y puede recibir y almacenar la secuencia y datos de calidad.

El dispositivo 105 de secuenciación interactúa con un centro 400 de análisis remoto, que puede recibir datos transferidos electrónicamente desde el componente 180 de comunicación del dispositivo 105 de secuenciación a través de un método de comunicación alámbrico y/o inalámbrico. El centro 400 de análisis remoto contiene una base de datos secuencia larga que incluye todas las secuencias de nucleótidos y aminoácidos y los datos de SNP disponibles hasta la fecha. Esta base de datos también contiene información epidemiológica y terapéutica asociada (por ejemplo, resistencia a los antibióticos). El centro 400 de análisis remoto incluye además un almacén 401 de datos. El almacén 401 de datos puede recibir información de datos de secuencia descomprimida a través de transmisión electrónica desde el componente 180 de comunicación del dispositivo 105 de secuenciación. Un montaje 402 de genoma está acoplado al almacén 401 de datos y puede y ensambla los datos de secuencia descomprimidos. El ADN contaminante obvio, como el ADN humano, se puede filtrar antes para un análisis más detallado.

El centro 400 de análisis remoto incluye además un procesador 403 equipado con tecnología de emparejamiento probabilístico y algoritmos de búsqueda de homología, que se pueden emplear para analizar datos de secuencia ensamblados para obtener las probabilidades de presencia de patógenos 403a diana, estructura 403 b comunitaria, Información 403c epidemiológica y terapéutica. Los datos de la secuencia del genoma de los patógenos diana se comparan con los de los genomas de no patógenos, incluidos los humanos y el metagenoma, para identificar las secuencias de nucleótidos y los sitios polimórficos de un solo nucleótido (SNP), que solo ocurren en los organismos diana. El análisis en el centro 400 de análisis a distancia se lleva a cabo sobre la marcha durante la transferencia de datos desde el dispositivo 105 de secuenciación. El centro 400 de análisis remoto puede incluir además una unidad 404 de comunicación desde la cual los resultados del análisis se transfieren electrónicamente de regreso al sistema 107 de alerta dentro del dispositivo 105 de secuenciación, así como otras autoridades (por ejemplo, DHS, CDC, etc.).

Clasificación probabilística: la presente invención puede proporcionar motores de base de datos, diseño de bases de datos, técnicas de filtrado y el uso de la teoría de la probabilidad como lógica extendida de acuerdo con las reivindicaciones.

Los métodos y el sistema actuales pueden utilizar los principios de la teoría de la probabilidad para hacer razonamientos plausibles (tomar decisiones) sobre los datos producidos por la secuenciación de ácidos nucleicos. Usando el enfoque de la teoría de la probabilidad, el sistema aquí descrito puede analizar datos tan pronto como alcance un número mínimo de nucleótidos de longitud (n), y calcular la probabilidad del entonces mero, además, cada aumento subsiguiente de longitud (n + par(es) de bases) se utiliza para calcular la probabilidad de una coincidencia de secuencia. El cálculo de cada n-mero y posteriores n-meros más largos pueden procesarse adicionalmente para recalcular las probabilidades de todas las longitudes crecientes para identificar la presencia de genoma(s). A medida que aumenta la longitud de la unidad, se comparan múltiples subunidades dentro del n-mero para el reconocimiento de patrones, lo que aumenta aún más la probabilidad de una coincidencia. Dicho método, incluidos otros métodos bayesianos, permite eliminar coincidencias e identificar un número significativo de muestras biológicas que comprenden con un fragmento o lectura de nucleótido muy corta sin tener que completar la secuenciación del genoma total completo o ensamblar el genoma. Como tal, asignar la probabilidad de coincidencia a los organismos existentes y pasar a la siguiente secuencia de ácido nucleico leída para mejorar aún más la probabilidad de la coincidencia. El sistema descrito en este documento aumenta velocidad, reduce el consumo de reactivos, permite la miniaturización y reduce significativamente el tiempo necesario para identificar el organismo.

Para construir clasificadores probabilísticos para tomar una decisión sobre secuencias cortas de ácido nucleico, se pueden utilizar una variedad de enfoques para filtrar primero y luego clasificar los datos de secuenciación entrantes. En el presente caso, se utiliza el formalismo de redes bayesianas. Una red bayesiana es un gráfico acíclico dirigido que representa de manera compacta una distribución de probabilidad. En dicho gráfico, cada variable aleatoria se indica mediante un nodo (por ejemplo, en un árbol filogenético de un organismo). Un borde dirigido entre dos nodos indica una dependencia probabilística de la variable denotada por el nodo padre a la del hijo. En consecuencia, la estructura de la red denota el supuesto de que cada nodo de la red es condicionalmente independiente de sus no descendientes dados sus padres. Para describir una distribución de probabilidad que satisfaga estos supuestos, cada nodo de la red está asociado con una tabla de probabilidad condicional, que especifica la distribución sobre cualquier posible asignación de valores dad a sus padres. En este caso, un clasificador bayesiano es una red bayesiana aplicada a una tarea de clasificación de calcular la probabilidad de cada nucleótido proporcionada por cualquier sistema de secuenciación. En cada punto de decisión, el clasificador bayesiano se puede combinar con una versión del algoritmo de gráfico de ruta más corta, como el de Dijkstra o Floyd.

El sistema actual puede implementar un sistema de clasificadores bayesianos (por ejemplo, clasificador bayesiano neófito, clasificador bayesiano y clasificador de estimación bayesiano recursivo) y fusionar los datos resultantes en la base de datos de decisiones. Después de fusionar los datos, cada clasificador puede recibir un nuevo conjunto de resultados con probabilidades actualizadas.

La Figura 8 muestra una ilustración esquemática de la arquitectura general del módulo de software probabilístico.

Fragmento de secuenciación de ADN: se puede utilizar cualquier método de secuenciación para generar la información del fragmento de secuencia. El módulo 160 en la figura 2 o 109 en la figura 7 es responsable de procesar los datos que ingresan desde el módulo de secuenciación en el casete intercambiable. Los datos están encapsulados con los

datos de secuenciación, así como información sobre el inicio y la parada de la secuencia, ID de secuencia, ID de cadena de ADN. El módulo formatea los datos y los pasa al módulo de filtro de taxonomía. El formateo incluye la adición de los datos del sistema y la alineación en fragmentos.

El módulo de secuenciación de ADN tiene 2 interfaces, está conectado al módulo de preparación de ADN y al filtro de taxonomía.

I. Interfaz de preparación de ADN: se pueden integrar varios métodos disponibles comercialmente para lograr la preparación de muestras mediante técnicas de microfluidos. La preparación típica de muestras se basa en una solución e incluye la lisis celular y la eliminación del inhibidor. Los ácidos nucleicos se recuperan o extraen y concentran. Realizaciones de la lisis incluye detergente/enzimas, métodos mecánicos, de microondas, de presión y/o ultrasónicos. Las realizaciones de extracción incluyen afinidad de fase sólida y/o exclusión de tamaño.

II. Filtro de taxonomía: El filtro de taxonomía tiene dos tareas principales: (i) Filtrar tantos organismos como sea posible para limitar el módulo clasificador a un espacio de decisión más pequeño, y (ii) ayudar a determinar la estructura de la red bayesiana, que implica el uso de técnicas de aprendizaje automático.

Filtro de árbol filogenético: este submódulo de filtro de taxonomía interactúa con la “Base de datos de decisiones” para conocer los resultados de la ronda anterior de análisis. Si no se encuentran resultados, el módulo pasa los nuevos datos al módulo de clasificación. Si los resultados son encontrados el filtro de taxonomía ajusta los datos del clasificador para limitar el posible espacio de decisión. Por ejemplo, si los datos anteriores indican que se trata de una secuencia de ADN viral que se está examinando, el espacio de decisión para el clasificador se reducirá a datos virales únicamente. Esto puede hacerse modificando los clasificadores bayesianos de datos recopilados durante el funcionamiento.

Aprendizaje automático: los algoritmos de aprendizaje automático se organizan en una taxonomía, en función del resultado deseado del algoritmo. (i) Aprendizaje supervisado, en el que el algoritmo genera una función que asigna las entradas a las salidas deseadas. Una formulación estándar de la tarea de aprendizaje supervisada es el problema de la clasificación: se requiere que el alumno aprenda (para aproximar) el comportamiento de una función que mapea un vector [X1, X2, ... XN] en una de varias clases al observar varios ejemplos de entrada-salida de la función. (ii) Aprendizaje semisupervisado, que combina ejemplos etiquetados y no etiquetados para generar una función o clasificador apropiada. (iii) Aprendizaje por refuerzo, en el que el algoritmo aprende una política de cómo actuar dada una observación del mundo. Cada acción tiene algún impacto en el entorno, y el entorno proporciona retroalimentación que guía el algoritmo de aprendizaje. (iv) Transducción: predice nuevos resultados en función de los insumos de capacitación, los resultados de capacitación y los insumos de prueba que están disponibles durante el entrenamiento. (v) Aprender a aprender, en el que el algoritmo aprende su propio sesgo inductivo basado en experiencias previas.

Módulo de caché de taxonomía: el módulo almacena en caché la información de taxonomía producida por el filtro de taxonomía. Puede actuar como una interfaz entre el filtro de taxonomía y la base de datos de taxonomía que contiene toda la información en la base de datos SQL. La caché de taxonomía se implementa como una base de datos en memoria con un tiempo de respuesta de microsegundo. Las consultas a la base de datos SQL se manejan en un hilo separado del resto del submódulo. La información de la caché incluye el gráfico de red creado por el módulo de filtro de taxonomía. El gráfico contiene la taxonomía completa cuando el sistema comienza el análisis. El análisis de la secuencia de ADN reduce el gráfico de taxonomía con la caché de taxonomía implementando las reducciones en el tamaño de los datos y la eliminación de los conjuntos de datos apropiados.

Selector de clasificador: el presente sistema puede utilizar múltiples técnicas de clasificación que se ejecutan en paralelo. El selector de clasificador puede actuar como árbitro de datos entre diferentes algoritmos de clasificación. El selector de clasificador puede leer información de la base de datos de decisiones y enviar dicha información a los módulos de clasificación con cada ADN. Unidad de secuenciación recibida para su análisis desde el Módulo de secuenciación de ADN. El filtro de taxonomía actúa como el paso de datos para los datos de secuenciación de ADN.

Clasificador bayesiano recursivo: el clasificador bayesiano recursivo es un enfoque probabilístico para estimar una función de densidad de probabilidad desconocida de forma recursiva a lo largo del tiempo utilizando medidas entrantes y un modelo de proceso matemático. El módulo recibe datos del selector de clasificador y de la base de datos de decisiones donde se almacenan las decisiones anteriores. El conjunto de datos se recupera de las bases de datos y la identificación de la decisión previa se coloca en la memoria local del módulo donde se produce el filtrado. El clasificador toma la secuencia de ADN e intenta emparejarla con o sin firmas, códigos de barras, etc., existentes, de la base de datos de taxonomía filtrando rápidamente las familias de organismos que no coinciden. El algoritmo funciona calculando las probabilidades de creencias múltiples y ajustando las creencias en función de los datos entrantes. Los algoritmos utilizados en este módulo pueden incluir métodos de Monte Carlo secuenciales y remuestreo de importancia de muestreo. Modelo de Markov oculto, Filtro Ensemble Kalman y otros filtros de partículas también se pueden utilizar junto con la técnica de actualización bayesiana.

Clasificador bayesiano neófito: clasificador probabilístico simple basado en la aplicación del teorema de Bayes. El clasificador toma todas las decisiones basándose en el conjunto de reglas predeterminado que se proporciona como

entrada del usuario al inicio. El módulo puede ser reinicializado con un nuevo conjunto de reglas mientras se ejecuta el análisis. El nuevo conjunto de reglas puede provenir del usuario o puede ser un producto de la fusión de reglas del módulo de fusiones resultantes.

Clasificador de red bayesiana: El clasificador de red bayesiana implementa una red bayesiana (o una red de creencias) como un modelo gráfico probabilístico que representa un conjunto de variables y sus independientes probabilísticas.

Base de datos de decisiones: La base de datos de decisiones es un caché de trabajo para la mayoría de los módulos del sistema. La mayoría de los módulos tienen acceso directo a este recurso y pueden modificar sus regiones individuales. Sin embargo, solo el módulo de fusión de resultados puede acceder a todos los datos y modificar los conjuntos de reglas bayesianas en consecuencia.

Datos de reglas bayesianas: el módulo recopila todas las reglas bayesianas en forma binaria y precompilada. Las reglas son de lectura y escritura en todos los clasificadores bayesianos, así como en los módulos de filtro de taxonomía y fusiones de resultados. Las reglas se recompilan dinámicamente a medida que se realizan los cambios.

Fusión de resultados: El módulo fusiona la fecha de múltiples clasificadores bayesianos, así como otros clasificadores estadísticos que se utilizan. El módulo de fusión de resultados analiza la varianza media entre las respuestas generadas para cada clasificador y fusiona los datos si es necesario.

Interfaz de base de datos: Interfaz para la base de datos SQL. La interfaz se implementa programáticamente con funciones de lectura y escritura separadas en diferentes hilos. MySQL es la base de datos de elección, sin embargo, sqLite puede usarse para una velocidad de base de datos más rápida.

Base de datos de taxonomía: la base de datos contendrá múltiples bases de datos internas: árbol de taxonomía, árbol preprocesado indexado, entrada del usuario y reglas.

Reglas almacenadas en caché: Caché en memoria de reglas posprocesadas proporcionadas por el usuario.

Gestión de reglas: Interfaz de gestión gráfica para el módulo

Entrada de usuario: reglas de inferencia creadas por el usuario. Las reglas son utilizadas por los clasificadores bayesianos para tomar decisiones.

Los sistemas y métodos de la invención se describen en este documento como incorporados en programas informáticos que tienen un código para realizar una variedad de funciones diferentes. El código puede incorporarse en un medio legible por ordenador no transitorio. Las mejores tecnologías particulares de su clase (presentes o emergentes) pueden ser componentes con licencia. Los métodos existentes para la extracción de ADN incluyen el uso de fenol/cloroformo, la salificación, el uso de sales caotrópicas y resinas de sílice, el uso de resinas de afinidad, la cromatografía de intercambio iónico y el uso de perlas magnéticas. Los métodos se describen en las Patentes de Estados Unidos Nos. 5.057.426, 4.923.978, Patentes EP 0512767 A1 y EP 0515484B y WO 95/13368, WO 97/10331 y WO 96/18731,

Debe entenderse, sin embargo, que los sistemas y métodos no se limitan a un medio electrónico, y que diversas funciones se pueden practicar alternativamente en una configuración manual. Los datos asociados con el proceso se pueden transmitir electrónicamente a través de una conexión de red utilizando Internet. Los sistemas y técnicas descritos anteriormente pueden ser útiles en muchos otros contextos, incluidos los que se describen a continuación.

Estudios de asociación de enfermedades: muchas enfermedades y afecciones comunes involucran factores genéticos complejos que interactúan para producir las características visibles de esa enfermedad, también llamado fenotipo. Múltiples genes y regiones reguladoras a menudo se asocian con una enfermedad o síntoma particular. Mediante la secuenciación de los genomas o genes seleccionados de muchos individuos con una condición determinada, puede ser posible identificar las mutaciones causales subyacentes a la enfermedad y la relación de agentes causantes de enfermedad específicos con la condición(es). Esta investigación puede conducir a avances en la detección, prevención y tratamiento de enfermedades.

Investigación sobre el cáncer: la genética del cáncer implica comprender los efectos de las mutaciones heredadas y adquiridas y otras alteraciones genéticas. El desafío de diagnosticar y tratar el cáncer se ve agravado por la variabilidad del paciente individual y las respuestas difíciles de predecir a la terapia con medicamentos. La disponibilidad de la secuenciación de genoma a bajo costo para caracterizar los cambios adquiridos del genoma que contribuyen al cáncer en base a muestras pequeñas o biopsias de células tumorales, y la identificación de agentes infecciosos asociados con y/o que influyen en el diagnóstico, pronóstico y resultado de la enfermedad puede permitir un mejor diagnóstico y tratamiento de cáncer.

Investigación y desarrollo farmacéutico: una promesa de la genómica ha sido acelerar el descubrimiento y desarrollo de nuevos fármacos más eficaces. El impacto de la genómica en esta área ha surgido lentamente debido a la

complejidad de las vías biológicas, los mecanismos de la enfermedad y las múltiples dianas de los fármacos. La secuenciación de una sola molécula podría permitir un cribado de alto rendimiento de manera rentable mediante el análisis de la expresión génica a gran escala para identificar mejor los fármacos potenciales prometedores. En el desarrollo clínico, la tecnología divulgada podría utilizarse para generar perfiles de genes individuales que puedan proporcionar información valiosa sobre la respuesta probable a la terapia, la toxicología o el riesgo de eventos adversos, y posiblemente para facilitar la selección del paciente y la individualización de la terapia.

Enfermedad infecciosa: Todos los virus, bacterias y hongos contienen ADN o ARN. La detección y secuenciación de ADN o ARN de patógenos a nivel de molécula única podría proporcionar información médica y ambientalmente útil para el diagnóstico, tratamiento y seguimiento de infecciones y para predecir la resistencia potencial a los medicamentos.

Afecciones autoinmunes: Se cree que varias afecciones autoinmunes, que van desde la esclerosis múltiple y el lupus hasta el riesgo de rechazo de trasplantes, tienen un componente genético. La monitorización de los cambios genéticos y los microorganismos probables asociados con estas enfermedades puede permitir un mejor manejo del paciente.

Diagnóstico clínico: los pacientes que presentan los mismos síntomas de la enfermedad a menudo tienen diferentes pronósticos y respuestas a los medicamentos en función de sus diferencias genéticas subyacentes. La entrega de información genética específica del paciente abarca diagnósticos moleculares que incluyen kits y servicios de diagnóstico basados en genes o expresión, productos de diagnóstico compañeros para seleccionar y monitorizar terapias particulares, así como también para la selección de pacientes para la detección temprana de enfermedades y la monitorización de enfermedades. La creación de diagnósticos moleculares y pruebas de cribado más efectivos y específicos requiere una mejor comprensión de los genes, los factores reguladores y otros factores relacionados con la enfermedad o los medicamentos, así como agentes microbianos asociados o causantes, que la tecnología de secuenciación de una sola molécula divulgada tiene el potencial de habilitar.

Agricultura: la investigación agrícola se ha vuelto cada vez más hacia la genómica para el descubrimiento, desarrollo y diseño de animales y cultivos genéticamente superiores. La industria de la agroindustria ha sido un gran consumidor de tecnologías genéticas, particularmente microarreglos, para identificar variaciones genéticas relevantes entre variedades o poblaciones. La tecnología de secuenciación divulgada puede proporcionar un enfoque más potente, directo y rentable para el análisis de la expresión génica y los estudios de población y la identificación de agentes microbianos comensales, patógenos y/o simbióticos para esta industria.

Más oportunidades estarán en el campo de las aplicaciones de secuencias repetidas donde los métodos se aplican a la detección de variaciones genéticas sutiles. El análisis genómico comparativo ampliado entre especies puede proporcionar grandes conocimientos sobre la estructura y función del genoma humano y, en consecuencia, la genética de la salud y las enfermedades humanas y las relaciones de los microorganismos con la salud o las enfermedades humanas. Se están ampliando los estudios de la variación genética humana y su relación con la salud y la enfermedad. La mayoría de estos estudios utilizan tecnologías que se basan en patrones de variación conocidos y relativamente comunes. Estos poderosos métodos proporcionarán nueva información importante, pero son menos informativos que la determinación de la secuencia completa y contigua de genomas humanos individuales. Por ejemplo, es probable que los métodos de genotipado actuales pasen por alto diferencias raras entre personas en cualquier ubicación genómica particular y tengan una capacidad limitada para determinar reordenamientos de largo alcance. La caracterización de cambios somáticos del genoma que contribuye al cáncer actualmente emplea combinaciones de tecnologías para obtener datos de secuencia (en muy pocos genes) además de información limitada sobre cambios en el número de copias, reordenamientos o pérdida de heterocigosidad. Estos estudios tienen una resolución deficiente y/o una cobertura incompleta del genoma. La heterogeneidad celular de las muestras tumorales presenta desafíos adicionales, así como la falta de conocimiento de la aminoflora humana. La secuenciación completa del genoma de bajo costo a partir de muestras extremadamente pequeñas, quizás incluso células individuales, alteraría la batalla contra el cáncer en todos los aspectos, desde el laboratorio de investigación a la clínica. El proyecto piloto del Atlas del genoma del cáncer (TCGA), lanzado recientemente, avanza en la dirección deseada, pero sigue estando drásticamente limitado por los costos de secuenciación. Se necesitan secuencias del genoma adicionales de animales y plantas de importancia agrícola para estudiar la variación individual, razas y varias variantes silvestres de cada especie. Análisis de secuencia de comunidades microbianas, cuyos muchos miembros pueden no ser cultivados, proporcionarán una rica fuente de información médica y ambientalmente útil. Y la secuenciación rápida y precisa puede ser el mejor enfoque para la monitorización microbiana de los alimentos y el ambiente, incluida la detección y mitigación rápida de las amenazas bioterroristas.

La secuenciación del genoma también podría proporcionar ácidos nucleicos aislados que comprenden regiones intrónicas útiles en la selección de secuencias de firma clave. Actualmente, las secuencias de firma clave están dirigidas a regiones exónicas.

Una aplicación fundamental de la tecnología del ADN implica diversas estrategias de marcado para marcar un ADN producido por una ADN polimerasa, lo que es útil en la tecnología de microarreglos: secuenciación de ADN, detección de SNP, clonación, análisis de PCR y muchas otras aplicaciones.

Si bien se han descrito anteriormente varias realizaciones de la invención, debe entenderse que se han presentado solo a modo de ejemplo, y no de limitación. Por lo tanto, la amplitud y el alcance de la invención no deben estar limitados por ninguna de las realizaciones descritas anteriormente, sino deben definirse solo de acuerdo con las siguientes reivindicaciones.

EJEMPLO 1

Objetivo: el uso de firmas clave y/o códigos de barras para permitir la identificación del genoma con tan solo 8-18 nucleótidos y el análisis de datos de secuencia muy corta (lecturas) en tiempo real.

