Rev Argent Microbiol. 2020;52(2):150–161

asociación argentina de microbiología

REVISTA ARGENTINA DE MICROBIOLOGÍA

www.elsevier.es/ram

ARTÍCULO ESPECIAL

Aplicación de la secuenciación masiva y la bioinformática al diagnóstico microbiológico clínico

Marta Hernández[a,b,*], **Narciso M. Quijada**[a,b], **David Rodríguez-Lázaro**[b] y **José María Eiros**[c]

[a] *Laboratorio de Biología Molecular y Microbiología, Instituto Tecnológico Agrario de Castilla y León, Valladolid, España*
[b] *Área de Microbiología, Universidad de Burgos, Burgos, España*
[c] *Servicio de Microbiología, Hospital Universitario del Río Hortega, Valladolid, España*

Recibido el 21 de febrero de 2019; aceptado el 13 de junio de 2019
Disponible en Internet el 26 de noviembre de 2019

PALABRAS CLAVE
ADN;
NGS;
Diagnóstico;
Bioinformática;
WGS;
Microbiota

Resumen La aparición de secuenciadores masivos que permiten leer en paralelo de millones a miles de millones de secuencias o fragmentos del ADN *(reads)* ha revolucionado la microbiología, la cual ha pasado de un ámbito exclusivamente laboratorial a uno computacional, con la aplicación ineludible de la bioinformática. La posibilidad de efectuar estudios de la microbiota, el microbioma y el metagenoma de una muestra clínica de manera rápida y a un coste reducido permite avanzar más rápidamente en el diagnóstico de enfermedades, en el conocimiento de la taxonomía y la epidemiología de los agentes involucrados, así como de su virulencia. También posibilita la realización de estudios de genómica comparada y el descubrimiento de genes o variantes de interés, lo que puede llevar a que enfermedades tradicionalmente consideradas como de carácter no microbiano sean asociadas a la presencia de microrganismos. En esta revisión se aclara la terminología usada en este campo, y se describen las principales tecnologías de secuenciación y su utilidad en el análisis microbiano. Asimismo, se señalan diversos programas de código libre, *pipelines* de análisis, bases de datos y plataformas web que permiten que la bioinformática se integre exitosamente al ámbito de la microbiología clínica y al estudio de las enfermedades infecciosas.

KEYWORDS
DNA;
NGS;
Diagnosis;
Bioinformatics;

Bioinformatics of next generation sequencing in clinical microbiology diagnosis

Abstract Massive parallel sequencing (High-Throughput Sequencing [HTS]) allows to read millions or billions of DNA sequences or fragments (reads) in parallel and is revolutionizing microbiology research, moving from laboratory methods to computed-assisted analyses, with the compelling use of Bioinformatics. The time and cost reduction in studies on the microbiota,

* Autor para correspondencia.
Correo electrónico: hernandez.marta@gmail.com (M. Hernández).

https://doi.org/10.1016/j.ram.2019.06.003

WGS;
Microbiota

microbiome and metagenome, allows to rapidly progress in diagnosis, taxonomy, epidemiology, comparative genomics, virulence, discovery of genes or variants of interest and the association of microorganisms with traditionally considered non-microbial diseases. In this review, the terminology, the sequencing technologies and their applications are described for microbial analysis using open-source bioinformatics software, analysis pipelines, databases and web platforms that allow a user-friendly bioinformatics approach affordable by the clinical microbiologist and infectious disease practitioners.

Desarrollo de la secuenciación masiva

En 1975, Sanger y Coulson[43] publicaron el primer método enzimático para secuenciar el ADN a través de la incorporación de dideoxinucleótidos terminales, y poco después secuenciaron por primera vez un genoma, el del bacteriófago MS2 o Phi-X174[42]. Una década más tarde aparecieron los secuenciadores automáticos, que primero empleaban geles (Applied Biosystems PRISM® 373) y después capilares recubiertos de polímero (ABI PRISM® 310), y en 1995 se completó la secuencia del primer genoma bacteriano, el de *Haemophilus influenzae*[18]. En la actualidad existen casi 120.000 genomas de especies bacterianas distintas depositados en diversas bases de datos, como PATRIC (www.patricbrc.org/view/DataType/Genomes, consultado a 10 de diciembre de 2018).

A comienzos del siglo XXI surgen nuevos métodos de secuenciación, basados en la pirosecuenciación y las denominadas plataformas de *Next Generation Sequencing* (NGS) que popularizó la compañía Life Sciences-Roche con su equipo 454 GS, comercializado por primera vez en 2004[36]. Hoy en día, es más apropiado hablar de *High-Throughput Sequencing* (HTS) o secuenciación masiva, porque han surgido nuevas generaciones de secuenciadores que aplican otras tecnologías de secuenciación en paralelo. Entre estas se encuentra la secuenciación por ligación *(Sequencing by Oligonucleotide Ligation and Detection)* del equipo SOLiD —introducido en el mercado en 2007 por Life Technologies y ya descatalogado—, la secuenciación por síntesis y semiconducción del Ion Torrent —tecnología que fue adquirida por Life Technologies—, la secuenciación por síntesis en *clusters* de la empresa Solexa —luego, Illumina—, que apareció en su primer equipo GAII, de 2006, y que es utilizada en los demás secuenciadores que fueron comercializados posteriormente, como el MiSeq, a partir de 2011 (este es el secuenciador más adecuado para microbiología, puesto que es el único secuenciador de mesa de laboratorio con una longitud de hasta 600 pb, con *paired end sequencing* y exactitud de lectura de más del 99,9%). Otros secuenciadores basados en esa tecnología son HiSeq, NextSeq, NovaSeq y MiniSeq. Una tercera generación de secuenciadores son los que usan la secuenciación de molécula única (*single molecule real-time* [SMRT]), como el equipo portátil MinION™ de Oxford Nanopore Technologies (2014) o el equipo PacBio® de Pacific Biosciences (2010), que permiten secuenciar moléculas mucho más largas, de hasta 30 kb. La tabla 1 recoge las características de los principales secuenciadores masivos[34,47].

La aplicación de la secuenciación masiva, además de reducir los costes y el tiempo de análisis, genera una gran cantidad de información, que está cambiando el modo de cómo se realiza la investigación en microbiología, y demanda, inexorablemente, la aplicación de la bioinformática en el diagnóstico y el análisis microbiológico. De hecho, en genómica no se cumple la ley de Moore, ya que se duplica la capacidad de secuenciar cada 6 a 9 meses (y no cada 2 años, como ocurre con los microprocesadores). La bioinformática aplica las matemáticas, la estadística y la computación al estudio biológico, por tanto, requiere de unos conocimientos básicos en lenguajes de programación (Bash, Perl, Python y R, entre otros) y, preferentemente, también en el uso de sistemas operativos basados en UNIX®, como Linux y MacOS®. Al ser de licencia y código libres, Linux se ha convertido en la opción de preferencia para los investigadores. Las distintas distribuciones de Linux (Ubuntu, CentOS, Mint, etc.) ofrecen interfaces sencillas para el usuario, que recuerdan a aquellas de los ordenadores de oficina. Sin embargo, el verdadero poder de estas plataformas reside en el uso de la terminal, que requiere, a su vez, conocimiento del lenguaje Bash como intérprete de comandos, de tal forma que se concentre todo el potencial del computador en la tarea que se quiera realizar.

La variedad y cantidad de programas informáticos que proporcionan una interfaz gráfica son limitados y le restan poder computacional al desarrollarla, por lo que se vuelve necesario el uso de programas ejecutados desde la terminal. Aunque se trata de una disciplina nueva para muchos microbiólogos, la comunidad científica ha desarrollado numerosos programas informáticos en código abierto, que se ejecutan desde la terminal y se encuentran disponibles en plataformas como GitHub (https://github.com/) o SourceForge (https://sourceforge.net/). Incluso hay plataformas web como RAST, MG-RAST, EnteroBase o Galaxy (https://usegalaxy.org/), que integran una serie de programas y automatizan el proceso para los usuarios, así como la plataforma PLACNETw para la búsqueda de plásmidos[48]. Además, los científicos desarrollan *pipelines* de análisis, que resultan útiles porque agrupan herramientas de *software*; por ejemplo, en nuestro grupo hemos desarrollado TORMES