Se utilizaron algoritmos de construcción de matrices de sufijos de tiempo lineal para calcular el análisis de unicidad. El análisis determinó el porcentaje de todas las secuencias que eran únicas en varios genomas modelo. Se analizaron todas las longitudes de secuencia en un genoma. Se cuentan las secuencias que ocurren solo una vez en un genoma. El algoritmo de matriz de sufijos funciona calculando una gráfica de puntuación repetida que analiza la frecuencia de subsecuencias específicas dentro de una secuencia que se producen en función de una ventana deslizante de dos pares de bases. La información del genoma almacenada en GenBank se utilizó para el análisis insilico. Se analizaron un genoma viral, el fago lambda, un genoma bacteriano, E. coli K12 MG1655 y el genoma humano. El porcentaje de lecturas únicas es función de la longitud de la secuencia. Se hizo una suposición con respecto a las secuencias que solo producen coincidencias inequívocas y que producen superposiciones inequívocas para reconstruir el genoma. Las lecturas únicas variaron en tamaño de 7 a 100 nucleótidos. La mayoría de los tamaños únicos eran más cortos de 9, 13 y 18 nucleótidos, respectivamente.

Resultados: Los resultados muestran que las secuencias aleatorias de 12 nt del genoma del fago son 98 % únicas para el fago. Esto aumenta lentamente de manera que las secuencias de 400 nt (nucleótidos) son 99 % únicas para los fagos. Esto disminuye al 80 % para secuencias de fagos de 10 nt. Para las bacterias (E. coli), las secuencias de 18 nt del genoma son 97 % exclusivas de E. coli. Para los genomas humanos, las secuencias de 25 nt son 80 % únicas para humanos y un aumento a 45 nt da como resultado que el 90 % del genoma sea único.

Como se señaló anteriormente, aunque la presente invención se ha descrito con realizaciones y ejemplos, la amplitud y el alcance de la invención no deberían estar limitados por ninguna de las realizaciones descritas anteriormente. Por ejemplo, aunque se pueden haber descrito realizaciones de la invención en las que la identificación de genomas presentes en el material biológico de una muestra se produjo instantáneamente o en tiempo real y con comunicación directa con un secuenciador, la presente invención no requiere que la identificación ocurra instantáneamente o en tiempo real o comunicación directa o contacto con un secuenciador. Sin embargo, el emparejamiento probabilístico para identificar fragmentos metagenómicos con una base de datos genómica se puede realizar por separado de la secuenciación en un momento posterior utilizando un archivo de salida generado por un secuenciador, y no se requiere comunicación directa o procesamiento conectado al secuenciador. Por tanto, aunque la identificación se puede realizar en paralelo (es decir, en línea) con la secuenciación del material biológico como se describe, esto tampoco es necesario y la identificación se puede realizar de forma escalonada. De hecho, cada una de las etapas de recogida de muestras, extracción de muestras, secuenciación e identificación se puede realizar por separado y de forma escalonada. Además, con respecto al secuenciador utilizado en o con la presente invención, no existe ningún requisito de que el secuenciador sea portátil o de mano.

La Figura 11 es una ilustración esquemática de un instrumento 1100 de acuerdo con una realización de la presente invención de acuerdo con las reivindicaciones.

El instrumento 1100 puede ser un dispositivo capaz de caracterizar poblaciones de microorganismos en una muestra. En particular, el instrumento 1100 puede ser un dispositivo capaz de caracterizar las identidades y poblaciones relativas de microorganismos, como patógenos y comensales, en una muestra a nivel de especie y/o subespecie (por ejemplo, Morfovares, serovares y biovares) y/o nivel de deformación.

El instrumento 1100 puede comprender un procesador 1101 y una memoria 1102 configurada para realizar la caracterización de poblaciones de microorganismos en una muestra. Alternativamente, el instrumento 1100 puede comprender unidades en forma de hardware y/o software, cada una configurada para realizar una o más porciones de la caracterización de poblaciones de microorganismos en una muestra. Además, cada una de las unidades puede comprender su propio procesador y memoria, o cada una de las unidades puede compartir un procesador y memoria con una o más de las otras unidades.

El instrumento 1100 utiliza lecturas de fragmentos metagenómicos, que pueden producirse recolectando una muestra, extrayendo fragmentos metagenómicos (por ejemplo, ácido nucleico y/o proteína y/o metabolitos) y secuenciando los fragmentos. La muestra puede contener una pluralidad de organismos microbianos, que incluyen bacterias, virus, parásitos, hongos, plásmidos y otros fragmentos de ADN o ARN exógenos disponibles en el tipo de muestra. El instrumento 1100 puede utilizar cientos, miles o millones de lecturas cortas de fragmentos metagenómicos. Las lecturas del fragmento metagenómico pueden tener la forma de un archivo 1103 metagenómico producido a partir de las lecturas del fragmento metagenómico.

Aunque la presente invención puede utilizar lecturas de fragmentos metagenómicos que pueden tener una longitud superior a 100 pares de bases, las lecturas de fragmentos metagenómicos utilizadas también pueden tener longitudes de aproximadamente 12 a 100 pares de bases. Por ejemplo, el instrumento 1100 puede caracterizar poblaciones de microorganismos usando lecturas de fragmentos metagenómicos que tienen longitudes de aproximadamente 12 a 15 pares de bases, 16 a 25 pares de bases, 25 a 50 pares de bases o 50 a 100 pares de bases. Por ejemplo, para el ADN, las lecturas de fragmentos metagenómicos pueden tener longitudes de lectura de menos de 100 pares de bases, y el archivo 1103 metagenómico producido a partir de ellas puede contener millones de lecturas de fragmentos de ADN.

En la realización ilustrada en la Fig. 11, el instrumento 1100 recibe un archivo 1103 metagenómico como entrada. Sin embargo, en otras realizaciones, el instrumento 1100 también puede comprender un extractor y secuenciador y ser capaz de recibir una muestra como entrada y producir un archivo 1103 metagenómico a partir de la misma (ver, por ejemplo, la Figura 2). En otras realizaciones más, el instrumento 1100 puede recibir lecturas de fragmentos metagenómicos individualmente y producir un archivo 1103 metagenómico que incluye las lecturas de fragmentos metagenómicos recibidos.

El instrumento 1100 puede acoplarse a un secuenciador y recibir un archivo 1103 metagenómico directamente desde el secuenciador, pero esto no es necesario. El instrumento 1100 también puede recibir el archivo 1103 metagenómico indirectamente de uno o más secuenciadores que no están acoplados al instrumento 1100. Por ejemplo, el instrumento 1100 puede recibir un archivo metagenómico a través de una red de comunicación desde un secuenciador, que puede estar ubicado remotamente. O bien, un archivo 1103 metagenómico, que se ha almacenado previamente en un medio de almacenamiento, como una unidad de disco duro o un medio de almacenamiento óptico, puede introducirse en el instrumento 1100.

Además, el instrumento 1100 puede recibir un archivo 1103 metagenómico o lecturas de fragmentos metagenómicos en tiempo real, inmediatamente después de la secuenciación por un secuenciador o en paralelo con la secuenciación por un secuenciador, pero esto tampoco es necesario. El instrumento 1100 también puede recibir un archivo 1103 metagenómico o fragmentos metagenómicos en un momento posterior. En otras palabras, la caracterización de poblaciones de microorganismos en una muestra realizada por el instrumento 1100 puede realizarse en línea con la recolección de muestras, extracción de fragmentos metagenómicos y secuenciación de fragmentos metagenómicos, pero todas las etapas pueden manejarse por separado y/o en forma escalonada.

El instrumento 1100 puede operar bajo el control de un secuenciador que secuencia los fragmentos metagenómicos extraídos de una muestra, pero no se requiere ningún procesamiento conectado o incluso comunicación directa entre 1100 y un secuenciador. En cambio, la caracterización de poblaciones de microorganismos en una muestra realizada por el instrumento 1100 puede realizarse por separado de la recogida de muestras, extracción de fragmentos metagenómicos y/o secuenciación de fragmentos metagenómicos.

La Fig. 12 es un diagrama de flujo de nivel superior que ilustra un proceso que se puede realizar para caracterizar poblaciones de microorganismos en una muestra. En la etapa S1201, se recoge una muestra. En la etapa S1202, se extraen los fragmentos metagenómicos, que pueden ser ácidos nucleicos y/o proteínas y/o metabolitos. En la etapa S1203, se secuencian los fragmentos metagenómicos y se obtienen las lecturas de los fragmentos metagenómicos. En la etapa S1204, se realiza un proceso de análisis metagenómico para caracterizar las identidades y poblaciones relativas y/o concentraciones de organismos contenidos dentro de la muestra en base a las lecturas del fragmento metagenómico, que puede tener la forma de un archivo metagenómico.

Como se describió anteriormente, el proceso de análisis metagenómico de la etapa S1204 se puede realizar en línea con la recolección de muestras, la extracción del fragmento metagenómico y la secuenciación del fragmento metagenómico de las etapas S1201-S1203, pero todas las etapas se pueden manejar alternativamente por separado y/o de forma escalonada.

En la etapa S1204, el proceso de análisis metagenómico para caracterizar el material biológico en la muestra se puede ejecutar con el instrumento 1100. El archivo 1103 metagenómico secuenciado de pares de bases aleatorios de las lecturas del fragmento metagenómico puede comprender la entrada para la ejecución del proceso de análisis metagenómico por instrumento 1100. La caracterización puede incluir la identificación de las especies y/o subespecies y/o cepas de organismos contenidos en la muestra.

El proceso de análisis metagenómico realiza métodos probabilísticos, que pueden incluir comparar probabilísticamente lecturas de fragmentos metagenómicos con una o más bases de datos genómicas de referencia calificadas para caracterizar la comunidad microbiana de la muestra. Los métodos probabilísticos se pueden realizar en paralelo (es decir, al mismo tiempo) con la secuenciación de los fragmentos metagenómicos, tan rápido como se secuencian los fragmentos metagenómicos (es decir, secuenciación en tiempo real), secuencialmente siguiendo la secuenciación de los fragmentos metagenómicos, o en cualquier momento después que ha completado la secuenciación de los fragmentos metagenómicos.

En algunas realizaciones, el instrumento de secuenciación puede continuar recopilando cadenas de información de secuencia más largas y en paralelo con la comparación. La información de la secuencia posterior también puede compararse y puede aumentar la confianza de la identificación de un genoma o de una especie en la muestra. Es posible que el método no necesite esperar a que la información de secuencia se ensamble de las lecturas cortas en cóntigos más grandes. Sin embargo, como se indicó anteriormente, en algunas realizaciones, todas las lecturas de fragmentos metagenómicos utilizadas en el proceso de análisis metagenómico de la etapa S1204 pueden introducirse como un único archivo metagenómico.

El proceso de análisis metagenómico realizado en la etapa S1204 puede caracterizar la comunidad microbiana de la muestra identificando la comunidad microbiana de la muestra a nivel de especie y/o subespecie y/o cepa con sus concentraciones relativas o abundancia. En particular, los genomas de los organismos contenidos dentro de la muestra pueden identificarse basándose en las lecturas de fragmentos metagenómicos realizando comparaciones probabilísticas para cada una de la pluralidad de lecturas de fragmentos metagenómicos frente a información de secuencia genómica contenida en una o más bases de datos genómicas de referencia.

Mientras que la identificación de microorganismos en una muestra metagenómica, basada en secuenciación directa y emparejamiento probabilístico, es independiente del método de secuenciación o del tipo de máquina, los resultados pueden verse afectados por el error de la máquina del secuenciador y la eficacia/eficiencia de la extracción de materiales para ser secuenciados. De acuerdo con lo anterior, para una precisión relativa en la determinación de poblaciones relativas de organismos a nivel de subespecie o cepa o especie, la etapa S1204 puede incluir normalizar el error de la máquina mediante el uso de un mayor número de lecturas de fragmentos metagenómicos estadísticamente significativos. Por ejemplo, en algunas realizaciones, la extensión de fragmentos (es decir, pasar de lecturas de fragmentos que tienen una longitud de n a lecturas de fragmentos que tienen una longitud de n + 1) y/o la creación de más fragmentos se pueden realizar para aumentar la precisión. Si la identificación de la cepa es crítica para una decisión de tratamiento, entonces el sistema y el método de la invención pueden retroalimentar una solicitud de más secuenciación.

Fig. 13 es un diagrama de flujo que ilustra un proceso ejemplar que se puede realizar en la etapa S1204 para identificar genomas de organismos contenidos dentro de una muestra. Las lecturas del fragmento metagenómico pueden introducirse en el proceso que se muestra en la Fig. 13 en forma de un archivo 1103 metagenómico. En la etapa S1301, el instrumento 1100 puede realizar comparaciones probabilísticas de las lecturas del fragmento metagenómico con la información de secuencia (es decir, lecturas del genoma) contenida en una o más bases de datos genómicas de referencia que contienen identidades genómicas de microorganismos. En particular, las comparaciones probabilísticas pueden comparar cada una de la pluralidad de lecturas de fragmentos metagenómicos con lecturas de genomas de una o más bases de datos de referencia para identificar coincidencias entre una lectura de fragmentos metagenómicos obtenida de la muestra y uno o más genomas de microorganismos contenidos en una base de datos genómica de referencia. Las coincidencias pueden adoptar la forma de correlaciones causales entre un fragmento metagenómico leído y un genoma de la base de datos genómica de referencia y, cuando se detectan, se conserva la correlación causal.

Las comparaciones probabilísticas pueden incluir, pero no se limitan a, emparejamiento perfecto, unicidad de subsecuencia, emparejamiento de patrones, emparejamiento de múltiples subsecuencias dentro de n longitudes, emparejamiento inexacto, semilla y extensión, mediciones de distancia y mapeo de árbol filogenético. Además, las comparaciones probabilísticas pueden utilizar el enfoque bayesiano, el enfoque bayesiano recursivo o el enfoque bayesiano neófito, pero no se limitan a ninguno de estos enfoques.

Las comparaciones probabilísticas pueden determinar que existe una correlación causal entre un fragmento metagenómico leído y una subsecuencia en una base de datos genética de referencia cuando las comparaciones probabilísticas determinan que la lectura de fragmento metagenómico y la subsecuencia de la base de datos genómica de referencia son lo suficientemente similares como para implicar una relación biológica. Además, la subsecuencia genética puede tener primos cercanos de cepas relacionadas o una función biológica similar y, como se supone que la base de datos genómica de referencia está incompleta, el emparejamiento probabilístico también puede considerar que la correlación de fragmentos con primos cercanos también es causal. Al comparar millones de lecturas de fragmentos metagenómicos con miles de millones de subsecuencias, en algunas realizaciones, uno solo espera alrededor de decenas de millones de fragmentos causales entre los millones de billones de comparaciones putativas.

Las comparaciones probabilísticas de la etapa S1301 producen resultados probabilísticos, que en algunas realizaciones pueden tener la forma de un mapa de probabilidad de probabilidades de que estén presentes en la muestra especies y/o cepas de microorganismos dentro de una base de datos genómica de referencia. El mapa de probabilidad puede permitir la correlación de las probabilidades del mapa de probabilidad con poblaciones relativas y/o concentraciones de microorganismos contenidos en la muestra.

En algunas realizaciones, el mapa de probabilidad puede tener una estructura basada en el recuento estadístico y el proceso de correlación con la estructura construida sobre la jerarquía del proceso de correlación. La estructura puede estar relacionada con el grado de parentesco de los genomas diana con otros genomas presentes (es decir, si Shigella está presente, existe un grado de parentesco con Escherichia no patógena presente en la muestra). La relación puede

estar escalonada por el nivel de taxonomía (por ejemplo, cepa, especie, género, etc.). La distancia de la relación se puede representar como color. Por ejemplo, los puntajes únicos (es decir, alta probabilidad) serían rojos y el azul representaría baja probabilidad. A su vez, por ejemplo, este “mapa de colores” puede describirse como un mapa de calor (o mapa de probabilidad) en gradaciones de probabilidad desde rojo (alto) a bajo (azul).

Además, en algunas realizaciones no limitantes, la etapa S1301 puede compensar el error de la máquina normalizando el mapa de probabilidad usando un número mayor de lecturas estadísticamente significativas que el número de lecturas necesarias para una caracterización precisa si no se asumiera ningún error de la máquina. Estadísticamente, la mayoría de las señales se pueden recuperar de un entorno ruidoso mediante la integración de más datos. En una realización, para el error de secuenciación, suponga que la probabilidad de que no haya errores en una secuencia de n largos es (1-p) ^ n. Si n = 20 y p = 0.01, entonces la probabilidad de que no haya errores en una secuencia de 20 pares de bases es del 82 %, mientras que si n = 20 y p = 0.1, esta probabilidad es del 12 %. En igualdad de condiciones, se requieren 7 veces más datos en el segundo caso para hacer declaraciones igualmente sólidas de contenido metagenómico basadas en muestras de fragmentos de n largos.

La compensación del error de la máquina puede complicarse mediante el análisis de la varianza de las distribuciones y la no independencia de las cepas genómicas relacionadas. En un ejemplo artificial, donde A y B son marcadores de genomas α y β; suponga que P (A|A) = P (B|B) = p y P (A|B) = P (B|A) = 1-p; la probabilidad de observar A dado α y la probabilidad de observar B dado β son cada una q; y la mezcla metagenómica de la muestra es rα + (1-r)β. Luego, al primer orden dadas las observaciones de T,

$$\mu(A) = (rqp + (1–r)q(1–p))T = (1\text{-}r\text{-}p+2rp)qT \quad \text{var}(A)$$
$$\text{sigmage}(A) \quad \text{sqrt}\{1\text{-}r\text{-}p–rp)qT\}$$

Por tanto, la importancia de las predicciones de los métodos probabilísticos aumenta exponencialmente con el cuadrado del número de lecturas. En algunas realizaciones, la predicción precisa puede utilizar una estimación de p (por ejemplo, la tasa de error de secuenciación), y el número de lecturas de fragmentos metagenómicos usados en el proceso de análisis metagenómico puede seleccionarse de modo que los errores de secuenciación se normalicen a cerca de cero.

En la etapa S1302, el instrumento 1100 puede integrar (es decir, agregar) las correlaciones causales retenidas mediante, por ejemplo, la cepa genómica para aumentar significativamente la relación señal-ruido de la cepa causalmente presente en la población metagenómica en relación con cepas no causales y así lograr una alta identificación de especies de los genomas de los microorganismos contenidos en la muestra. En otras palabras, las correlaciones causales retenidas para cada una de las cepas genómicas de la base de datos genómica de referencia pueden sumarse.

El método también puede incluir etapas para eliminar la ambigüedad adicional entre cepas estrechamente relacionadas y para detectar la presencia de múltiples cepas relacionadas de concentraciones diferentes. En la etapa S1303, el instrumento 1100 puede crear conjuntos de patrones independientes de inclusión de subconjuntos y exclusión de subconjuntos del conjunto de genomas estrechamente relacionados. Los conjuntos de patrones independientes son patrones n-mero que ocurren en algunos, pero no en todos los genomas de los genomas estrechamente relacionados. Por ejemplo, si los genomas G1-G3 estrechamente relacionados se incluyen en una base de datos genómica de referencia, el instrumento 1100 puede crear los siguientes conjuntos de patrones independientes: (1) patrones n-mero A y B, que ocurren en el genoma G1 pero no en los genomas G2 y G3; (2) patrón C n-mero, que ocurre en el genoma G2 pero no en los genomas G1 y G3; (3) patrones n-mero D, E, F y G, que se encuentran en el genoma G3 pero no en los genomas G1 y G2; (4) patrón H n-mero que ocurre en los genomas G1 y G2 pero no en el genoma G3; y (5) patrones de n-meros G, I y H que ocurren en los genomas G2 y G3 pero no en el genoma G1. Tenga en cuenta que los genomas G1 y G3 pueden no tener marcadores únicos y, como resultado, el conjunto de patrones n-mero que se producen en los genomas G1 y G3 pero no en el genoma G2 es el conjunto nulo.

En su forma más simple, un conjunto de patrones puede ser aquellos n-meros que ocurren en un genoma X pero no en ningún otro. Suponiendo que existen K tales patrones con multiplicidad entre M patrones totales en el genoma y una concentración de C, si hay L patrones en las lecturas de fragmentos, entonces se esperan L*C*K/M = H aciertos entre los K patrones o C ~ H*M/(K*L). En algunas realizaciones, esta estimación puede ajustarse para el error de muestreo y la precisión de la base de datos. Se pueden usar métodos simples de programación lineal, escalada y mínimos cuadrados con conjuntos de patrones que involucran múltiples genomas y para combinar múltiples estimadores para una concentración genómica dada. Un conjunto de patrones también puede ser aquellos n-meros que ocurren en una pluralidad de genomas (por ejemplo, genomas X, Y y Z) pero no en ningún otro.

En la etapa S1304, para cada conjunto de patrones independientes, el instrumento 1100 puede emparejar el conjunto con lecturas diana. Por ejemplo, en la etapa S1304, para cada conjunto de patrones independientes {P_i}, el instrumento 1100 puede comparar los patrones dentro del conjunto con las lecturas del fragmento diana. Si un genoma causal (es decir, un genoma encontrado en la muestra metagenómica) se encuentra entre los genomas del conjunto, entonces todos los patrones dentro del conjunto {P_i} deben encontrarse en las lecturas del fragmento diana, sujeto a un potencial sesgo de muestreo, concentración genómica y variación de tensión. El grado de cobertura del conjunto

de patrones {P _i} por las lecturas diana proporciona, por tanto, una estimación de la concentración para cada genoma que se incluyó en la creación del conjunto de patrones. El resultado de cada emparejamiento entre un conjunto de patrones independiente y el objetivo es una estimación independiente. Además, el resultado de cada uno de los emparejamientos es una estimación independiente de la concentración del genoma en el conjunto que proporciona una estimación detallada de las concentraciones de cepas genómicas incluso para comunidades microbianas estrechamente relacionadas.

En las etapas S1301 y S1302, el instrumento 1100 puede realizar un filtrado primario para determinar qué especies y cepas de una o más bases de datos de referencia pueden estar presentes en la muestra metagenómica. Luego, en las etapas S1303 y S1304, el instrumento 1100 puede realizar un filtrado secundario y terciario para eliminar tanto los falsos negativos como los falsos positivos y para identificar a nivel de deformación lo que está presente en la muestra.