Tabla 1 Tipos y características de los principales secuenciadores masivos

Secuenciador	Compañía	Año[a]	Química	Máxima longitud de lectura	Máx. rendimiento/carrera	Ratio de error (%)	Máx. n.º de secuencias/carrera
454 GS FLX	Roche	2004	Pirosecuenciación-emulsión	700-800 pb	0,7 Gb	1	5×10^5
Genome Analyzer GA	Solexa	2006	Síntesis-terminadores-*clusters*	2×100 pb	1 Gb	~0,1	1×10^8
SOLiD®	Life Technologies	2007	Ligación-emulsión	75 + 35 pb	150 Gb	~5	1.400×10^6
Ion Torrent PGM™	Life Technologies	2010	Síntesis-emulsión-detección protones	200-400 pb	2 Gb	~1	$0,4\text{-}5,5 \times 10^6$
Proton™	Life Technologies	2012	Síntesis-emulsión-detección protones	200 pb	10 Gb (actualmente ~100 Gb)	~1	$60\text{-}80 \times 10^6$
PacBio®	Pacific Biosciences	2010	SMRT	6-8 kb (máx. 15 kb)	3 Gb/día	~13	370.000×10^6
HiSeq	Illumina	2010	Síntesis-terminadores-*clusters*	2×150 pb (250 pb en modo *rapid run*)	1.500 Gb	~0,1	5.000×10^6
MiSeq	Illumina	2011	Síntesis-terminadores-*clusters*	2×300 pb	15 Gb	~0,1	25×10^6
Nextseq	Illumina	2014	Síntesis-terminadores-*clusters*	2×150 pb	120 Gb	~0,1	260×10^6
Novaseq	Illumina	2017	Síntesis-terminadores-*clusters*	2×250 pb	4.800-6.000 Gb	~0,1	20.000×10^6
MinION™	Oxford Nanopore	2014	SMRT	300 kb	42 Gb	≥ 4	$4,4 \times 10^6$

SMRT: single molecule real-time.

[a] Año de introducción en el mercado.

(https://github.com/nmquijada/tormes[40]), un *pipeline* de código libre para el análisis de genomas bacterianos, que permite ensamblar y anotar el genoma, tipificar el microorganismo por *Multilocus Sequence Typing* (MLST), obtener los factores de virulencia y de resistencia a los antibióticos, y realizar un análisis pangenómico comparativo de los aislados. Los resultados generados son recopilados en un archivo interactivo en formato web, que puede visualizarse en cualquier navegador, y ser compartido y analizado entre los distintos usuarios de una manera sencilla.

Por su parte, el *National Center for Biotechnology Information* (NCBI) dispone de un proyecto llamado *Reference Sequence* (RefSeq), que incluye el *Prokaryotic Genome Annotation Pipeline* (PGAP), ya en su versión 4.1, y un repositorio de genomas procariotas curados y anotados (www.ncbi.nlm.nih.gov/refseq/)[21].

Existen diversas plataformas web con genomas depositados y curados, como EnteroBase[3] (https://enterobase.warwick.ac.uk/), donde se pueden analizar y tipificar genomas de enterobacterias por MLST, *core genome* MLST (cgMLST) o *whole genome* MLST (wgMLST). Otra interesante plataforma es Pathogenwatch (https://pathogen.watch/), que permite acceder a una serie de análisis automáticos a partir de secuencias propias (ensamblados *de novo* en formato FASTA) y determinar la especie bacteriana por MLST, así como predecir la resistencia a antibióticos y acceder a secuencias públicas de genomas bacterianos completos (también de hongos y virus), que, en última instancia, pueden integrarse con genomas propios para generar colecciones, que sirven para comparar estas secuencias mediante árboles filogenéticos. Microreact[4] (https://microreact.org/showcase) es otra plataforma que permite una visualización integrada de los conjuntos de datos de secuenciación (árboles filogenéticos, determinantes de resistencia y virulencia), datos geográficos y temporales, y otras variables de interés para realizar epidemiología genómica.

Además del *software*, es necesario disponer de un *hardware* o estructura de cómputo de gran capacidad. Consideramos que para poder realizar algunos de los tipos de análisis bioinformáticos, como el mapeo de referencia o el ensamblaje *de novo* (ambos descritos más adelante), se pueden emplear ordenadores de escritorio o portátiles con *software* y configuración adecuados (16 GB de memoria RAM, 4 núcleos de procesamiento y 1 TB de espacio de almacenamiento). Sin embargo, el uso de máquinas de alta gama ubicadas en grandes centros de datos (miles de GB de memoria, hasta 64 núcleos de procesamiento y acceso a cientos o miles de TB de almacenamiento), incluso vinculadas entre sí para construir grupos de cómputo de alto rendimiento o *clusters* (capaces de analizar simultáneamente cientos o miles de genomas), constituye una práctica eficiente en cuanto al uso de los recursos. Algunos países han creado una infraestructura común para ejecutar los análisis con un gran poder computacional, lo que requiere un conocimiento informático mínimo para el acceso, además de la instalación y el uso de distintos programas o *software* ya preinstalados en los equipos. Por ejemplo, el Reino Unido creó en 2016 *Cloud Infrastructure for Microbial Bioinformatics* (CLIMB, www.climb.ac.uk.), un recurso informático para la comunidad científica en microbiología clínica, que provee al investigador de capacidad de procesamiento (múltiples *cores* y memoria RAM) y memoria de almacenamiento, con numerosos programas preinstalados[13]. Con el fin de ahorrar al usuario la tarea de instalar programas y entornos, se han creado distintos repositorios, como es la iniciativa Bio-Linux (http://environmentalomics.org/bio-linux/), que permite descargar un sistema operativo (Linux) en el ordenador, con una gran variedad de lenguajes y programas bioinformáticos. Con objeto de iniciar al microbiólogo en la bioinformática y facilitarle su comprensión, en la tabla 2 se presenta un pequeño glosario de terminología básica.

En los apartados subsiguientes se describe la aplicación de la secuenciación masiva en la microbiología clínica; asimismo, se desarrollan en extenso dos de sus principales usos.

Aplicaciones bioinformáticas para la secuenciación masiva en la microbiología clínica

La secuenciación masiva aplicada a la microbiología se puede realizar secuenciando pequeños fragmentos del ácido desoxirribonucleico (ADN) o amplicones previamente amplificados *(targeted sequencing)*, o bien secuenciando todo el ADN previamente fragmentado de forma aleatoria *(shotgun sequencing)*. La aproximación de *targeted sequencing* permite obtener la secuencia del mismo gen en muchas muestras; por ejemplo, en el caso de genes considerados «relojes moleculares» (como los genes ARNr 16S y 18S), este análisis revela la composición microbiana de cada muestra (bacteriana y fúngica, respectivamente). Mediante *shotgun sequencing* puede obtenerse el genoma completo de una bacteria. Debido a la corta longitud de las secuencias generadas por algunas plataformas de secuenciación, la obtención de un genoma bacteriano completamente cerrado es una tarea compleja y, en algunos casos, imposible, por lo que las *reads* se ensamblan en un *draft genome* o borrador del genoma formado por un número de fragmentos más grandes o *contigs*, mediante un proceso de ensamblado de las lecturas secuenciadas (*de novo* o *reference-based genome assemblies*). El conocimiento del genoma facilita la detección de diferencias (mutaciones, SNPs, InDels) entre genomas (genómica comparada), lo que es de gran utilidad en la identificación de los mecanismos de resistencia a antibióticos y en estudios de epidemiología molecular y dinámica de transmisión, así como de filogenética y filogeografía, y también en la estimación de la velocidad de mutación.

Un experimento de secuenciación masiva consta de 4 etapas principales: la extracción del ADN de la muestra o aislado, la preparación de las bibliotecas o librerías, la secuenciación propiamente dicha y el análisis bioinformático e interpretación de los resultados. Antes de introducir una muestra en el secuenciador, es necesario la preparación de bibliotecas (que denominaremos librerías) de un tamaño determinado, es decir, hay que preparar los fragmentos que van a ser secuenciados. Ello implica fragmentar el ADN por métodos bioquímicos (enzimáticos) o físicos (nebulización o ultrasonido), o bien amplificar fragmentos del ADN (por reacción en cadena de la polimerasa [PCR]). Posteriormente se realiza el marcado de dichos fragmentos con