El proceso ejemplar de la etapa S1204 para identificar genomas de microorganismos contenidos dentro de una muestra ilustrada en la Fig. 13 incluye las etapas S1301-S1304,

De hecho, las etapas son separables. Las etapas S1303 y s1304 pueden sembrarse con genomas sin tener que recurrir a las etapas S1301 y S1302. Por ejemplo, las etapas S1303 y s1304 pueden, en cambio, ser sembradas a través de una lista de cepas de V. Cholerae para un estudio de V. Cholerae. Además, la etapa S1303 puede realizarse a ciegas en toda la base de datos como una etapa de precálculo muy grande, aunque también se pueden usar enfoques más dinámicos, como la inclusión de las etapas 1301 y 1302, que requieren menos hardware y almacenamiento.

Además, las etapas mostradas en la Fig. 13 son capaces de determinar cambios en la biodiversidad y capaces de detectar patógenos modificados. Por ejemplo, el método puede tener en cuenta la biodiversidad al identificar (a) elementos genéticos móviles mediante transferencia lateral de genes, recombinación o plásmido u otra inserción de mobiloma; (b) inserciones y eliminaciones; y (c) identificación y detección de cepas parientes cercanas relacionadas por mutación, inserción, eliminación.

Aunque el proceso de análisis metagenómico probabilístico de la etapa S1204 para caracterizar las identidades y poblaciones relativas y/o concentraciones de organismos contenidos en una muestra basada en lecturas de fragmentos metagenómicos se puede realizar mediante muchos procesos, una realización particular no limitante de una prueba El proceso de análisis metagenómico que puede realizar el instrumento 1100 para llevar a cabo la etapa S1204 se describe a continuación con referencias a las Figs. 14-16. La realización particular no limitativa se denomina Motor Comparador.

El Motor Comparador se compone de tres componentes principales: (1) la base de datos del Motor Comparador, (2) procesos de catalogación de sustancias y (3) procesos de identificación y análisis sintáctico de muestras desconocidas. La premisa básica detrás del Motor Comparador es que los datos de secuencia de una sustancia se pueden dividir en palabras y que se puede utilizar un subconjunto de estas palabras para identificar la sustancia original. En un nivel alto, el Motor Comparador toma las sustancias conocidas y crea un catálogo de palabras. Luego, para analizar muestras de contenido desconocido, el Motor Comparador toma la información de secuencia de las muestras de contenido desconocido y divide esa información de secuencia en una lista de palabras. A continuación, el Motor Comparador toma las palabras de la muestra y las empareja con las palabras del catálogo. El resultado de la coincidencia se resume al contar el número de palabras que coinciden por sustancia conocida.

La base de datos del Motor Comparador se puede construir en un servidor, tal como Microsoft SQL Server 2005. Las herramientas de catalogación, análisis sintáctico y análisis de palabras pueden estar escritas en un lenguaje de programación, tal como Java. Se puede utilizar la interfaz de usuario principal, como las Java Server Pages que se ejecutan en un servidor de aplicaciones Tomcat. Además, todo el sistema puede estar alojado en una única máquina servidor, como en una única máquina Windows Server.

En general, el diseño de la base de datos del sistema Motor Comparador puede ser muy simple. La simplicidad de la base de datos facilita el rendimiento. Por consiguiente, la base de datos puede diseñarse para mantener la sobrecarga asociada con las relaciones de datos subyacentes al mínimo.

El modelo de datos puede tener las siguientes estructuras de datos de alto nivel: (1) Catálogo, (2) Sustancia, (3) Categoría, (4) Categoría de Sustancia, (5) Palabras Sustanciales, (6) Desconocido y (7) Palabras Desconocidas. La base de datos se puede dividir en Catálogos lógicos. Por ejemplo, un Catálogo podría representar información de secuencia que represente las secuencias asociadas con la lista de 8 patógenos, o podría ser todas las especies entéricas. Una Sustancia es una secuencia conocida (por ejemplo, identidad genómica) a la que se le asigna un nombre. Normalmente, una Sustancia puede ser una secuencia asociada con una especie individual, como Vibrio cholerae, pero una Sustancia también puede ser una secuencia asociada con una cepa o subespecie individual conocida. Una Categoría puede definir un nombre que puede asignarse a una o más Sustancias con el fin de agruparlas. Por ejemplo, una Categoría puede permitir que las especies se agrupen en un género. Una Categoría de Sustancia puede definir la relación entre una Sustancia y una Categoría. Una Sustancia puede tener una o más Categorías y una Categoría puede tener una o más Sustancias. La estructura de Palabras Sustanciales puede ser,

esencialmente, el diccionario de palabras conocidas. Puede contener todas las palabras identificadas en los datos de secuencia para cada Sustancia. La estructura desconocida puede etiquetarse como datos de secuencia de contenido desconocido. Esto puede representar las lecturas del fragmento metagenómico obtenidas de una muestra que contiene material genético de una pluralidad de organismos, que puede tener la forma de un archivo 1103 de entrada metagenómico.

El archivo 1103 de entrada metagenómico puede contener datos de secuencia en un formato particular. Por ejemplo, se puede utilizar 454, Illumina o cualquier formato FASTA o FASTQ. La estructura de Palabras Desconocidas puede ser palabras contenidas en los datos de secuencia del contenido desconocido (es decir, una muestra que contiene material genético de una pluralidad de organismos).

El proceso de catalogación de sustancias en una realización se refiere al procesamiento que rodea la carga de archivos de secuencia alineados y secuenciados, asociados con especies conocidas, en la base de datos. Una realización ilustrativa del proceso de catalogación de sustancias se muestra en la Fig. 14. Las entradas al proceso de catalogación de sustancias pueden ser (1) datos de secuencia que representan una sola Sustancia conocida, (2) un Catálogo en el que se agregarán los registros de la Sustancia y (3) una Categoría (por ejemplo, género o especie o cepa a la que pertenece la Sustancia. Las entradas en el proceso de catalogación de sustancias también pueden incluir parámetros de construcción de palabras, pero también es posible que los parámetros de construcción de palabras se establezcan de forma predeterminada. Los datos de secuencia para la sustancia pueden tener la forma de un archivo de entrada, que puede tener un formato FASTA o FASTQ. Los datos de secuencia para la sustancia pueden ser lecturas del genoma obtenidas de una o más bases de datos genómicas de referencia que contienen identidades genómicas de organismos.

En la etapa S1401, el proceso de catalogación de sustancias determina si el Catálogo identificado en el que se agregarán los registros de la Sustancia es un nuevo Catálogo. Si el Catálogo identificado es un Catálogo nuevo, en la etapa S1402, se crea el Catálogo identificado. La etapa S1402 puede generar los detalles del catálogo creado. Luego, en la etapa S1403, se selecciona el catálogo creado. De lo contrario, si el Catálogo identificado no es un Catálogo nuevo y ya existe, el Catálogo existente se selecciona en S1403. En la etapa S1404, el proceso de catalogación de sustancias determina si la Categoría identificada a la que pertenece la sustancia es una Categoría nueva. Si la Categoría identificada es una Categoría nueva, en la etapa S1405, se crea la Categoría identificada. La etapa S1405 también puede generar los detalles de la categoría creada. Luego, en la etapa S1406, se selecciona la Categoría creada. De lo contrario, si la Categoría identificada no es una categoría nueva y ya existe, la Categoría existente se selecciona en S1406.

La etapa S1406 también agrega la Sustancia a la Categoría seleccionada leyendo los datos de secuencia que representan la Sustancia conocida y asociándola con el catálogo seleccionado. Luego, en la etapa S1407, el proceso inicia el proceso de búsqueda de palabras preparándose para la detección de palabras. La etapa S1407 puede leer cualquier parámetro de construcción de palabras introducido en el proceso. A continuación, en la etapa S1408, los datos de la secuencia de lectura que representan la Sustancia conocida se analizan en busca de palabras. En esta etapa, el proceso divide la secuencia en fragmentos, llamados palabras, interroga la lista de palabras o vector resultante y selecciona palabras para agregar a la estructura, tabla o archivo de Palabras Sustanciales. El proceso de búsqueda de palabras se describe con mayor detalle a continuación. La etapa S1408 puede generar la tabla de Palabras Sustanciales. Luego, en la etapa S1409, el proceso puede actualizar el Catálogo agregando la tabla de Palabras Sustanciales y pueden generarse varios informes de comparación. Por ejemplo, en la etapa S1409, el proceso puede generar informes de propiedades de Sustancia y similitudes de Sustancia a Sustancia.

El proceso de catalogación de sustancias puede repetirse para varias Sustancias y, de esta manera, la base de datos puede llenarse con uno o más Catálogos, cada uno de los cuales tiene una o más Categorías y una o más Sustancias. Los Catálogos resultantes de Palabras de sustancias conocidas pueden utilizarse para identificar sustancias desconocidas utilizando palabras generadas a partir de sustancias desconocidas.

En otras realizaciones, el proceso de catalogación de sustancias también puede almacenar todas las palabras de referencia generadas a partir de todas las sustancias conocidas catalogadas (es decir, todas las palabras de cada una de las estructuras de Palabras Sustanciales generadas para cada Sustancia catalogada) en una tabla hash. Una tabla hash de este tipo poblada con todas las palabras de referencia puede indicar cuántas de las sustancias catalogadas tienen cada una de las palabras de referencia. Por tanto, la tabla hash puede informar qué palabras de referencia son únicas (es decir, que pertenecen a una sola sustancia) y qué palabras de referencia son comunes a más de una sustancia conocida. De esta manera, se pueden usar una o más tablas hash para indicar las palabras de referencia que son exclusivas de un género, especie, subespecie y/o cepa. Saber si una palabra es única o común puede usarse en la identificación de las sustancias desconocidas de la muestra.

El proceso de análisis e identificación de muestras desconocidas se refiere al procesamiento que rodea al análisis sintáctico de los datos de secuencia del contenido desconocido en palabras y la comparación de las palabras del contenido desconocido con las palabras de las sustancias conocidas para identificar el contenido desconocido. Una realización ilustrativa del proceso de identificación y análisis sintáctico de muestras desconocidas se muestra en la Fig. 15. Los datos de secuencia que representan el contenido desconocido se ingresan en el proceso de análisis e

identificación de muestras desconocidas. Los datos de secuencia pueden representar las lecturas del fragmento metagenómico obtenidas de una muestra que contiene material genético de una pluralidad de organismos y pueden tener la forma de un archivo 1103 de entrada metagenómico. El archivo 1103 de entrada metagenómico puede contener datos de secuencia en un formato particular. Por ejemplo, se puede utilizar 454, Illumina o cualquier formato FASTA o FASTQ. Las entradas también pueden incluir parámetros de construcción de palabras, pero también es posible que los parámetros de construcción de palabras se establezcan de forma predeterminada.

En la etapa S1501, el archivo de secuencia que representa el contenido desconocido de una muestra se lee y se agrega a la base de datos del Motor Comparador como Desconocido. La etapa S1501 también puede generar detalles de la muestra desconocida. Los detalles de la muestra desconocida pueden incluir la fecha y hora en que se tomó la muestra, la ubicación geográfica en la que se tomó la muestra y/u otros metadatos similares. En la etapa S1502, el proceso inicia el proceso de búsqueda de palabras preparándose para la detección de palabras. La etapa S1502 puede leer cualquier parámetro de construcción de palabras introducido en el proceso. A continuación, en la etapa S1503, los datos de la secuencia de lectura que representan los contenidos desconocidos en la estructura Desconocida se analizan en busca de palabras. En esta etapa, el proceso divide la secuencia en fragmentos, llamados palabras, interroga la lista de palabras o vector resultante y selecciona palabras para agregar a la estructura, tabla o archivo de Palabras Desconocidas. El proceso de búsqueda de palabras se describe con mayor detalle a continuación. La etapa S1503 puede generar la tabla de Palabras Desconocidas.

Una vez que se completa el análisis sintáctico de palabras de la etapa S1503, en la etapa S1504, las palabras asociadas con el Desconocido se comparan con palabras en el diccionario de sustancias conocidas catalogadas (es decir, las palabras en las estructuras de Palabras Sustanciales de un Catálogo). En la etapa S1504, el proceso crea un índice para hacer referencias cruzadas de palabras desconocidas a palabras catalogadas en el diccionario. La etapa S1504 puede generar palabras desconocidas que se han emparejado con palabras del diccionario de sustancias conocidas. Las coincidencias pueden ser coincidencias exactas y/o coincidencias inexactas (es decir, parciales o similares).

En algunas realizaciones, el proceso de comparación de S1504 busca coincidencias exactas en palabras únicas para una sola Sustancia conocida, así como coincidencias totales con cualquier palabra en el diccionario. La etapa S1504 también puede calcular una puntuación de relación ponderada asignando un factor de ponderación a los tipos de coincidencias. Por ejemplo, las coincidencias exactas con palabras únicas para una Sustancia en particular tendrían el mayor peso, las coincidencias exactas con palabras exclusivas de Categoría tendrían un peso menor, las coincidencias exactas con palabras muy comunes tendrían un peso aún menor, etc. De esta manera, el proceso de comparación de S1504 puede calcular una puntuación que considera el valor relativo de una coincidencia en particular. En otras realizaciones, el proceso de comparación de S1504 puede realizar un emparejamiento parcial usando, por ejemplo, el enfoque de Levenshtein. Varias técnicas de búsqueda que podrían utilizarse incluyen un motor de base de datos de búsqueda basada en SQL, una búsqueda de comparación de cadenas Simple, una Búsqueda de Expresión Regular con la capacidad de buscar coincidencias inexactas; el Enfoque de Distancia de Levenshtein, un Clasificador Bayesiano, un Clasificador de Vectores y/o un Motor de Búsqueda Tipo Google. Algunas realizaciones que utilizan el enfoque de Distancia de Levenshtein pueden construir primero un conjunto de coincidencias que son similares y luego usar el enfoque de Distancia de Levenshtein para calcular la puntuación de similitud.

En la etapa S1505, el proceso compila las coincidencias entre las palabras de la muestra desconocida y las palabras de las sustancias conocidas catalogadas. Por ejemplo, la etapa S1505 puede calcular un rango para las sustancias conocidas que ayuda a separar las cepas que aparecen debido a la similitud a una mejor coincidencia de las que están realmente presentes en los contenidos desconocidos de la muestra. La clasificación puede, por ejemplo, basarse en, para cada sustancia conocida, la suma del número total de coincidencias entre las palabras de la muestra desconocida y las palabras de la sustancia conocida, una suma del número de coincidencias distintas, el total número de palabras únicas y/o número de coincidencias únicas distintas. La etapa S1505 también puede producir y sacar las coincidencias compiladas y/o palabras estrechamente relacionadas en una hoja de cálculo de salida.

El Motor Comparador también puede realizar un análisis comparativo entre dos o más elementos para calcular la relación entre dos o más secuencias. Las secuencias pueden ser entre cepas de sustancia conocida o entre muestras desconocidas. El análisis comparativo compara listas de palabras y se informa el recuento superpuesto. El análisis comparativo puede calcular una puntuación de relación relativa basada en cuántas palabras están disponibles en cada conjunto y cuántas coinciden. Por ejemplo, el parentesco de conjuntos de palabras puede vincularse a las cepas (palabras) en la base de datos de referencia para producir identidades relativas. El valor de unicidad de la palabra se puede utilizar para la probabilidad, y se puede incluir el significado biológico de la palabra. Las palabras compartidas entre organismos pueden tener una puntuación más baja, y en la puntuación se puede considerar el grado de compartición entre el número de cepas diferentes.

Una realización ilustrativa del proceso de búsqueda de palabras, que puede usarse en el proceso de catalogación de sustancias y/o el proceso de análisis sintáctico de muestras desconocidas, se muestra en la Fig. 16. El proceso de búsqueda de palabras puede comenzar en la etapa S1601 preparándose para la detección de palabras. La etapa S1601 corresponde a la etapa S1407 del proceso de catalogación de sustancias mostrado en la Fig. 14 y a la etapa S1502 del proceso de análisis sintáctico de muestras desconocidas mostrado en la Fig. 15. La etapa S1601 puede

leer cualquier parámetro de construcción de palabras introducido en el proceso. En esta realización, los parámetros de construcción de palabras pueden incluir una longitud de palabra mínima. El proceso de búsqueda de palabras también puede establecer la longitud mínima de palabras en una longitud mínima predeterminada de palabras, que puede ser, por ejemplo, 19 letras.

A continuación, en la etapa S1602, el proceso de búsqueda de palabras lee datos de secuencia, que pueden representar una sustancia conocida o una muestra desconocida. En la etapa S1603, el proceso determina si hay más registros (es decir, datos de secuencia) que necesitan procesamiento. Si hay más registros, el proceso pasa a la etapa S1604, donde los datos de secuencia se analizan en busca de palabras. En la etapa S1604, el proceso puede realizar, por ejemplo, cuatro pasadas, S1604a a S1604d, a través de los datos de secuencia. En cada una de las pasadas S1604a a S1604d, se utiliza una letra de secuencia diferente como carácter de límite. Por ejemplo, la pasada S1604a puede usar “A” como carácter límite, la pasada S1604b puede usar “C” como carácter límite, la pasada S1604c puede usar “T” como carácter límite y la pasada S1604d puede usar “G” como carácter límite. En las pasadas S1604a a S1604d, el proceso de búsqueda de palabras puede dividir la secuencia en fragmentos, llamados palabras, en el carácter límite de la palabra.

En la etapa S1605, cada una de las palabras generadas en las pasadas S1604a a S1604d de la etapa de análisis sintáctico S1604 se selecciona para procesamiento adicional. Por ejemplo, en la etapa S1606, el proceso determina si cada palabra tiene una longitud igual o mayor que la longitud mínima de palabra, que puede establecerse de acuerdo con los parámetros de construcción de palabras de entrada o por defecto. Si la etapa S1606 determina que una palabra cumple el requisito de tamaño mínimo, la etapa S1607 puede determinar si la palabra se encontró previamente. Si las etapas S1606 y S1607 determinan que una palabra tiene una longitud igual o mayor que la longitud mínima de palabra y no se ha encontrado previamente, la palabra se agrega a una lista de palabras en la etapa S1608. En la etapa S1609, el proceso determina si hay más palabras que no se han procesado. Si hay más palabras, el proceso repite las etapas S1605 a S1609 hasta que se hayan procesado todas las palabras generadas por las pasadas S1604a a S1604d. Una vez que se han procesado todas las palabras, y la etapa S1609 determina que no hay más palabras, el proceso vuelve a la etapa S1602, donde se lee cualquier dato de secuencia adicional.

Si la etapa S1603 determina que no hay más registros (es decir, datos de secuencia) que necesitan procesarse, el proceso pasa a la etapa S1510, donde la lista de palabras se guarda en la base de datos del Motor Comparador. Si el proceso de búsqueda de palabras se ejecuta mediante el proceso de catalogación de sustancias, la lista de palabras guardadas se puede utilizar como archivo o tabla de Palabras Sustanciales. Si el proceso de búsqueda de palabras se ejecuta mediante el proceso de análisis de muestras desconocidas, la lista de palabras guardadas se puede utilizar como archivo o tabla de Palabras Desconocidas.

En algunas realizaciones, la longitud mínima de palabra utilizada por el proceso de búsqueda de palabras es mayor o igual a 18 letras y menor o igual a 27 letras. En algunas realizaciones, la longitud mínima de palabra utilizada por el proceso de búsqueda de palabras es mayor o igual a 19 letras y menor o igual a 25 letras. En una realización particular, la longitud mínima de palabra utilizada por el proceso de búsqueda de palabras es de 19 letras.

En algunas realizaciones, los datos de secuencia de la entrada de contenido desconocido al problema serán una gran colección (por ejemplo, 10 ^ 9) de secuencias cortas de k letras, k-meros, de la muestra ambiental, y la salida será ser los organismos que puedan estar en la muestra ambiental. Si bien algunos k-meros se comparten entre múltiples organismos, los que son exclusivos de una especie, subespecie y/o cepa de organismo, o de su género, serán los más valiosos.

El número de k-meros de la muestra que pueden provenir de una sustancia desconocida (por ejemplo, un objetivo peligroso) para identificar esa sustancia y si es probable que el objetivo se pueda identificar con un solo k-mero se puede tener en cuenta al seleccionar la longitud a la que establecer la longitud mínima de palabra. Como se señaló anteriormente, los k-meros se pueden comparar en una tabla hash de todos los k-meros que se encuentran en todas las sustancias conocidas catalogadas, que podrían ser, por ejemplo, todas las secuencias bacterianas en una base de datos genómica de referencia, como GenBank.

En algunas realizaciones no limitantes, los datos experimentales sugieren que una longitud de palabra k igual a 19 letras (es decir, k = 19) es eficaz como longitud mínima de palabra. En un ejemplo no limitativo, se analizó una muestra de 30 secuencias completas de diferentes cepas de bacterias de GenBank. Aproximadamente el 5 % de los 17-meros se compartieron entre múltiples organismos. Sin embargo, para los organismos que no tienen un pariente común en la muestra, la especie podría ser identificada de forma única por más del 99 % de los 19-meros en su genoma. Por ejemplo, el 99.4 % de todos los 19-meros eran exclusivos de Chlamydia trachomatis (1.1 Mb) y el 99.9 % eran exclusivos de Synechocystis (3.5 Mb). Las longitudes un poco más largas, como 25 letras, también pueden funcionar bien. Sin embargo, en este ejemplo no limitante, longitudes de longitudes mucho más largas no funcionaron tan bien porque diferencias menores entre cepas similares pueden haber impedido las coincidencias. Por lo tanto, en una realización no limitante, solo se necesita un 19-mero para identificar las bacterias en la gran mayoría de los casos, y la tabla hash podría informar cuántas bacterias tienen el 19-mero, en los raros casos en los que no lo es en una cepa única.

En este ejemplo no limitante, de estas 30 bacterias, había dos pares de organismos estrechamente relacionados: E. coli W3110 y E. coli K-12 y Helicobacter pylori 26695 y Helicobacter pylori J99. En ambos casos, los genomas de organismos relacionados tenían casi la misma longitud (4.5 Mb para E. coli y 1.7 Mb para H. pylori). La diversidad entre estos pares fue bastante variada. Las dos secuencias de H. pylori compartían sólo el 40 % de sus 19-meros con la otra, mientras que las dos cepas de E. coli eran prácticamente idénticas.

Los resultados preliminares de este ejemplo no limitante indican que existe una capacidad sustancial para determinar al menos el género de bacterias en una muestra ambiental y en muchos casos la especie con solo un 19-mero. Cuando se utilizan múltiples 19-meros, la detección mejorada da como resultado una posibilidad mucho mayor de detección precisa a nivel de especie, subespecie o cepa. Incluso para casos extremos como E. coli, donde el 99.9 % de los 19-meros se comparten entre los miembros, múltiples 19-meros mejorarán en gran medida la detección de la cepa de E. coli de la que provienen los 19-meros. Por ejemplo, si se extraen 1000 19-meros distintos de una muestra con solo una de las dos especies de E. coli, existe un 63 % de probabilidad (1 - 1/e) de que al menos uno de ellos sea exclusivo de la especie E. coli. En comparación, para dos secuencias de H. pylori, cinco 19-meros permitirán la especificación de una sola elección con una probabilidad del 99 %.

En una realización, cuando se usan 19 letras como longitud mínima de palabra, el proceso de catalogación de sustancias crea un Catálogo de 19-meros (es decir, palabras que tienen una longitud mínima de 19 letras) para cada genoma conocido (por ejemplo, sustancia). Las palabras se construyen a partir de combinaciones de 19-meros de 3 nt (nucleótidos). Usando la realización del proceso de búsqueda de palabras mostrado en la Fig. 16, se realizan cuatro pases, por ejemplo, a través de cada genoma para tener en cuenta la ruptura de la secuencia de secuencia por cada uno de los cuatro nucleótidos (por ejemplo, A, C, T y G). El cuarto nt puede ser la etiqueta de alternancia de pares de bases para la palabra. Esto crea interrupciones en la secuencia para comparación. Esto limita el tamaño de los datos para la coincidencia y aumenta la velocidad de las comparaciones entre los catálogos de los diccionarios del genoma de referencia y el archivo de secuencia de muestra analizado en palabras. En algunas realizaciones, no se cuentan las duplicaciones de palabras. En algunas realizaciones, una palabra se puntúa cuando hay una coincidencia perfecta entre una palabra de archivo de secuencia de muestra y una palabra de referencia. La probabilidad de una puntuación correcta es una de cuatro posibilidades: a) objetivo correcto, b) genoma estrechamente relacionado, c) acierto accidental, d) secuencia transferida lateralmente entre genomas.

Las concentraciones relativas de los organismos en la muestra pueden determinarse, por ejemplo, comparando el número de coincidencias para cada una de las sustancias conocidas catalogadas. Las concentraciones relativas pueden determinarse adicional o alternativamente considerando la frecuencia con la que aparece una palabra coincidente.

En la identificación de genomas de microorganismos contenidos dentro de una muestra basada en lecturas de fragmentos metagenómicos usando los procesos de análisis metagenómico probabilístico descritos anteriormente, el número de longitudes de lectura desplegadas para identificar poblaciones relativas depende de la abundancia relativa y concentración de los microorganismos de la muestra, el tamaño del genoma microbiano, la profundidad de cobertura y la precisión de la secuencia. La precisión de la secuencia se ve afectada por múltiples factores que pueden incluir, por ejemplo, la preparación de la muestra, el contexto de la secuencia, la química de secuenciación, el software de llamada de base y el error de la máquina del método de secuenciación implementado para producir el archivo metagenómico (es decir, la precisión del instrumento).

En una realización, para un error de máquina de menos del 1 %, para un genoma que contiene 5 millones de pares de bases, el sistema y el método de la presente invención pueden ser capaces de (i) identificar con precisión cepas que tienen una concentración de 0.001 o mayor utilizando un archivo metagenómico que contenga tan solo mil lecturas cortas de fragmentos metagenómicos de una muestra metagenómica, y (ii) realizar la identificación de especies utilizando un archivo metagenómico que contenga tan solo aproximadamente cien lecturas cortas de fragmentos metagenómicos. De acuerdo con lo anterior, el sistema y método de la presente invención es capaz de realizar la identificación del nivel de cepa para un nivel de detección del 0.1 % usando un archivo metagenómico que contiene tan solo aproximadamente un millón de lecturas totales de fragmentos metagenómicos de la muestra metagenómica. Además, tan solo diez mil lecturas pueden dar una granularidad de nivel de especie a niveles de detección del 1 %. Sin embargo, se observa que la fidelidad de la base de datos genómica de referencia puede ser una fuente de error. Si una cepa de un microorganismo presente en la muestra no está presente en la base de datos de referencia, es posible que no sea posible identificar la cepa. Además, si la cepa de microorganismo presente en la muestra está en la base de datos de referencia, entonces el número de cepas para la especie de microorganismo puede afectar el nivel de error. El número de cepas y sus concentraciones relativas encontradas dentro del objetivo también puede afectar el nivel de error.

Para un error de máquina de menos del 10 %, siendo todo lo demás igual, el sistema y el método de la presente invención pueden ser capaces de realizar (a) la identificación de especies usando un archivo metagenómico que contenga tan solo varios cientos de lecturas de fragmentos metagenómicos y (b) identificación del nivel de cepa con estimaciones de concentración utilizando un archivo metagenómico que contiene tan solo decenas de miles de lecturas de fragmentos metagenómicos. Sin embargo, en algunas realizaciones, dependiendo de la longitud, se puede usar un

archivo metagenómico que contiene cientos de miles o millones de fragmentos metagenómicos para aumentar la identificación de la cepa y la confiabilidad de la estimación de la concentración.

En determinadas realizaciones, la precisión estimada del sistema y método de la presente invención para la identificación a nivel de especie puede ser superior al 98 %. La precisión estimada del sistema y método de la presente invención para la identificación a nivel de cepa puede ser superior al 92 % siempre que la base de datos genómica de referencia contenga la cepa. En un ejemplo extremo, cuando hay varias cepas muy estrechamente relacionadas que difieren en menos de 100 pares de bases en 1 millón (por ejemplo, cepas relacionadas de B. anthracis), la precisión de la identificación cercana al vecino puede ser muy alta. Además, se pueden utilizar más iteraciones con herramientas algorítmicas adicionales para identificar la cepa precisa con una precisión superior al 92 %. Por ejemplo, los métodos probabilísticos pueden utilizar firmas específicas integradas de las cepas estrechamente relacionadas.

Actualmente no hay disponibles otros métodos y sistemas para proporcionar niveles similares de precisión.

En un ejemplo no limitativo, el sistema y los métodos de la presente invención se usaron para identificar patógenos en muestras metagenómicas de pacientes ingresados en un hospital para ser tratados por enfermedad diarreica. La causalidad de la enfermedad se identificó en todos los casos mediante secuenciación directa de ADN de la muestra metagenómica seguida de la identificación del genoma de acuerdo con una realización de la presente invención usando emparejamiento probabilístico.

Los resultados de la identificación del genoma a través de la comparación probabilística se compararon ciegamente con métodos ortogonales de 26 bioensayos estándar que demostraron una precisión del 100 %. La Tabla 1 a continuación muestra los resultados de la comparación de 9 muestras tomadas de los pacientes ingresados para el tratamiento de una enfermedad diarreica. Algunos patógenos, como Helicobacter pylori o comunidad de patógenos, que se caracterizaron por la identificación del genoma de la presente invención, pero no fueron detectados por la batería de métodos convencionales que constituyen un conjunto de 26 bioensayos basados en dianas bacterianas, virales y parasitarias predeterminadas. Además, en relación con la secuenciación directa y la identificación del genoma de la presente invención, los métodos convencionales requieren mucho trabajo y tiempo. Por ejemplo, los métodos convencionales incluyen serotipificación clásica, cultivo, moleculares, microscópicos e inmunoensayos y requieren un tiempo sustancialmente mayor en comparación con la secuenciación metagenómica y la identificación de poblaciones de la presente invención. Incluso para las muestras en las que se identificaron etiologías probables, el análisis convencional no proporciona una modalidad de tratamiento definible y eficiente. Por el contrario, la amplitud y profundidad de la secuenciación directa del ADN y el perfil completo de la muestra a lo largo de la etiología lograda por la presente invención demuestran el agente causal dominante y, en muchos casos, los agentes causales que son una comunidad de patógenos en lugar de un solo patógeno. Este hallazgo de infección polimicrobiana reitera la importancia de los métodos de diagnóstico dinámicos y sus ventajas sobre los métodos estáticos concurrentes que se utilizan para el diagnóstico de enfermedades.

Tabla 1: Etiología de las muestras de enfermedades diarreicas identificadas por métodos convencionales frente a la secuenciación directa e identificación de la presente invención

Muestra No.	Etiología identificada por una batería de 26 bioensayos	Etiología determinada por secuenciación directa en combinación con identificación bioinformática para un diagnóstico médico rápido
1	V. parahaemolyticus	Vibrio parahaemolyticus
		Shigella dysenteriae Sd 197
		Shigella CDC 3083-94
		Escherichia coli HS
		Eubacterium eligens ATCC 27750
		Veillonella parvula DSM 2008
2	V. parahaemolyticus	Leuconostoc citreum KM20
		Bacteroides vulgatus ATCC 8482
		Citrobacter rodentium ICC 168
		Eubacterium rectale ATCC 33656
		Escherichia coli SMS-3-5

		Shigella boydii CDC 3083-94
		Leuconostoc citreum KM20
		Eubacterium eligens ATCC 27750
		Bacteroides vulgatus ATCC 8482
		Klebsiella pneumoniae MGH 78578
		Veillonella parvula DSM 2008
3	Shigellae	Shigella sonnei Ss046
		Shigella flexneri 2a 2457T
		Escherichia coli 55989
		Klebsiella pneumoniae MGH 78578
		Citrobacter rodentium
4	Shigellae	Shigella sonnei Ss046
		Escherichia coli B REL606
		Escherichia coli K 12 substr W3110
		Bacteroides vulgatus ATCC 8482
		Escherichia coli ATCC 8739
		Klebsiella pneumoniae MGH 78578
		Eubacterium rectale ATCC 33656
		Eubacterium eligens ATCC 27750
		Parabacteroides distasonis 8503
5	Vibrio cholerae O1 & Shigellae	Vibrio cholerae O1
		Shigella sonnei Ss046
		Escherichia coli 55989
		Escherichia coli B str. REL606
		Shigella flexneri 2a str. 2457T
		Escherichia coli DH1
6	Ningún patógeno identificado	Bacteroides fragilis YCH46
		Escherichia coli DH1
		Klebsiella pneumoniae MGH 78578
		Streptococcus pyogenes MGAS2096
		Bacteroides vulgatus ATCC 8482
		Eubacterium rectale ATCC 33656
		Parabacteroides distasonis ATCC 8503
		Bacteroides thetaiotaomicron VPI-5482
7	Ningún patógeno identificado	Klebsiella pneumoniae MGH 78578
		Shigella boydii CDC 3083 94

		Escherichia coli SE11
		Escherichia coli BL21 DE3
		Eubacterium rectale ATCC 33656
		Clostridium novyi NT
		Cytophaga hutchinsonii ATCC 3340
		Mycoplasma pulmonis UAB CTIP
		Psychromonas ingrahamii 37
8	Ningún patógeno identificado	Bacteroides fragilis YCH46
		Klebsiella pneumoniae MGH 78578
		Escherichia coli UMN026
		Eubacterium rectale ATCC 33656
		Eubacterium eligens ATCC 27750
		Bacteroides thetaiotaomicron VPI-5482
		Bacteroides vulgatus ATCC 8482
		Parabacteroides distasonis ATCC 8503
		Veillonella parvula DSM 2008
		Citrobacter rodentium ICC 168
9	Ningún patógeno identificado	Helicobacter pylori G27
		Escherichia coli SMS-3-5
		Escherichia coli UMN026
		Candidatus Sulcia muelleri SMDSEM
		Buchnera aphidicola str. Cc

A partir de un análisis separado de muestras de diarrea (principalmente cólera) de pacientes que experimentaban diarrea más grave, la caracterización de los microorganismos mediante la presente invención permitió concluir que, para estos pacientes, había una mezcla de patógenos (por ejemplo, Vibrio cholerae y Giardia lamblia), y la gravedad de la enfermedad presentada probablemente no fue causada por un único patógeno (por ejemplo, Vibrio cholerae), sino por una mezcla de patógenos que actúan de manera sinérgica. Este tipo de etiología sinérgica habilitada por la caracterización realizada por la presente invención, donde un patógeno atribuye la causalidad de la enfermedad y otros elevan la gravedad de la enfermedad, es capaz de proporcionar información invaluable en el manejo y control de la enfermedad que está más allá del alcance de cualquier tecnología de diagnóstico existente.

Uno de los métodos estándar para identificar bacterias es usar la robustez del rDNA 16S para la colocación taxonómica. La evaluación de la población 16S se dirige a la región del espaciador transcrito intergénico (ITS) del gen ARNr 16S-23S. Sin embargo, el rDNA 16S se limita a la resolución a nivel de género. En otras palabras, el rDNA 16S identifica bacterias solo a nivel de género, familia y orden. Cada género, familia y orden constituye un gran número de especies que representan tanto a los comensales como a los patógenos. Además, el rDNA 16S solo es específico típicamente a nivel familiar y puede incluso tener un rendimiento inferior a la identificación a nivel de género.

Las Figs. 17A-17E ilustran las medidas de población relativa de rDNA 16S en comparación con la secuenciación directa de DNA con identificación genómica de la presente invención. La Fig. 17A muestra mediciones de población usando rDNA 16S, y las Figs. 17B-17E muestran mediciones de poblaciones usando la secuenciación directa de ADN con identificación genómica de la presente invención. Las Figs. 17C-17E muestran especies y cepas del género Clostridium, género Bacteroides y género Escherichia/Shigella, respectivamente, identificadas mediante la secuenciación directa de ADN con identificación genómica de la presente invención, junto con sus concentraciones relativas.

Como se muestra en las Figs. 17A-17E, la secuenciación directa de ADN con identificación metagenómica de la presente invención clasificó además cada género identificado por 16S según los niveles de especie y/o cepa. Por ejemplo, el género Clostridium incluye especies comensales, pero también incluye cuatro especies principales responsables de enfermedades en humanos: Clostridium diffcile, Clostridium botulinum, Clostridium perfringens y Clostridium tetani. Aunque 16S identificó que la muestra contenía bacterias del género Clostridium, como se muestra en la Fig. 17C, la secuenciación directa de ADN con identificación metagenómica de la presente invención identificó Clostridium phyiofermentans ISDg, Clostridium diffcile 630, Clostridium beijerincki beijerincki NCIMB 8052, Clostridium thermocellum ATCC 27405, Clostridium kluyveri DSM A strumidium 555, Clostridium kluyveri DSM struminium 555. ATCC 3502, Clostridium acetobutylicum ATCC 824, Clostridium perfringens str. 13 y Clostridium tetani E88. Además, la secuenciación directa de ADN con identificación metagenómica de la presente invención identificó adicionalmente las concentraciones de las especies/cepas de Clostridium identificadas con respecto a cada una y con respecto a otros microorganismos de la muestra.

Con respecto a los Bacteroides, los Bacteroides son normalmente mutualistas y solo unas pocas especies (por ejemplo, B. fragilis) son patógenos humanos oportunistas. Similar a los resultados con el género Clostridium, mientras que 16S identificó que la muestra contenía bacterias del género Bacteroides, como se muestra en la Fig. 17D, la secuenciación directa de ADN con identificación genómica de la presente invención identificó Bacteroides thetaiotaomicron VPI-5482, Bacteroides vulgatus ATCC 8482, Bacteroides fragiles 638R, Bacteroides fragiles ATCC 25285, Bacteroides fragiles YCH46, Bacteroides fragiles NCTC 9343 y Bacteroides uniforomicides. Aquí nuevamente, la secuenciación directa de ADN con identificación metagenómica de la presente invención identificó adicionalmente las concentraciones de las especies/cepas de Bacteroides identificadas con respecto a cada una y con respecto a otros microorganismos de la muestra.

Las concentraciones relativas en las Figs. 17B-17E se basaron en el número de “aciertos” observados para un conjunto particular de longitudes de lectura pertenecientes a un microorganismo específico. Cuanto menor sea el número de aciertos, menor será la concentración relativa. La precisión a concentraciones relativas bajas se basa en puntuaciones de probabilidad determinadas comparando la longitud de lectura con las bases de datos genómicas. Dada una estimación del error de procesamiento del sistema, el porcentaje de aciertos observados de un conjunto dado de patrones esperados produce una estimación de concentración con barras de error. La presente invención puede emplear múltiples conjuntos independientes que permitan obtener una desambiguación metagenómica mejorada, a saber, la detección e identificación de especies y cepas.

La Figura 18 compara la concentración relativa observada y la concentración real en una muestra con el número relativo de lecturas. Se utilizó una dilución en serie de un genoma para crear un conjunto de muestras de diferentes concentraciones. Las series de dilución (DS) 1-12 representan concentraciones de 0.8 %, 1.6 %, 4 %, 7.5 %, 11 %, 14 %, 17 %, 29 %, 38 %, 45 %, 55 % y 89 %, respectivamente. Los resultados mostrados en la Fig. 18 también demuestran que el sistema y los métodos de la presente invención pueden identificar con precisión una especie microbiana con su concentración relativa a partir de una mezcla metagenómica compleja incluso si está representada por solo 1000 lecturas cortas (-72 pb), y las concentraciones relativas observadas están en buena concordancia con las concentraciones reales incluso cuando las especies diana están presentes en concentraciones muy bajas.

Las mayores fuentes de error para determinar la sensibilidad, las concentraciones relativas y los niveles más bajos de abundancia son la integridad de la (s) base de datos (s) de referencia y el grado de proximidad entre la cepa causal y su prima más cercana. Por ejemplo, si solo hay un centenar de diferencias de sitios entre dos cepas, es poco probable que se produzca una discriminación estadística entre estas cepas putativas. La siguiente fuente de error más grande en la estimación de la concentración ocurre cuando dos cepas relacionadas están presentes en diferentes niveles de concentración. Los errores de lectura del sistema de la cepa en mayor concentración pueden sangrar en el conjunto de desambiguación de la cepa menor, elevando artificialmente la concentración estimada de la cepa secundaria. En algunas realizaciones, se estima que la cantidad de ADN de muestra requerida para analizar poblaciones de microorganismos es menor o igual a aproximadamente 0.4 ng de ADN.

En la Figura 19 se muestra otro ejemplo no limitante del sistema y método de la presente invención aplicado para medir poblaciones del microbioma a nivel de especie para un paciente con enfermedad de Crohn. La Fig. 19 muestra una comparación de los resultados del Motor Comparador (“CE”) con los resultados de BLAST y demuestra que la secuenciación metagenómica directa del ADN y el emparejamiento probabilístico de acuerdo con la presente invención fue capaz de caracterizar, con mayor precisión, poblaciones de bacterias en comparación con el método BLAST conocido.

Las ventajas de un sistema y método de secuenciación directa de ADN y análisis metagenómico que utiliza emparejamiento probabilístico de acuerdo con la presente invención pueden incluir, pero no se limitan a, lo siguiente:

(i) el sistema y el método son capaces de identificar poblaciones enteras de microorganismos en una sola muestra;

(ii) el sistema y el método son universales y generales para todos los microorganismos, ya sean bacterias, virus, parásitos, hongos o fragmentos de ADN, plásmidos, elementos móviles u otros;

(iii) el sistema y el método son capaces de detectar e identificar simultáneamente todos los tipos de microorganismos presentes mediante el empleo de bases de datos genómicas de referencia;

(iv) la precisión y sensibilidad del sistema y método excede todos los métodos convencionales;

(v) el sistema y el método no requieren la amplificación del ADN, y la extracción y secuenciación directas minimizan los errores que de otro modo podrían resultar de la amplificación;

(vi) el sistema y el método son independientes de la tecnología de secuenciación y normalizan el error de máquina de los secuenciadores;

(vii) el sistema y el método dan cuenta de la mutación genética, incluidas inserciones y eliminaciones, islas de patogenicidad, inserción de plásmidos y genomas móviles u otros;

(viii) el sistema y el método son capaces de rastrear el origen de poblaciones específicas de microorganismos, incluida la detección de agentes patógenos en los alimentos y el agua;

(ix) el sistema y el método son capaces de determinar cambios en la biodiversidad y pueden detectar patógenos manipulados; y

(x) el sistema y el método pueden tener una sensibilidad estimada para identificar un solo organismo en una población de microorganismos de 1 en 20 mil millones.

Las realizaciones de la presente invención se han descrito completamente anteriormente con referencia a las figuras de los dibujos.