Tabla 2 Secuenciación masiva: terminología básica

Término o sigla	Definición
Amplicón	Producto generado por amplificación de ADN en una reacción en PCR usando un par de *primers* o cebadores.
Amplicon sequencing	Secuenciación masiva de productos de PCR previamente amplificados.
ANI	*Average Nucleotide Identity* es el método de análisis de la identidad de nucleótidos entre regiones o genomas.
Anotación	Asignación de una función a un gen conocido.
CDS	*Coding Sequence* es parte del ARNm o secuencia genómica que codifica una secuencia de proteína.
Contig	*Contiguous sequence:* secuencia de ADN que procede de dos o más secuencias que se superponen en sus extremos y se pueden juntar en una sola secuencia no redundante.
De novo assembly	Ensamblado de un genoma basado únicamente en la información que contienen las *reads*, sin necesidad de comparación con un genoma de referencia.
Draft genome	Borrador del genoma es la versión no cerrada del genoma que cubre entre el 95 y el 9% de aquel, obtenido a partir de la secuenciación de lecturas cortas que se solapan formando *contigs*.
Ensamblado	Proceso por el cual los fragmentos cortos del ADN secuenciados se juntan en fragmentos más grandes hasta reconstruir el genoma.
HTS	*High Throughput Sequencing* es la tecnología de secuenciación masiva que permite obtener gran cantidad de información de forma rápida.
InDel	Inserciones y deleciones de pequeño tamaño (< 10 pb) que existen en el genoma y pueden usarse como marcadores moleculares para el mapeo de determinados caracteres.
k-mers	Subcadenas de caracteres en las que se compone cada secuencia y que son utilizadas por distintos procesos bioinformáticos, como el ensamblado *de novo* basado en gráficos de Bruijn por SPAdes o la asignación taxonómica de Kraken.
Mapping o mapeado	Alineamiento de cada *read* a una posición en el genoma de referencia.
Mate-pair sequencing	Tecnología de secuenciación del ADN en ambos sentidos 5′ y 3′ a partir de la ligación de fragmentos del ADN biotinilados de 2-5 kb.
MLST	*Multilocus Sequence Typing* es el método de tipado de microorganismos basado en la secuenciación de una serie de genes que se comparan con unos esquemas definidos para cada especie.
NGS	*Next generation sequencing* es la tecnología de secuenciación masiva que surgió después de la de Sanger.
Paired-end sequencing	Secuenciación de lectura pareada que consiste en la secuenciación de fragmentos del ADN en ambos sentidos 5′ y 3′.
Phred score	Valor de calidad de cada base secuenciada. Un valor de 30 indica que la precisión de la base secuenciada es del 99,9%.
Pipeline	Es una serie de múltiples *software* que se combinan para llevar a cabo un análisis automático determinado de forma secuencial o en paralelo.
Read	Lecturas o secuencia del ADN continua obtenida de un secuenciador.
Sequence coverage	Es la cobertura o la parte estimada del genoma que ha sido secuenciada.
Sequence depth	Profundidad de cobertura es la media del número de veces que cada base de un genoma secuenciando tiene una *read* que alinea en esa posición.
Single-end sequencing	Secuenciación de lectura simple que consiste en la secuenciación de fragmentos del ADN que son secuenciados solo en una dirección o sentido.
SNP	*Single Nucleotide Polymorphism:* polimorfismo de un único nucleótido que puede interrumpir la función de un gen o bien ser fuente de variación del ADN y usarse como marcador.

ADN: ácido desoxirribonucleico; PCR: reacción en cadena de la polimerasa.

índices *(barcoding)*, para poder secuenciar a la vez múltiples muestras equimolecularmente *(multiplexear)*. También conlleva el reparado de los extremos y la incorporación en los fragmentos del ADN de 2 tipos, un índice que permite la secuenciación de múltiples muestras y un pequeño adaptador del ADN complementario a aquel que existe en el secuenciador, que permite que los fragmentos puedan adherirse a un soporte *(flow cell)* para ser secuenciados, como es el caso de la tecnología Illumina. La forma de secuenciar puede ser en un sentido de la doble hebra del ADN *(single-end sequencing)* o en ambos, es decir, 5′ y 3′, en lo que se denomina *paired-end sequencing*. Este último procedimiento es el más recomendable en microbiología, porque permite obtener fragmentos más largos, de casi 600 pb en el caso del MiSeq (Illumina), lo que mejora el ensamblado ulterior.