REIVINDICACIONES

1. Un método para caracterizar material biológico en una muestra que contiene material genético de una pluralidad de organismos, el método comprende:

(a) generar una pluralidad de lecturas de fragmentos metagenómicos secuenciando fragmentos metagenómicos extraídos de la muestra;

(b) realizar una comparación probabilística para cada una de las lecturas del fragmento metagenómico con una pluralidad de lecturas del genoma de una base de datos genómica de referencia que contiene identidades genómicas del organismo para detectar correlaciones causales entre cada una de las lecturas del fragmento metagenómico y la pluralidad de lecturas del genoma en las que existe una correlación causal entre un fragmento metagenómico leído y una subsecuencia en la base de datos genómica de referencia cuando las comparaciones probabilísticas determinan que la lectura del fragmento metagenómico y la subsecuencia de la base de datos genómica de referencia son lo suficientemente similares como para implicar una relación biológica, en donde la comparación probabilística comprende:

(i) crear una lista de palabras de secuencia de muestra para cada una de la pluralidad de lecturas de fragmentos metagenómicos que comprenden:

dividir cada una de las lecturas de fragmentos metagenómicos en palabras de secuencia de muestra en un carácter de límite, en el que el carácter de límite es uno de A, C, T o G, y

crear la lista de palabras de secuencia de muestra a partir de las palabras de secuencia de muestra para cada una de las lecturas de los fragmentos metagenómicos;

ii) crear una lista de palabras de referencia para una pluralidad de lecturas del genoma de la base de datos genómica de referencia que contiene identidades genómicas de organismos que comprenden:

dividir cada una de las lecturas del genoma en palabras de referencia en un carácter de límite en el que el carácter de límite es uno de A, C, T o G,

crear las listas de palabras de referencia a partir de las palabras de referencia para cada una de las lecturas del genoma, y

almacenar las listas de palabras de referencia en un catálogo de listas de palabras de referencia, en el que cada lista de palabras de referencia está asociada con una o más categorías, y cada una de las una o más categorías es un género, especie o cepa de un organismo conocido;

(c) para cada una de las palabras de secuencia de muestra de la lista de palabras de secuencia de muestra, comparar la palabra de secuencia de muestra con las palabras de referencia de cada una de las listas de palabras de referencia e identificar coincidencias entre la palabra de secuencia de muestra y una o más de las palabras de referencia usando emparejamiento probabilístico; y

d) retener las correlaciones causales detectadas;

e) agregar las correlaciones causales retenidas por cepa genómica y especie para identificar un conjunto de genomas de microorganismos contenidos en la muestra;

f) crear conjuntos de patrones independientes de inclusión de subconjuntos y exclusión de subconjuntos para dicho conjunto de genomas de microorganismos, en el que los conjuntos de patrones independientes son patrones de palabra de referencia que ocurren en algunos, pero no en todos los genomas del conjunto de genomas;

g) para cada conjunto de patrones independientes, emparejar el conjunto con la pluralidad de lecturas de fragmentos metagenómicos en el que un resultado de emparejamiento indica el grado de cobertura del conjunto de patrones por la pluralidad de lecturas de fragmentos metagenómicos y en el que cada uno de los emparejamientos da como resultado una estimación independiente de concentración del genoma en el conjunto;

en el que las etapas del método se realizan utilizando un procesador y una memoria.

2. El método de acuerdo con la reivindicación 1, en el que el método comprende rellenar una tabla hash con las palabras de referencia de cada una de las listas creadas de palabras de referencia.

3. El método de acuerdo con la reivindicación 1, en el que las comparaciones probabilísticas:

a) producir resultados probabilísticos en forma de un mapa de probabilidad de probabilidades de que las especies y/o cepas de microorganismos contenidas en la base de datos genómica de referencia estén presentes en la muestra. o

b) incluir mapeo de árboles filogenéticos.

4. Un aparato para caracterizar material biológico en una muestra que contiene material genético de una pluralidad de organismos, el aparato comprende un procesador y

memoria, en la que el procesador y la memoria están configurados para realizar el método de la reivindicación 1.

5. El aparato de la reivindicación 4, que comprende además un secuenciador configurado para generar la pluralidad de lecturas de fragmentos metagenómicos secuenciando los fragmentos metagenómicos extraídos de la muestra.

6. El aparato de la reivindicación 5, que comprende además un extractor configurado para extraer los fragmentos metagenómicos de la muestra.

7. El método de acuerdo con cualquiera de las reivindicaciones 1-3, en el que la muestra es una muestra clínica.

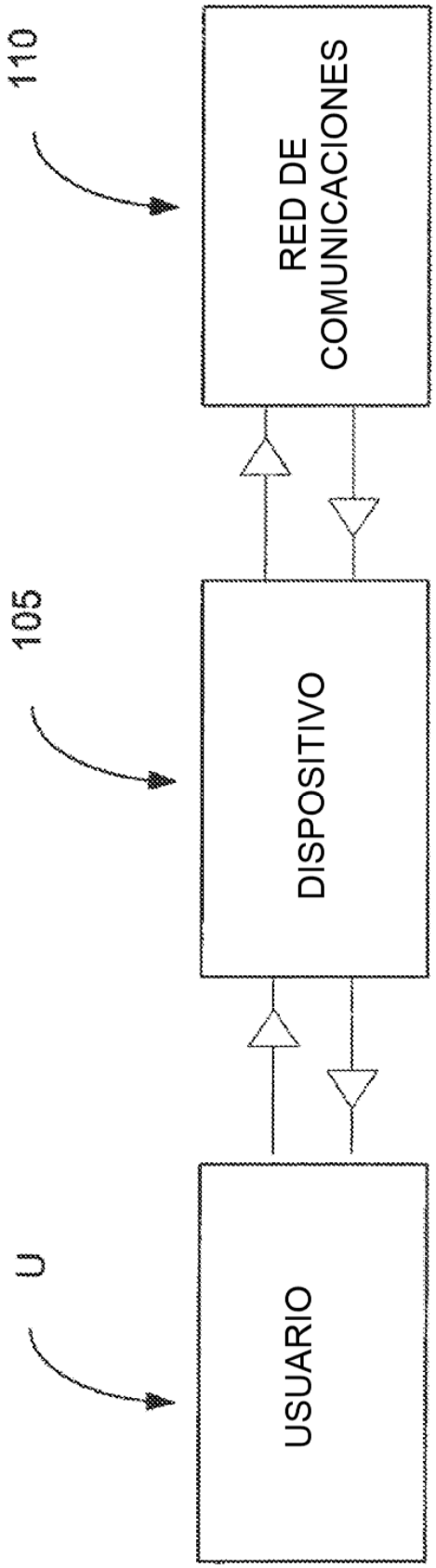

FIG. 1

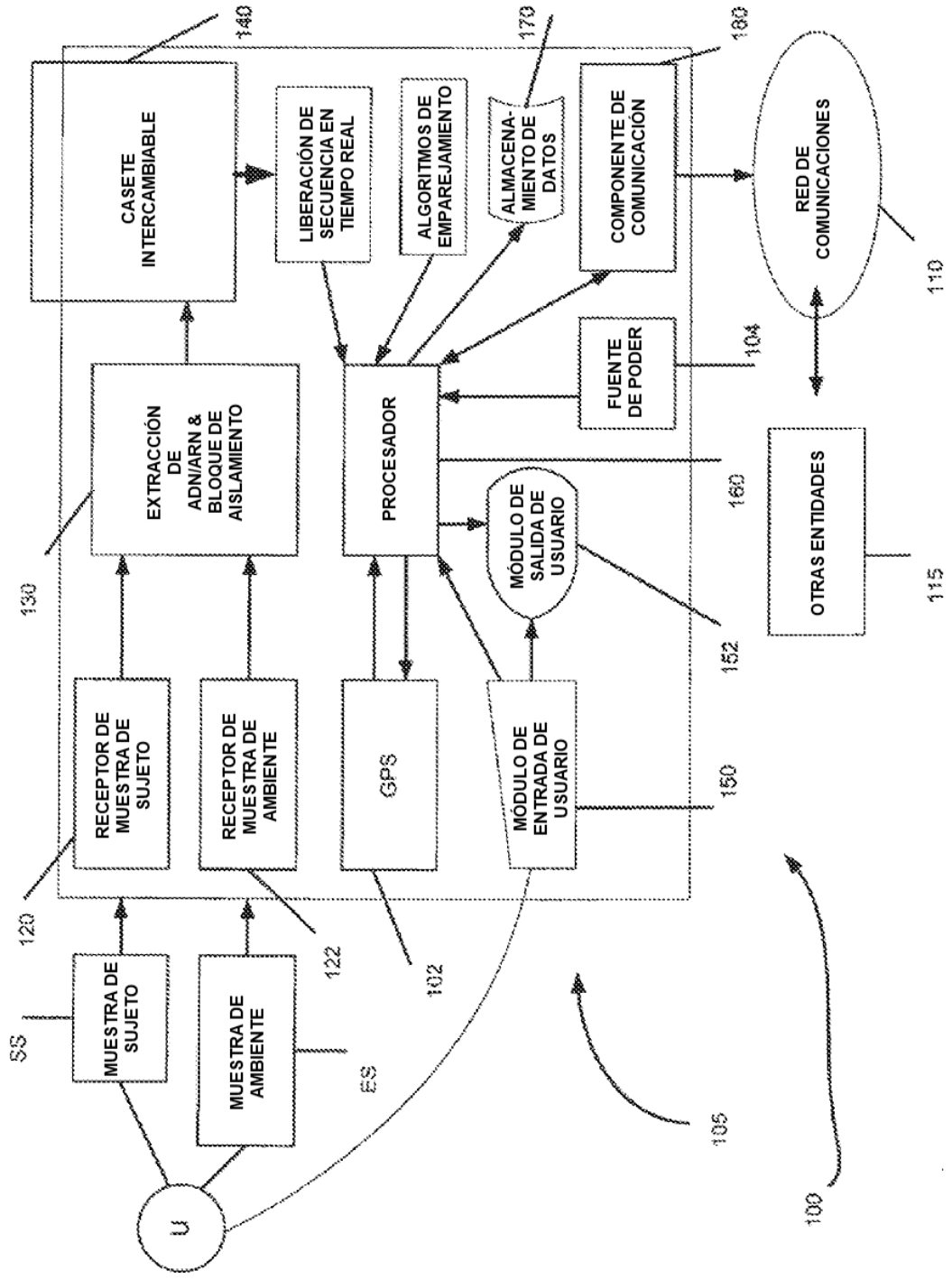
CASETE INTERCAMBIABLE
LIBERACIÓN DE SECUENCIA EN TIEMPO REAL
ALGORITMOS DE EMPAREJAMIENTO
ALMACENA-MIENTO DE DATOS
COMPONENTE DE COMUNICACIÓN
RED DE COMUNICACIONES
EXTRACCIÓN DE ADN/ARN & BLOQUE DE AISLAMIENTO
PROCESADOR
FUENTE DE PODER
MÓDULO DE SALIDA DE USUARIO
OTRAS ENTIDADES
RECEPTOR DE MUESTRA DE SUJETO
RECEPTOR DE MUESTRA DE AMBIENTE
GPS
MÓDULO DE ENTRADA DE USUARIO
MUESTRA DE SUJETO
MUESTRA DE AMBIENTE
U
SS
ES
140
170
180
130
104
110
160
115
120
152
122
102
150
105
100

FIG. 2

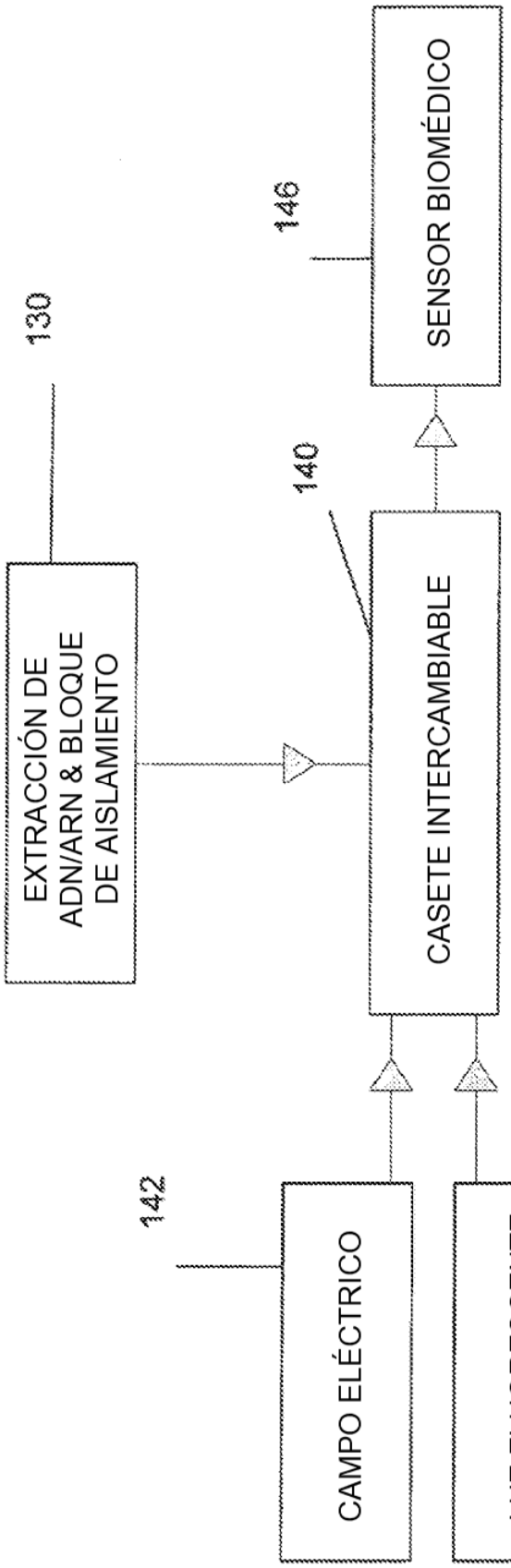
130
EXTRACCIÓN DE ADN/ARN & BLOQUE DE AISLAMIENTO
140
CASETE INTERCAMBIABLE
146
SENSOR BIOMÉDICO
142
CAMPO ELÉCTRICO
144
LUZ FLUORESCENTE

FIG. 3

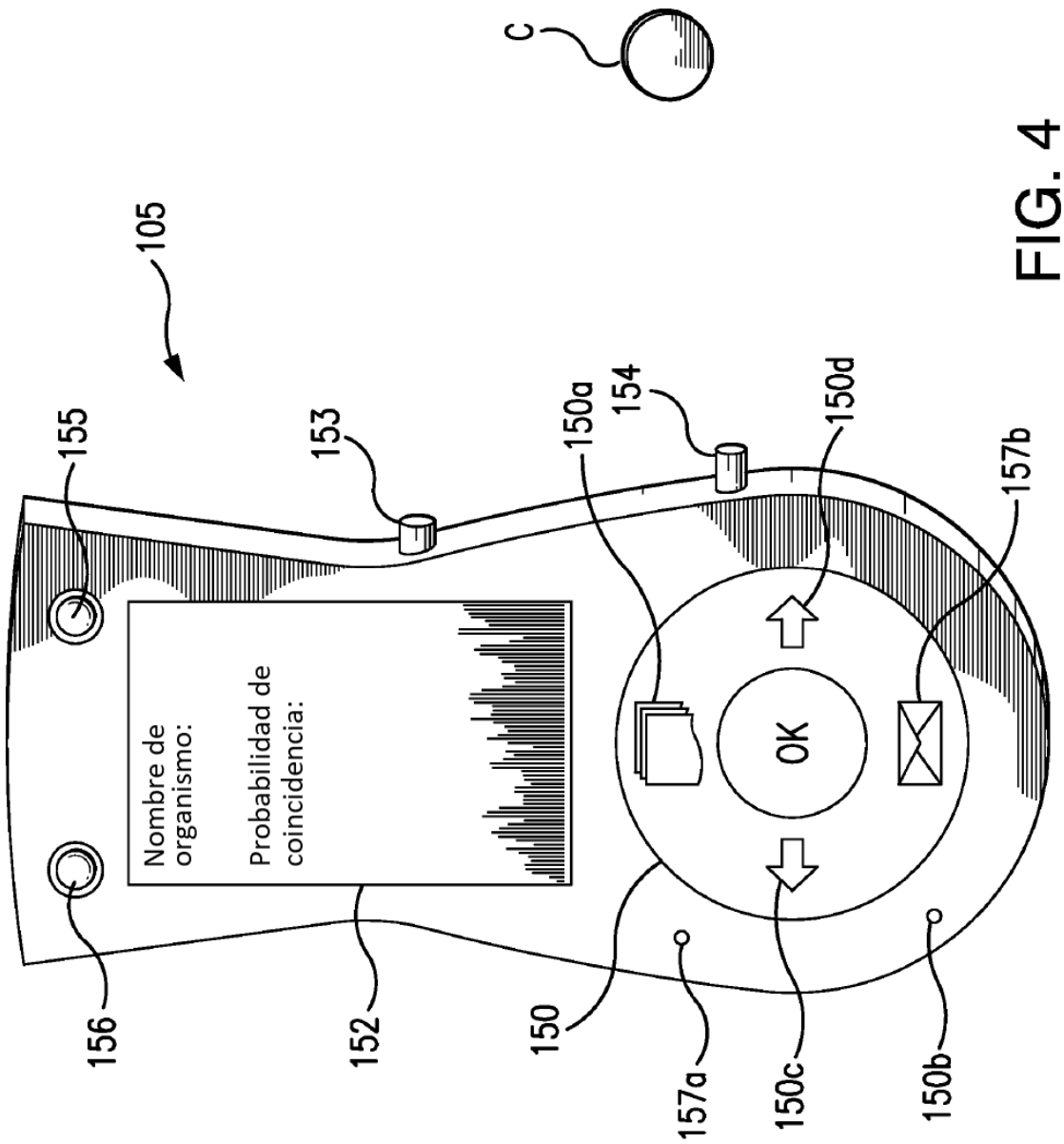
C
105
155
153
150a
154
150d
157b
Nombre de organismo:
Probabilidad de coincidencia:
OK
156
152
150
157a
150c
150b

FIG. 4

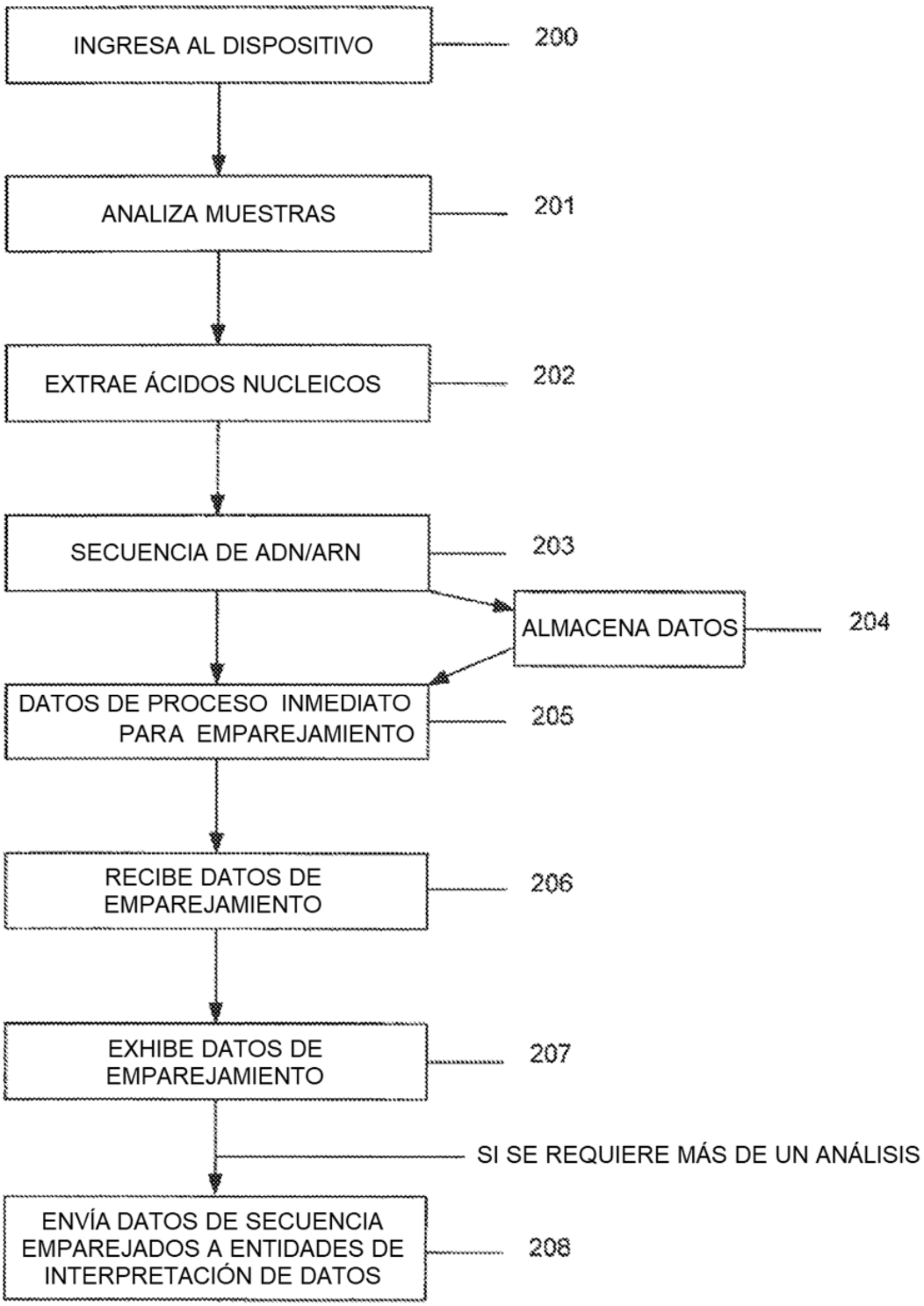
INGRESA AL DISPOSITIVO
200
ANALIZA MUESTRAS
201
EXTRAE ÁCIDOS NUCLEICOS
202
SECUENCIA DE ADN/ARN
203
ALMACENA DATOS
204
DATOS DE PROCESO INMEDIATO PARA EMPAREJAMIENTO
205
RECIBE DATOS DE EMPAREJAMIENTO
206
EXHIBE DATOS DE EMPAREJAMIENTO
207
SI SE REQUIERE MÁS DE UN ANÁLISIS
ENVÍA DATOS DE SECUENCIA EMPAREJADOS A ENTIDADES DE INTERPRETACIÓN DE DATOS
208

FIG. 5

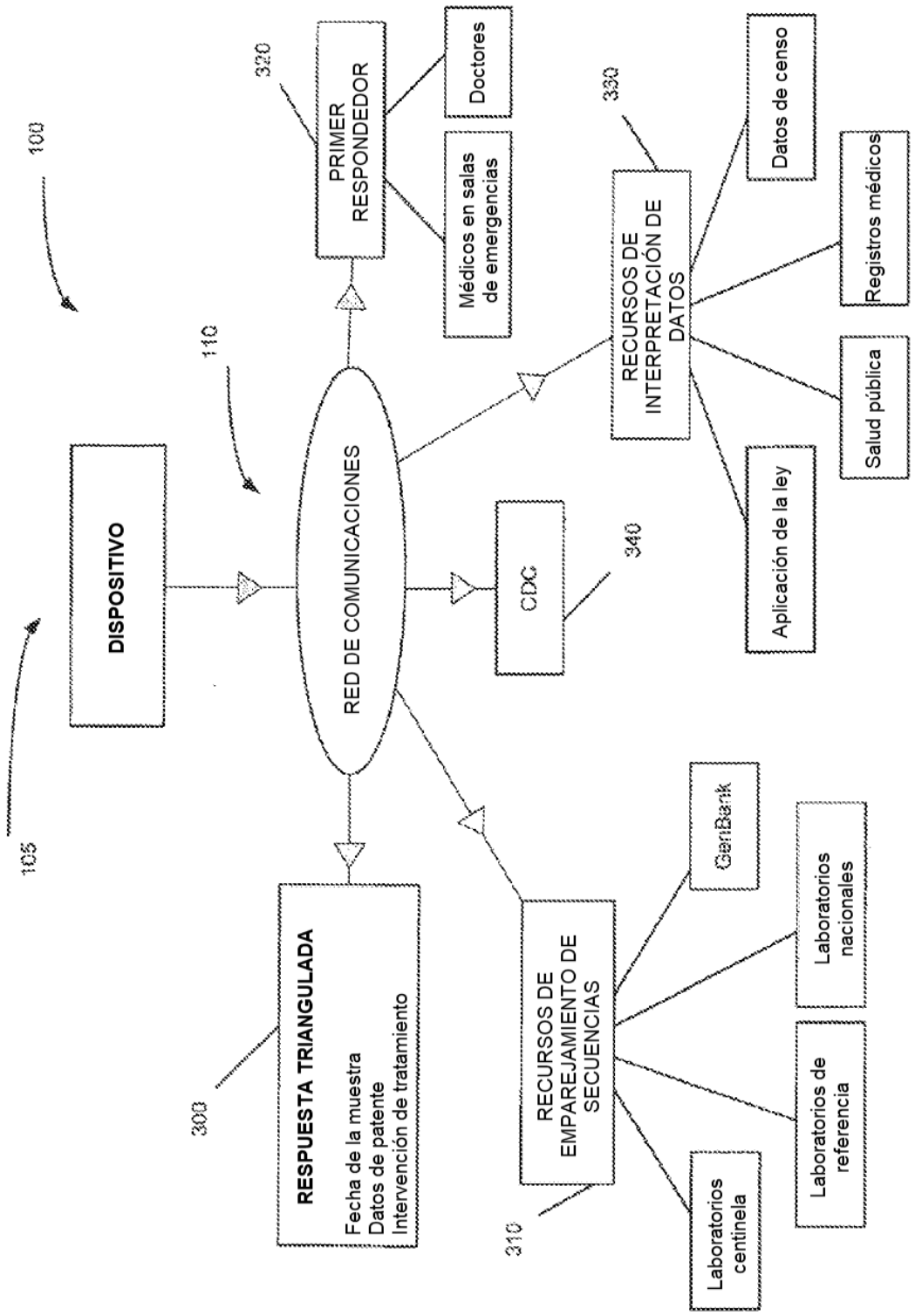
100
105
DISPOSITIVO
110
RED DE COMUNICACIONES
RESPUESTA TRIANGULADA
Fecha de la muestra
Datos de patente
Intervención de tratamiento
300
PRIMER RESPONDEDOR
320
Doctores
Médicos en salas de emergencias
CDC
340
RECURSOS DE INTERPRETACIÓN DE DATOS
330
Datos de censo
Registros médicos
Salud pública
Aplicación de la ley
RECURSOS DE EMPAREJAMIENTO DE SECUENCIAS
310
GenBank
Laboratorios nacionales
Laboratorios de referencia
Laboratorios centinela

FIG. 6

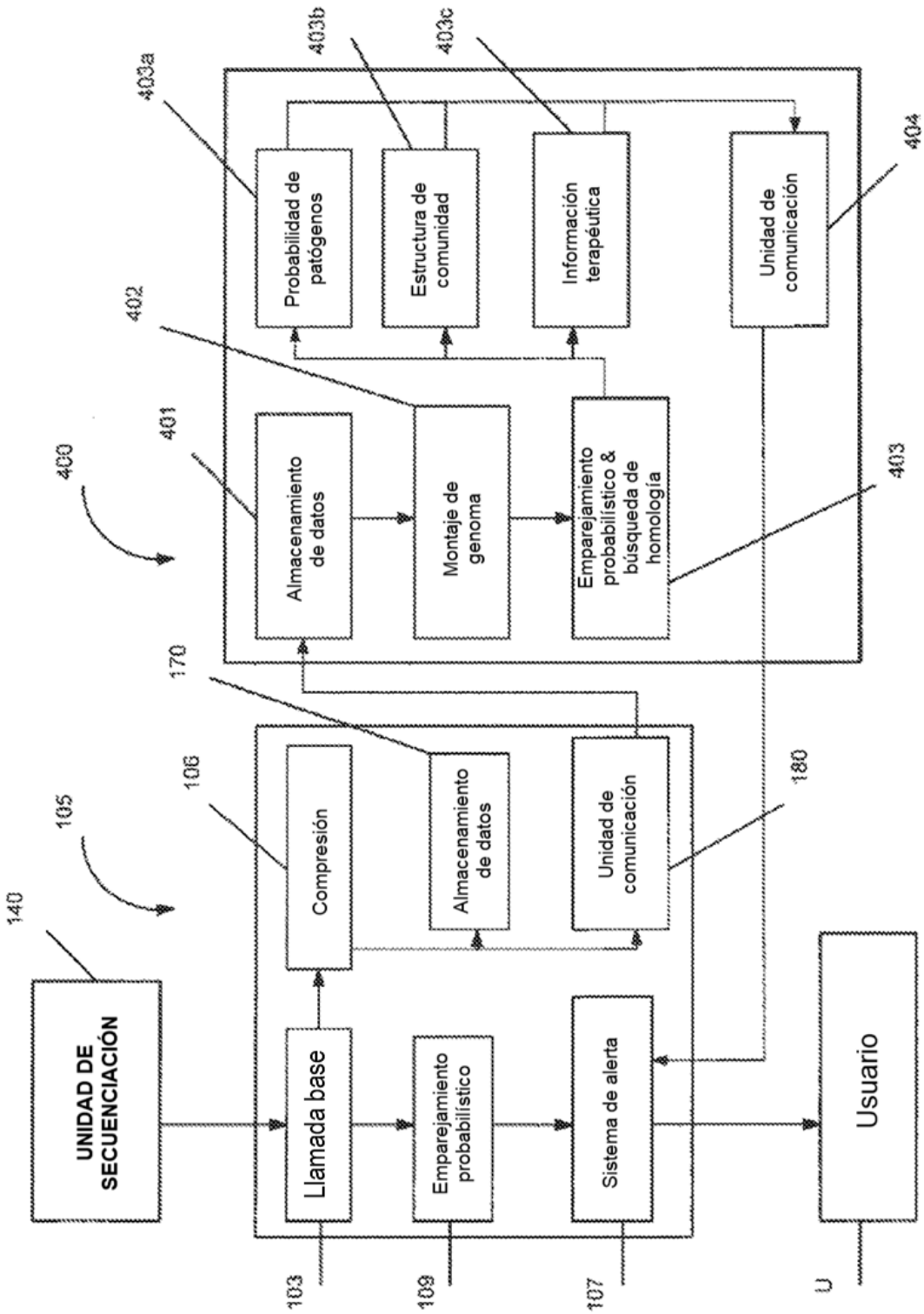
UNIDAD DE SECUENCIACIÓN
140
105
Llamada base
Compresión
106
Emparejamiento probabilístico
Almacenamiento de datos
Sistema de alerta
Unidad de comunicación
170
180
Usuario
U
103
109
107
400
401
Almacenamiento de datos
402
Montaje de genoma
Emparejamiento probabilístico & búsqueda de homología
403
403a
Probabilidad de patógenos
403b
Estructura de comunidad
403c
Información terapéutica
Unidad de comunicación
404

FIG. 7

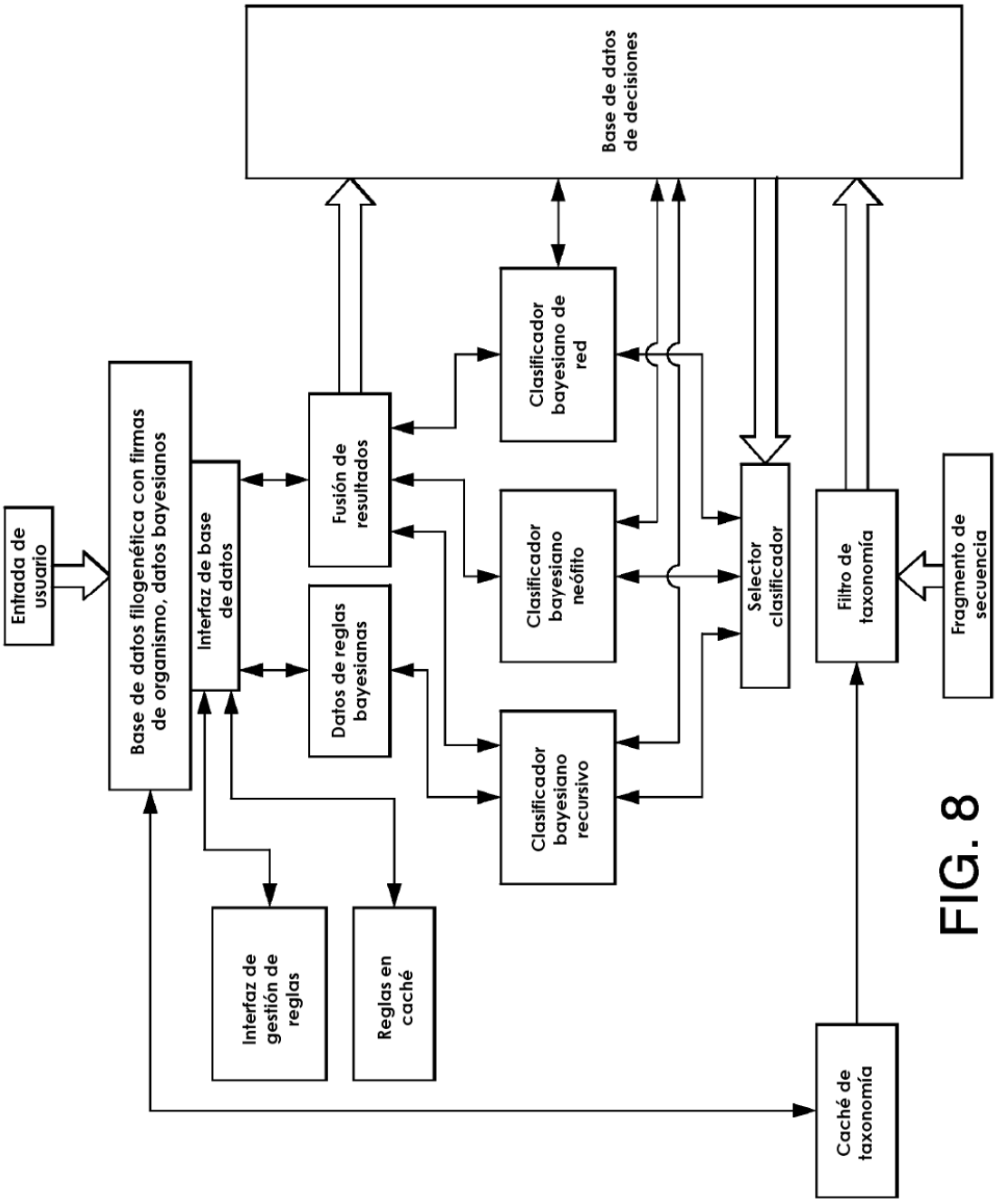
Base de datos de decisiones
Clasificador bayesiano de red
Fusión de resultados
Entrada de usuario
Base de datos filogenética con firmas de organismo, datos bayesianos
Interfaz de base de datos
Clasificador bayesiano neófito
Selector clasificador
Filtro de taxonomía
Fragmento de secuencia
Datos de reglas bayesianas
Clasificador bayesiano recursivo
Interfaz de gestión de reglas
Reglas en caché
Caché de taxonomía

FIG. 8

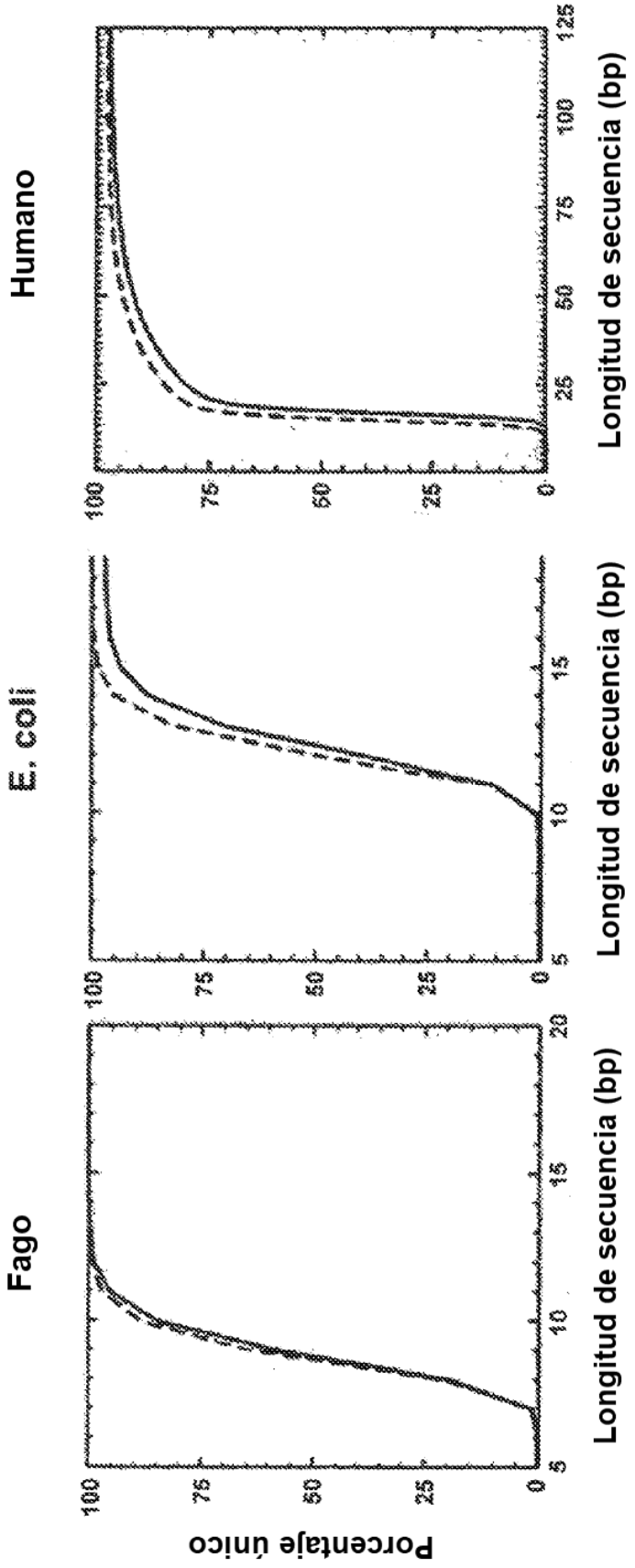
Fago
E. coli
Humano
Porcentaje único
Longitud de secuencia (bp)
Longitud de secuencia (bp)
Longitud de secuencia (bp)

FIG. 9

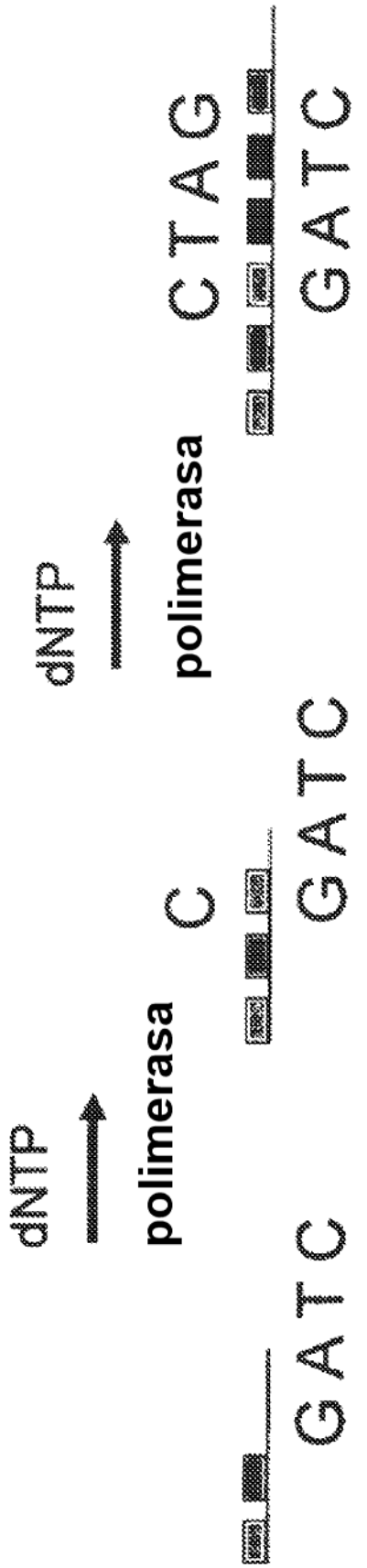

FIG. 10

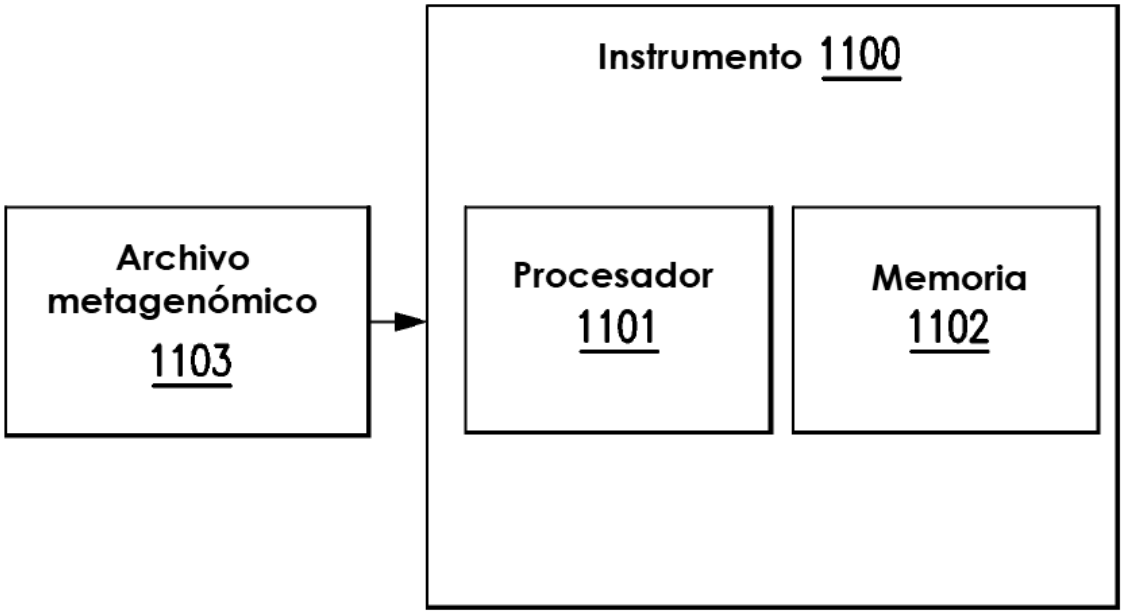
Instrumento 1100
Archivo metagenómico 1103
Procesador 1101
Memoria 1102

FIG. 11


Recolecta muestra
S1201
Extrae fragmentos metagenómicos
S1202
Secuencia de fragmentos metagenómicos
S1203
Métodos probabilísticos para caracterizar materiales biológicos en la muestra
S1204

FIG. 12

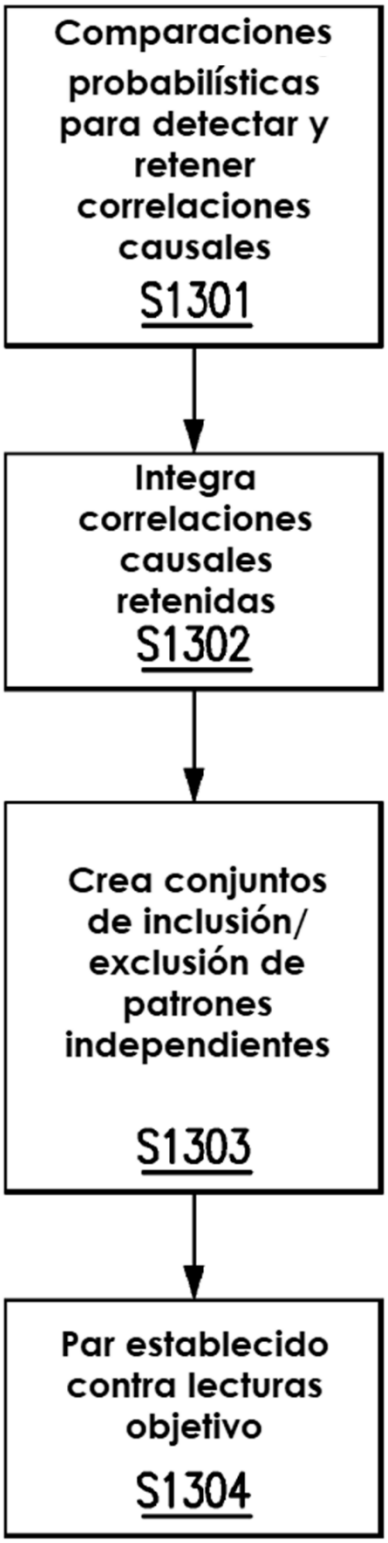
Comparaciones probabilísticas para detectar y retener correlaciones causales
S1301
Integra correlaciones causales retenidas
S1302
Crea conjuntos de inclusión/ exclusión de patrones independientes
S1303
Par establecido contra lecturas objetivo
S1304