La etapa de análisis de datos conlleva, a su vez, una serie de pasos: el análisis primario (la generación y el control de calidad), el secundario (alineamiento contra bases de datos específicas, ensamblado de referencia o *de*

novo) y el terciario (generación de datos a partir de los resultados de la etapa de análisis secundario: anotación, búsqueda de SNPs, determinantes de resistencia y/o virulencia, etc.). Una vez concluida la secuenciación, se obtiene un archivo de datos FASTQ para cada uno de los *paired-ends* (si fue este el tipo de secuenciación realizada), es decir, uno para el *forward* o secuencia sentido y otro para el *reverse* o secuencia antisentido correspondientes a cada muestra; dicho archivo contiene las secuencias o *reads* y los datos de calidad *(phred score)* basados en código ASCII (en el que cada letra representa un valor numérico), de manera que una secuencia con *phred score* = 10 tiene un 90% de eficacia (una base mal secuenciada cada 10) y una secuencia con *phred score* = 30 tiene un 99,9% de eficacia (una base mal secuenciada cada 1.000). A partir de este archivo FASTQ se realiza en todos los casos el filtrado por calidad, en el que diferentes parámetros como el *phred*, la longitud, el contenido en GC, los *primers*, los *barcodes*, etc., deben ser evaluados con programas como FastQC, en tanto que otros programas como Trimmomatic, Prinseq o Sickle se usan para eliminar restos de *barcodes* y *primers*, y filtrar las secuencias basándose en su calidad.

Se han desarrollado numerosos programas; en la tabla 3 se recogen algunos de ellos. Una vez que se obtiene la secuencia filtrada por calidad, se debe analizar mediante alineamiento contra una base de datos específica, para identificar los taxones microbianos de una muestra, o mediante ensamblado y anotado de los genes, para construir el borrador del genoma que permita detectar diferencias entre genomas o identificar los genes de interés, tal y como se va a describir a continuación. En nuestro grupo, hemos realizado trabajos de secuenciación masiva tanto para caracterizar la población microbiana de una muestra con el objeto de identificar marcadores[1], como también para asociar la disbiosis intestinal a una enfermedad[22]. En cuanto a la caracterización de genomas completos, nos ha permitido descubrir nuevos determinantes de resistencia a antibióticos[23] y realizar estudios de epidemiología molecular[24].

Análisis de la microbiota de una muestra

La identificación y la caracterización de los microorganismos que pueden estar causando una infección o una enfermedad es importante para el tratamiento y la recuperación del paciente, y también para la seguridad del resto de los individuos. Se define como microbiota la composición taxonómica microbiana de una muestra, mientras que el término metagenoma alude al conjunto de genes y genomas de la microbiota. Bajo el término microbioma se engloba el conjunto de genes y, además, sus productos o metabolitos. En todos los casos, se trata de establecer las relaciones ecológicas microbianas en un determinado ambiente o muestra.

Tradicionalmente, la aproximación empleada en estudios de ecología microbiana se basaba en el aislamiento de los microorganismos en cultivo axénico y la purificación de su ADN, o bien la extracción directa del ADN a partir de la muestra y la posterior amplificación por PCR, seguida de la separación de los fragmentos en geles de agarosa o de acrilamida (como sucede en la electroforesis en gel en gradiente desnaturalizante, DGGE) y la identificación de los aislados por secuenciación Sanger. Sin embargo, estos métodos no ofrecen las ventajas de la secuenciación masiva: esta última es independiente del cultivo, se necesita poco ADN de la muestra, tiene suficiente profundidad para identificar microorganismos que estén poco representados y permite *multiplexear* varias muestras a un bajo coste y en un tiempo reducido.

Como se ha comentado, la mayoría de los métodos de ecología bacteriana se basan en la secuenciación, en concreto, del gen que codifica la subunidad menor del ribosoma 16S (16S ARNr). El tamaño de este gen es de aproximadamente 1.540 pares de bases, dependiendo de la especie bacteriana, y está formado por 9 regiones hipervariables (V1 a V9) y regiones altamente conservadas. Sobre esta base se han definido numerosas combinaciones de *primers* o cebadores[8,29], que permiten amplificar un fragmento del ADN de forma inespecífica en todas las especies bacterianas, pero a su vez es posible realizar la identificación inequívoca de cualquier bacteria, porque la secuenciación de la región interna es característica de cada especie.