FIG. 13

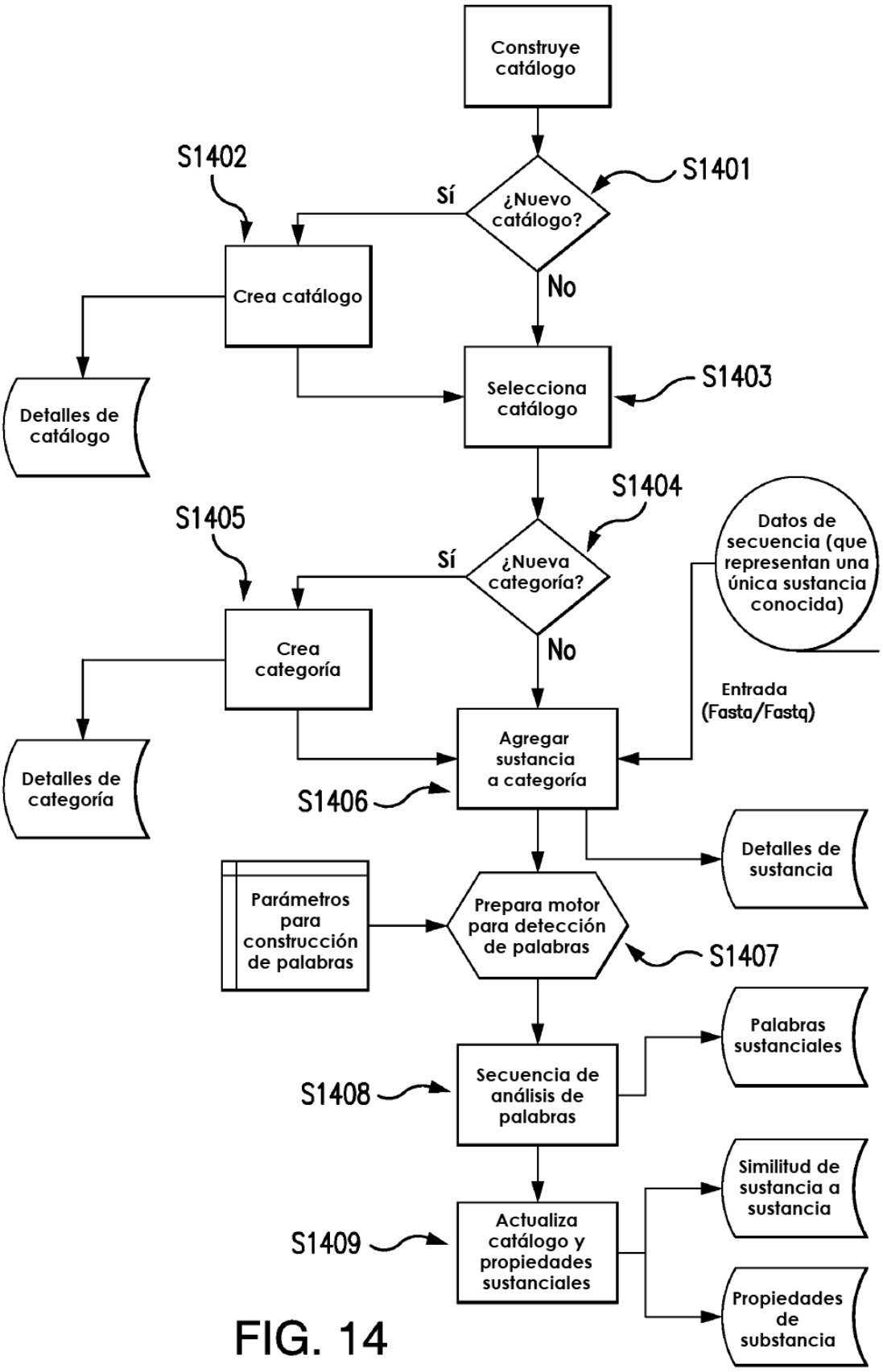
Construye catálogo
S1401
¿Nuevo catálogo?
Sí
No
S1402
Crea catálogo
Detalles de catálogo
Selecciona catálogo
S1403
S1404
¿Nueva categoría?
Sí
No
S1405
Crea categoría
Detalles de categoría
Datos de secuencia (que representan una única sustancia conocida)
Entrada (Fasta/Fastq)
Agregar sustancia a categoría
S1406
Detalles de sustancia
Parámetros para construcción de palabras
Prepara motor para detección de palabras
S1407
Palabras sustanciales
Secuencia de análisis de palabras
S1408
Similitud de sustancia a sustancia
Actualiza catálogo y propiedades sustanciales
S1409
Propiedades de substancia

FIG. 14

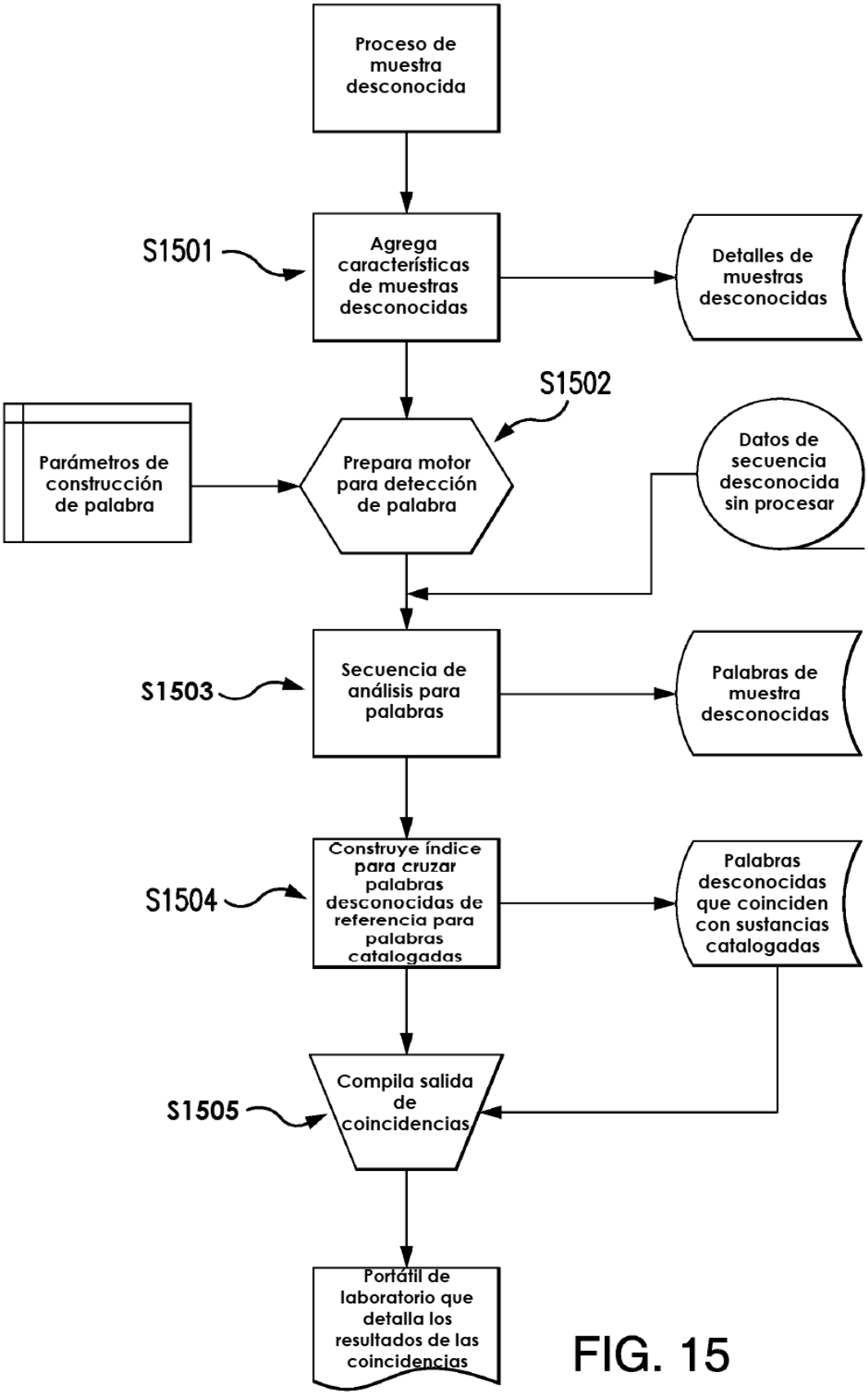
Proceso de muestra desconocida
S1501
Agrega características de muestras desconocidas
Detalles de muestras desconocidas
S1502
Parámetros de construcción de palabra
Prepara motor para detección de palabra
Datos de secuencia desconocida sin procesar
S1503
Secuencia de análisis para palabras
Palabras de muestra desconocidas
S1504
Construye índice para cruzar palabras desconocidas de referencia para palabras catalogadas
Palabras desconocidas que coinciden con sustancias catalogadas
S1505
Compila salida de coincidencias
Portátil de laboratorio que detalla los resultados de las coincidencias

FIG. 15

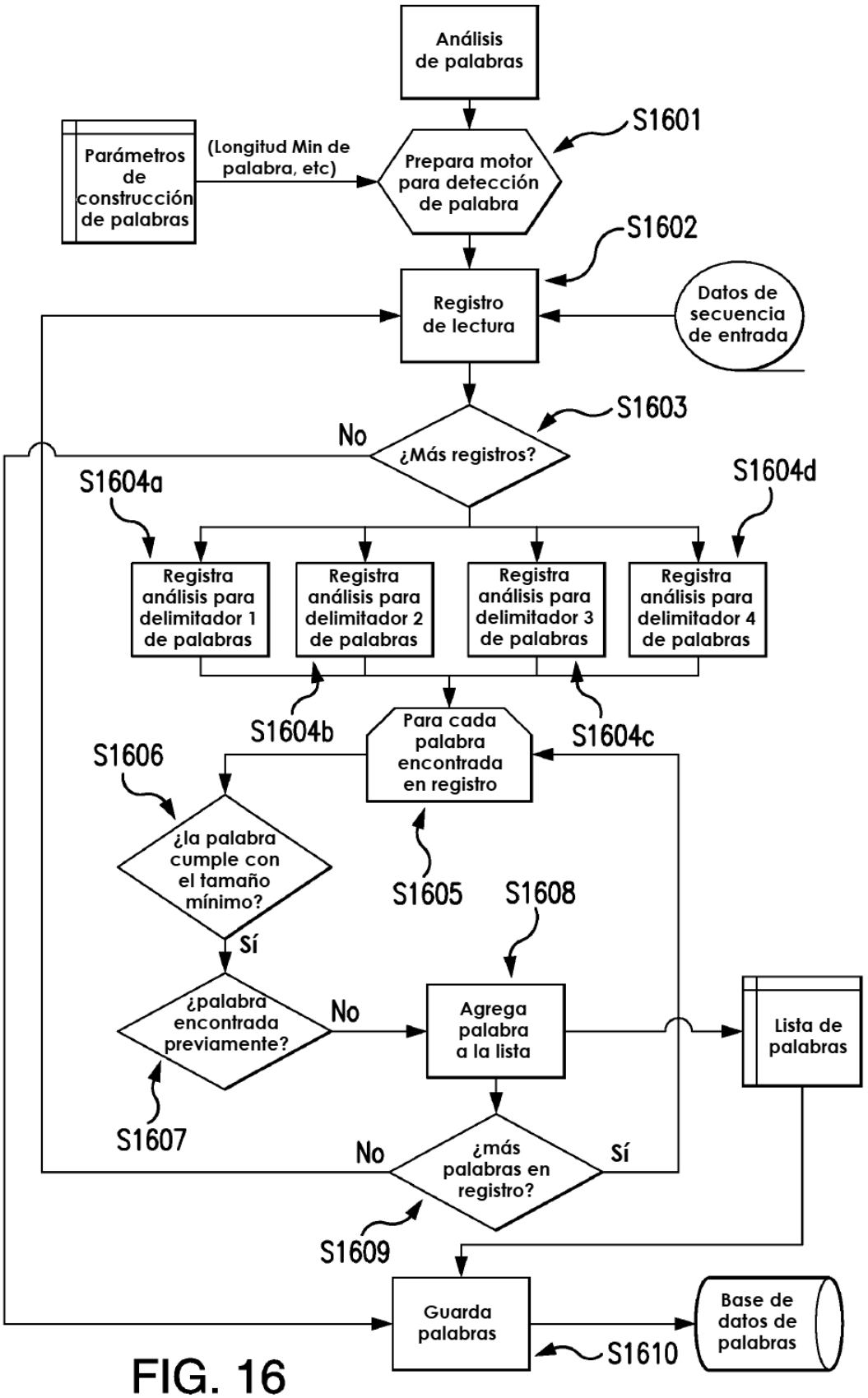
Análisis de palabras
S1601
Parámetros de construcción de palabras
(Longitud Min de palabra, etc)
Prepara motor para detección de palabra
S1602
Registro de lectura
Datos de secuencia de entrada
S1603
No
¿Más registros?
S1604a
S1604d
Registra análisis para delimitador 1 de palabras
Registra análisis para delimitador 2 de palabras
Registra análisis para delimitador 3 de palabras
Registra análisis para delimitador 4 de palabras
S1604b
S1604c
Para cada palabra encontrada en registro
S1606
¿la palabra cumple con el tamaño mínimo?
S1605
S1608
Sí
¿palabra encontrada previamente?
No
Agrega palabra a la lista
Lista de palabras
S1607
No
¿más palabras en registro?
Sí
S1609
Guarda palabras
Base de datos de palabras
S1610

FIG. 16

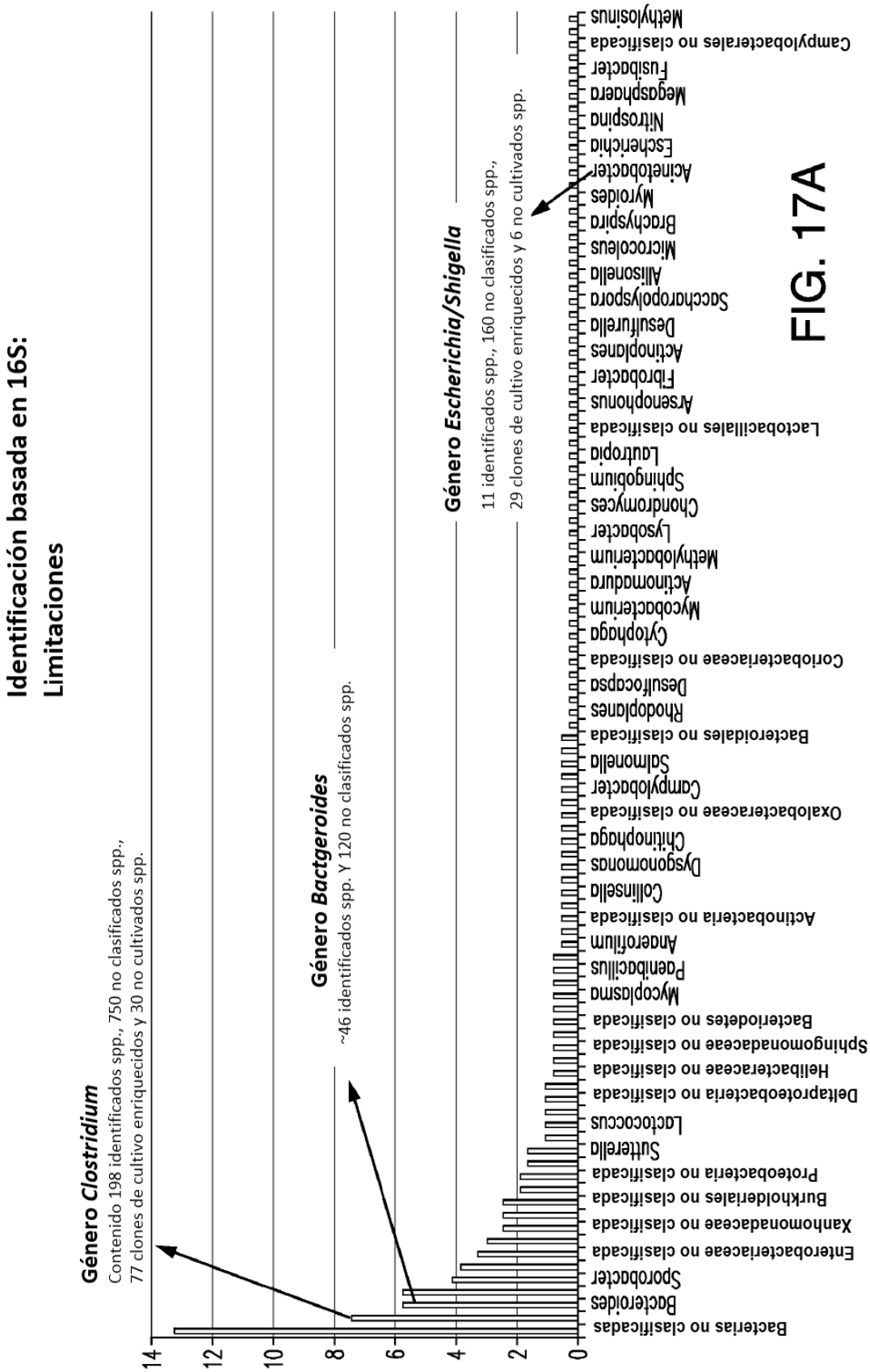
Identificación basada en 16S:
Limitaciones
Género Clostridium
Contenido 198 identificados spp., 750 no clasificados spp.,
77 clones de cultivo enriquecidos y 30 no cultivados spp.
Género Bactgeroides
~46 identificados spp. Y 120 no clasificados spp.
Género Escherichia/Shigella
11 identificados spp., 160 no clasificados spp.,
29 clones de cultivo enriquecidos y 6 no cultivados spp.
14
12
10
8
6
4
2
0
Bacterias no clasificadas
Bacteroides
Sporobacter
Enterobacteriacee no clasificada
Xanhomonadaceae no clasificada
Burkholderiales no clasificada
Proteobacteria no clasificada
Sutterella
Lactococcus
Deltaproteobacteria no clasificada
Helibacteraceae no clasificada
Sphingomonadaceae no clasificada
Bacteriodetes no clasificada
Mycoplasma
Paenibacillus
Anaerofilum
Actinobacteria no clasificada
Collinsella
Dysgonomonas
Chitinophaga
Oxalobacteraceae no clasificada
Campylobacter
Salmonella
Bacteroidales no clasificada
Rhodoplanes
Desulfocapsa
Coriobacteriaceae no clasificada
Cytophaga
Mycobacterium
Actinomadura
Methylobacterium
Lysobacter
Chondromyces
Sphingobium
Lautropia
Lactobacillales no clasificada
Arsenophonus
Fibrobacter
Actinoplanes
Desulfurella
Saccharopolyspora
Allisonella
Microcoleus
Brachyspira
Myroides
Acinetobacter
Escherichia
Nitrospina
Megasphaera
Fusibacter
Campylobacterales no clasificada
Methylosinus

FIG. 17A

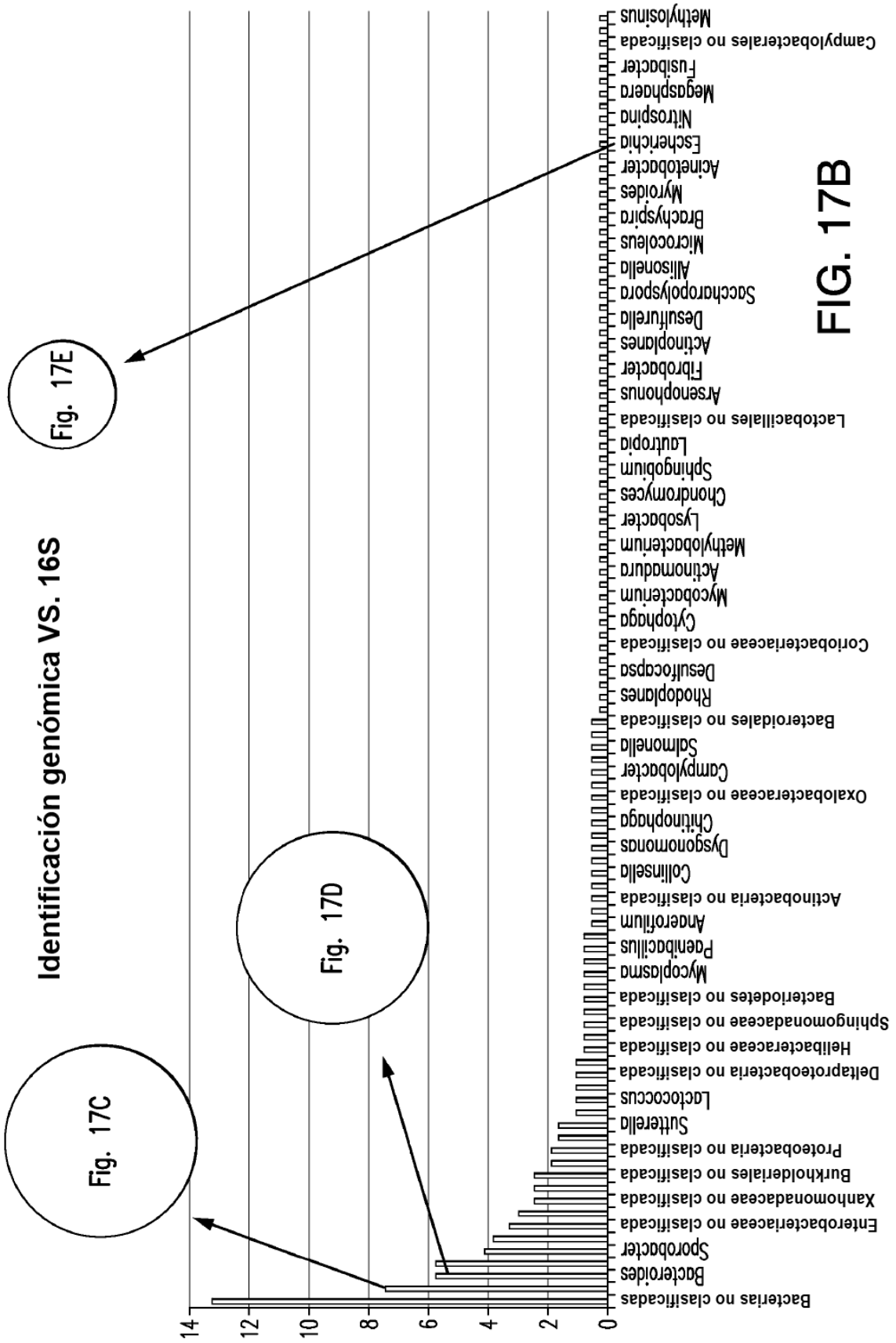
Identificación genómica VS. 16S
Fig. 17C
Fig. 17D
Fig. 17E
0
2
4
6
8
10
12
14
Bacterias no clasificadas
Bacteroides
Sporobacter
Enterobacteriaceae no clasificada
Xanthomonadaceae no clasificada
Burkholderiales no clasificada
Proteobacteria no clasificada
Sutterella
Lactococcus
Deltaproteobacteria no clasificada
Helibacteraceae no clasificada
Sphingomonadaceae no clasificada
Bacteriodetes no clasificada
Mycoplasma
Paenibacillus
Anaerofilum
Actinobacteria no clasificada
Collinsella
Dysgonomonas
Chitinophaga
Oxalobacteraceae no clasificada
Campylobacter
Salmonella
Bacteroidales no clasificada
Rhodoplanes
Desulfocapsa
Coriobacteriaceae no clasificada
Cytophaga
Mycobacterium
Actinomadura
Methylobacterium
Lysobacter
Chondromyces
Sphingobium
Lautropia
Lactobacillales no clasificada
Arsenophonus
Fibrobacter
Actinoplanes
Desulfurella
Saccharopolyspora
Allisonella
Microcoleus
Brachyspira
Myroides
Acinetobacter
Escherichia
Nitrospina
Megasphaera
Fusibacter
Campylobacterales no clasificada
Methylosinus

FIG. 17B

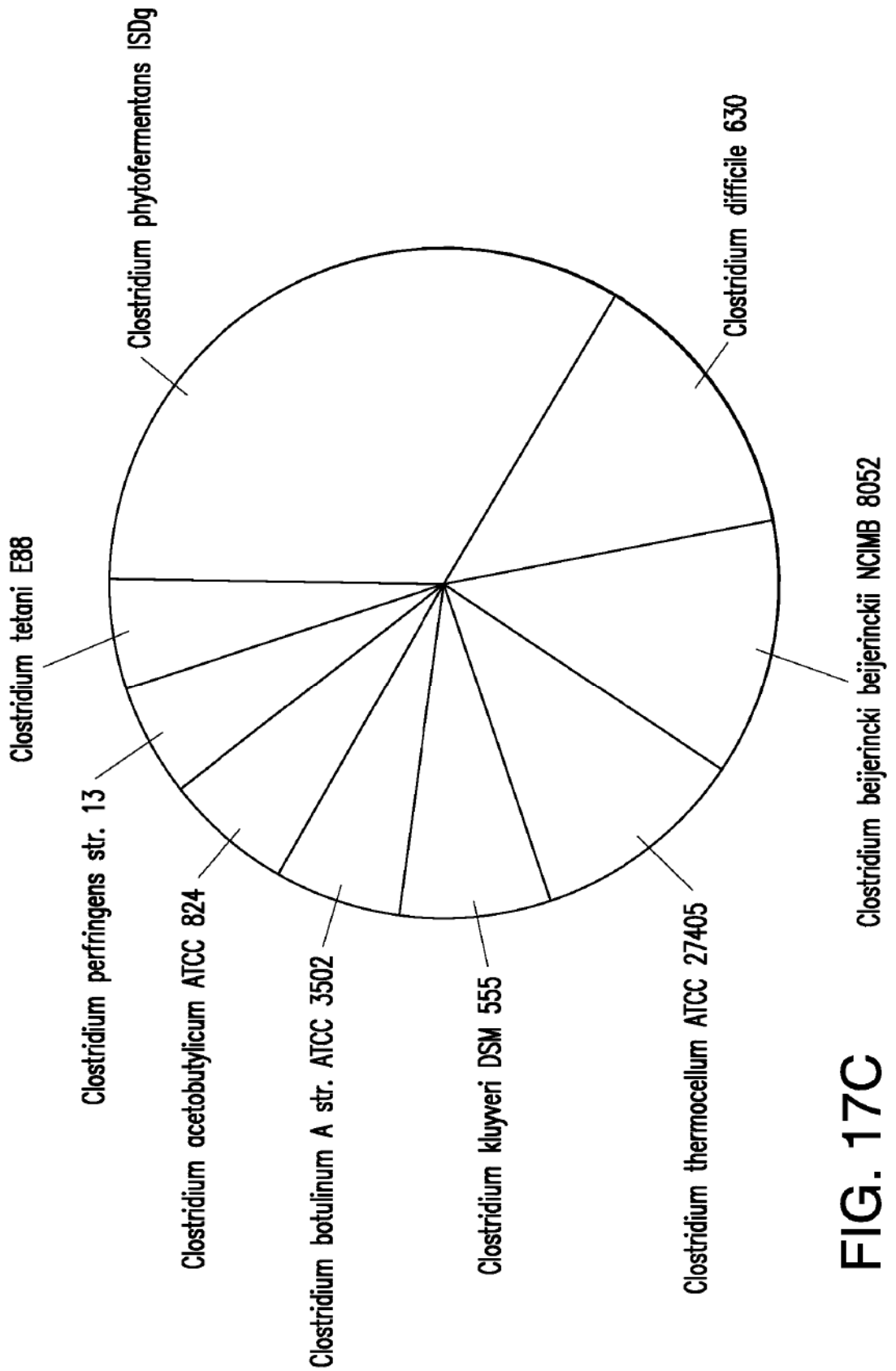
Clostridium phytofermentans ISDg
Clostridium difficile 630
Clostridium beijerinckii NCIMB 8052
Clostridium thermocellum ATCC 27405
Clostridium kluyveri DSM 555
Clostridium botulinum A str. ATCC 3502
Clostridium acetobutylicum ATCC 824
Clostridium perfringens str. 13
Clostridium tetani E88

FIG. 17C

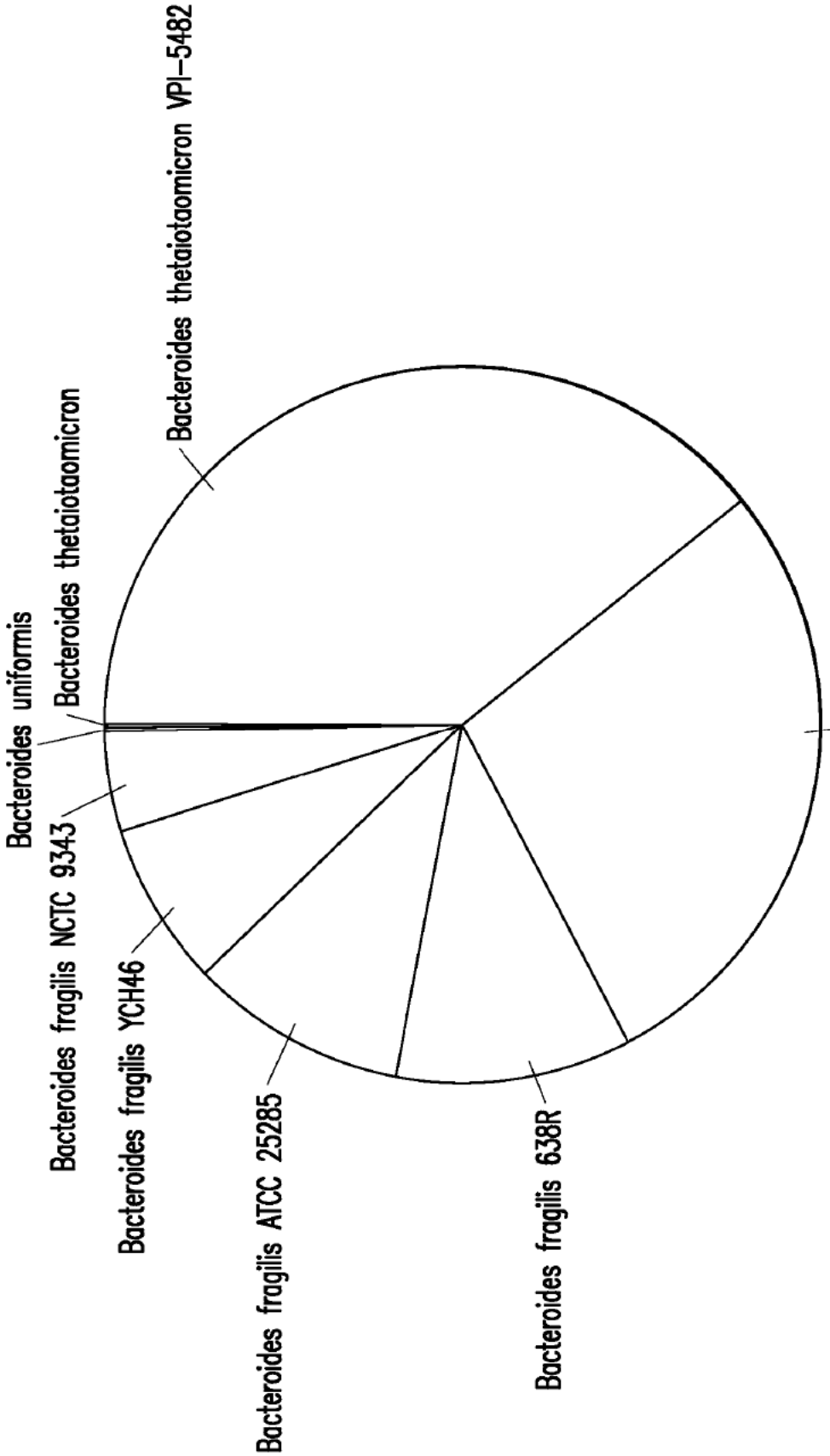
Bacteroides thetaiotaomicron VPI-5482
Bacteroides uniformis
Bacteroides thetaiotaomicron
Bacteroides fragilis NCTC 9343
Bacteroides fragilis YCH46
Bacteroides fragilis ATCC 25285
Bacteroides fragilis 638R
Bacteroides vulgatus ATCC 8482

FIG. 17D

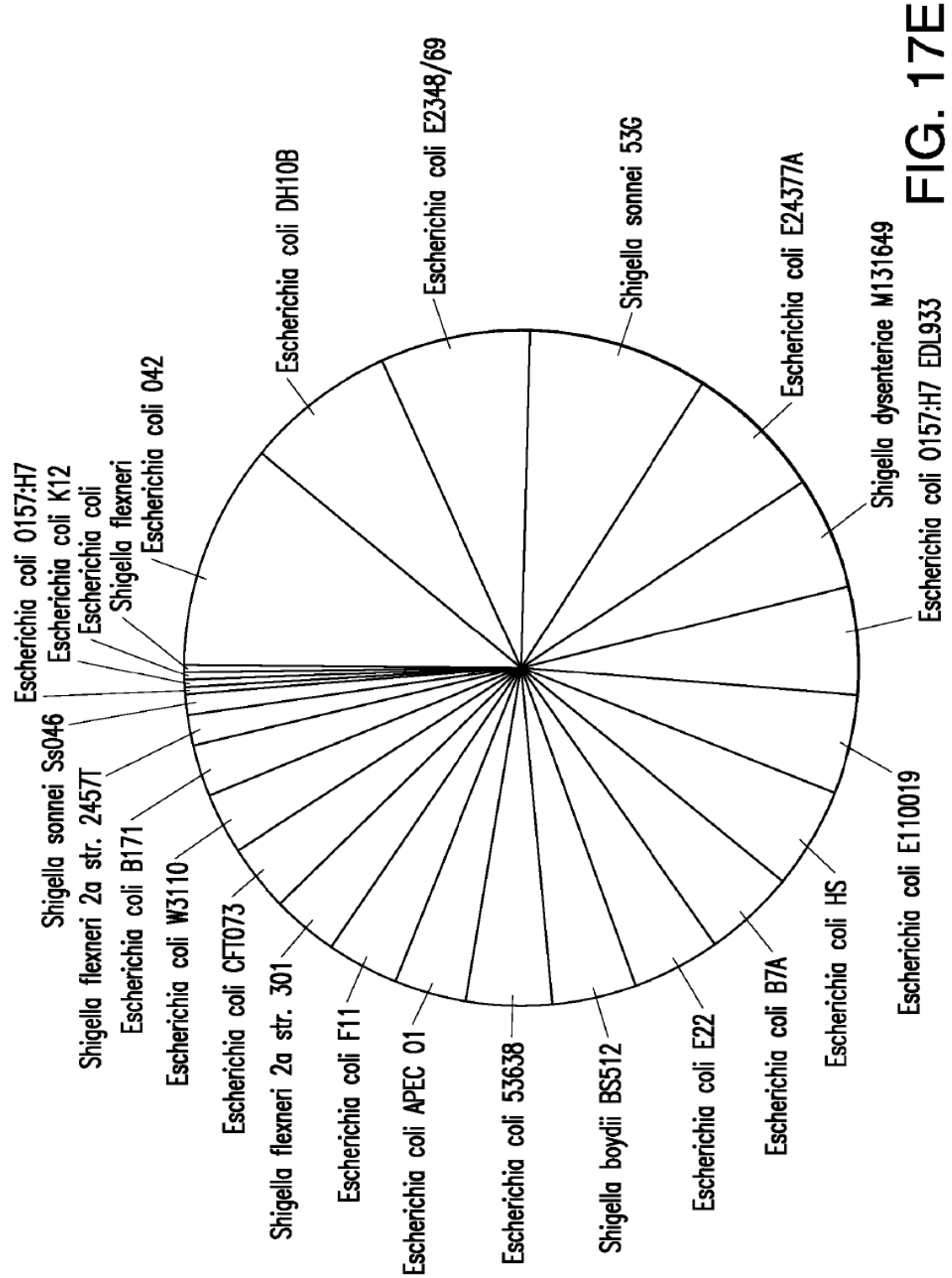
Escherichia coli 0157:H7
Escherichia coli K12
Escherichia coli
Shigella flexneri
Escherichia coli 042
Escherichia coli DH10B
Escherichia coli E2348/69
Shigella sonnei 53G
Escherichia coli E24377A
Shigella dysenteriae M131649
Escherichia coli 0157:H7 EDL933
Escherichia coli E110019
Escherichia coli HS
Escherichia coli B7A
Escherichia coli E22
Shigella boydii BS512
Escherichia coli 53638
Escherichia coli APEC 01
Escherichia coli F11
Shigella flexneri 2a str. 301
Escherichia coli CFT073
Escherichia coli W3110
Escherichia coli B171
Shigella flexneri 2a str. 2457T
Shigella sonnei Ss046

FIG. 17E

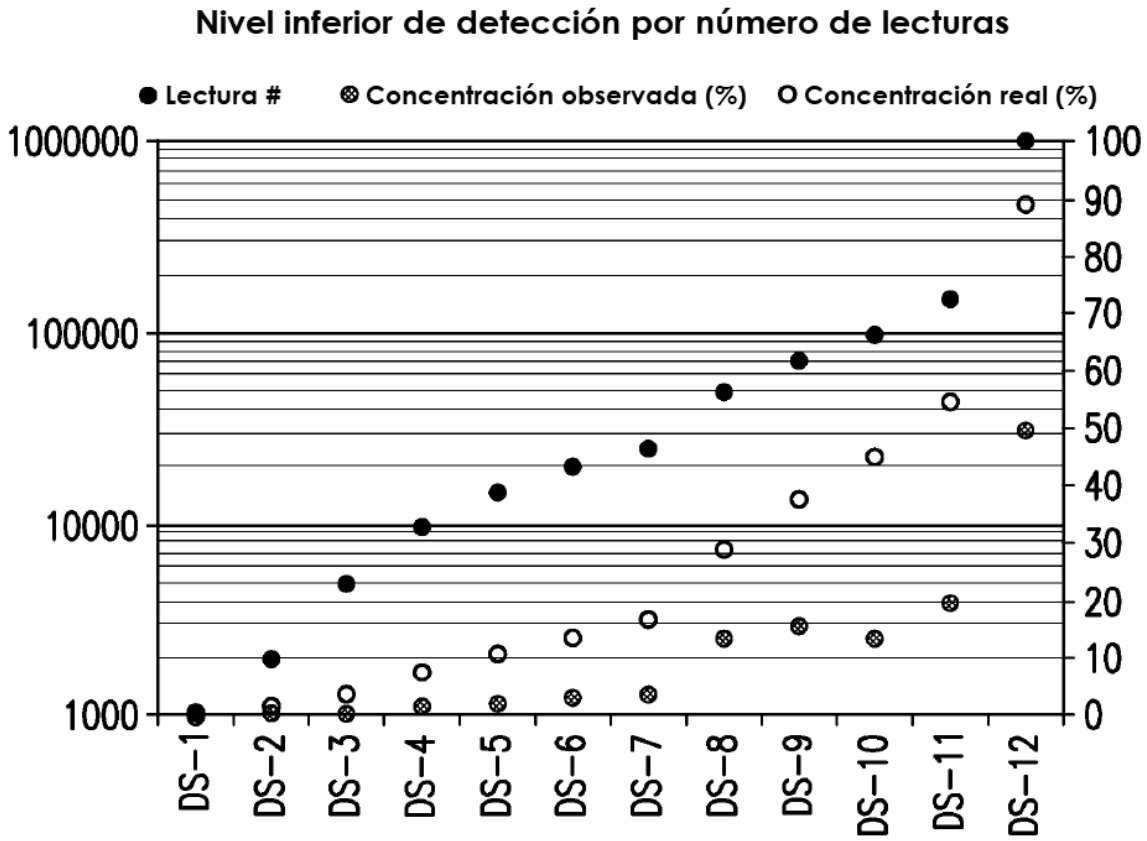
Nivel inferior de detección por número de lecturas
Lectura #
Concentración observada (%)
Concentración real (%)
1000000
100000
10000
1000
100
90
80
70
60
50
40
30
20
10
0
DS–1
DS–2
DS–3
DS–4
DS–5
DS–6
DS–7
DS–8
DS–9
DS–10
DS–11
DS–12

FIG. 18

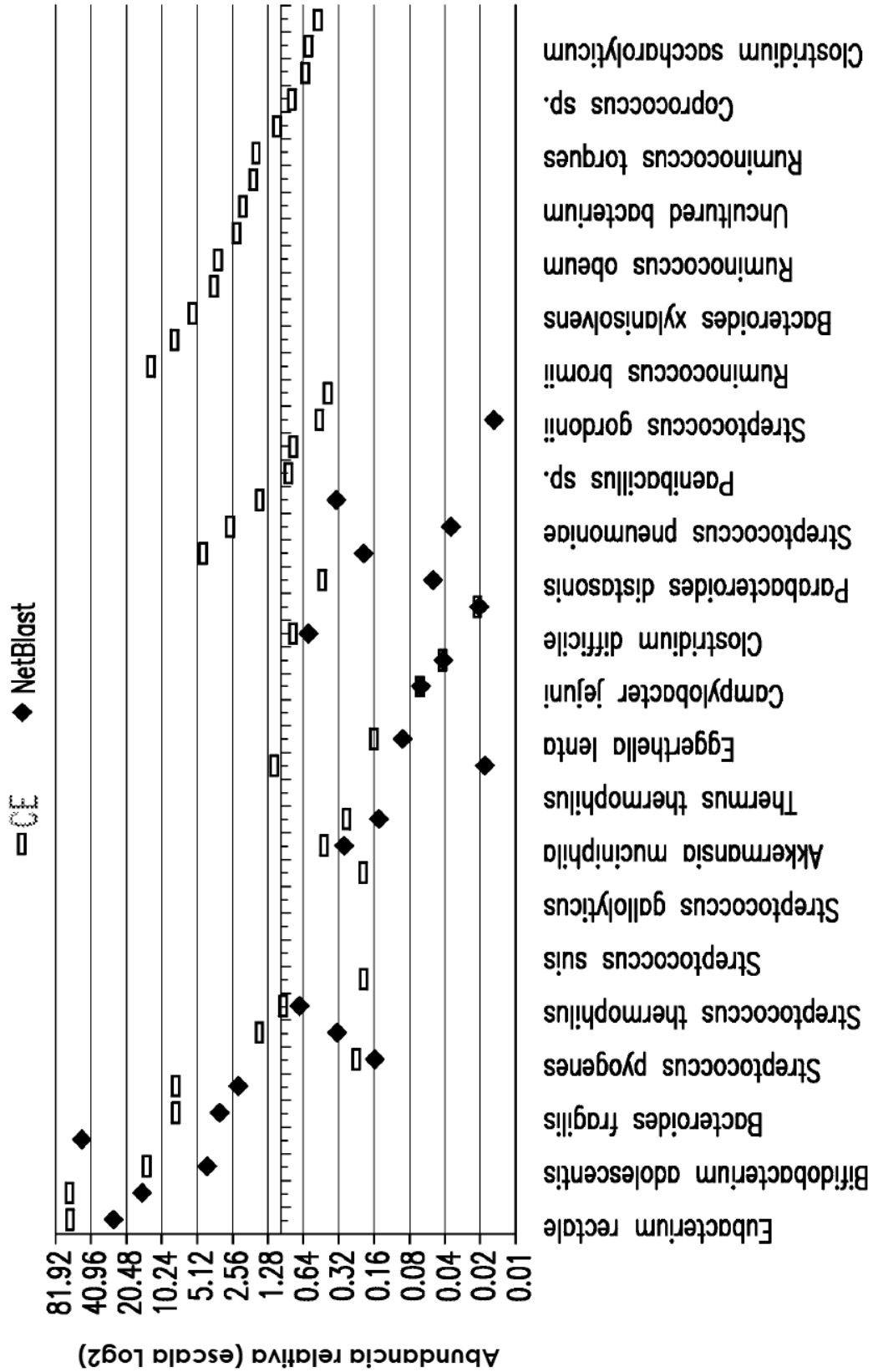
Abundancia relativa (escala log2)
81.92
40.96
20.48
10.24
5.12
2.56
1.28
0.64
0.32
0.16
0.08
0.04
0.02
0.01
NetBlast
Eubacterium rectale
Bifidobacterium adolescentis
Bacteroides fragilis
Streptococcus pyogenes
Streptococcus thermophilus
Streptococcus suis
Streptococcus gallolyticus
Akkermansia muciniphila
Thermus thermophilus
Eggerthella lenta
Campylobacter jejuni
Clostridium difficile
Parabacteroides distasonis
Streptococcus pneumoniae
Paenibacillus sp.
Streptococcus gordonii
Ruminococcus bromii
Bacteroides xylanisolvens
Ruminococcus obeum
Uncultured bacterium
Ruminococcus torques
Coprococcus sp.
Clostridium saccharolyticum

FIG. 19