La secuenciación Sanger no permitía la mezcla de amplicones ya que el cromatograma no podía ser leído si no era puro, pero en la secuenciación masiva sí, porque esta tecnología individualiza la lectura de cada amplicón. El estudio ecológico molecular de la microbiota mediante NGS consiste en la coamplificación de una región del gen 16S ARN. Por ejemplo, los *primers* descritos por Klindworth et al.[29] amplifican las regiones variables V3-V4 del gen 16S ARNr (~464 pb) a partir del ADN extraído directamente de la muestra. Estos fragmentos amplificados, una vez secuenciados, permiten caracterizar la comunidad bacteriana coexistente, con la posibilidad de identificar todas las especies de una muestra; incluso se puede realizar el análisis subespecie de los taxones mediante herramientas como *oligotyping*[16] o la resolución de *amplicon sequence variants* (ASVs)[6]. La asignación de filo, clase, orden, familia, género o especie se realiza por comparación con bases de datos como la *Ribosomal Database Project* (RDP)[11], *Greengenes*[37], Silva (www.arb-silva.de), el servidor web del CGE (https://cge.cbs.dtu.dk/services/SpeciesFinder/), o bien realizando un alineamiento empleando la *Basic Local Alignment Search Tool* (BLAST) de la base de datos del NCBI[17]. No solo podemos obtener la caracterización de la composición taxonómica, sino también la abundancia relativa de las comunidades microbianas de una muestra. La secuenciación masiva permite el análisis ecológico de la microbiota mediante secuenciación de amplicones, pero también se puede realizar este análisis mediante la secuenciación completa de todos los genomas, es decir, del metagenoma. Esto implica mayor costo y requiere un análisis más complejo y detallado, por lo que esta revisión no profundiza en el análisis metagenómico *sensu stricto*[41].

El proceso de análisis de la microbiota de una muestra dada se describe en la figura 1. Una vez purificado el ADN microbiano de la muestra, realizada la PCR y la librería, y secuenciados los amplicones, se obtiene el archivo FASTQ. De los datos de este archivo de secuenciación no se obtienen especies *sensu stricto*, sino que se agrupan las secuencias idénticas con un 97 al 99% de identidad usando el método uclust[15] en lo que se denominan *operational taxonomic units* (OTUs). Sin embargo, hoy ya se utilizan las *exact sequence variants* (ESVs) —también denominadas *amplicon sequence*

Tabla 3 *Software* de uso libre para el análisis de datos de secuenciación masiva

Acción	Programa	Sitio web
Asignación taxonómica	RDP	https://rdp.cme.msu.edu/
	Greengenes	http://greengenes.secondgenome.com/
	Silva	www.arb-silva.de
Calidad de secuencia	FastQC	www.bioinformatics.babraham.ac.uk
	TRIMMOMATIC	http://www.usadellab.org/cms/?page=trimmomatic
	PRINSEQ	http://prinseq.sourceforge.net/
Identificación	K-merFinder	www.genomicepidemiology.org
	BLAST	www.ncbi.nlm.nih.gov/blast
	MEGAN	http://ab.inf.uni-tuebingen.de/software/malt
	Kraken	https://ccb.jhu.edu/software/kraken/
Ensamblado	SPAdes	http://bioinf.spbau.ru/spades
	Velvet	www.ebi.ac.uk/~zerbino/velvet
	MIRA	www.chevreux.org/projects_mira.html
Alineamiento	Mauve	http://darlinglab.org/mauve/mauve.html
Visualización	ACT	www.sanger.ac.uk/science/tools
	Artemis	www.sanger.ac.uk/science/tools
	ClustalW	www.genome.jp/tools/clustalw
	BRIG	http://brig.sourceforge.net/
Anotación	Prokka	https://github.com/tseemann/prokka
	RAST	http://rast.nmpdr.org
	Prodigal	https://github.com/hyattpd/prodigal/wiki
Mapeado	Bowtie	http://bowtie-bio.sourceforge.net/index.shtml
	BWA	http://bio-bwa.sourceforge.net/
Filogenia	FastTree	www.microbesonline.org/fasttree/
	SNPTree	www.genomicepidemiology.org
	iTOL	http://itol.embl.de/
Resistencia	ARDB	https://ardb.cbcb.umd.edu
	CARD	https://card.mcmaster.ca
	ResFinder	https://cge.cbs.dtu.dk/services/ResFinder/
	Abricate	https://github.com/tseemann/abricate
	ARG-ANNOT	http://www.mediterranee-infection.com/article.php?laref=282&titre=arg-annot
	PointFinder	https://bitbucket.org/genomicepidemiology/pointfinder
SNP	Samtools	www.htslib.org
	VarScan	http://dkoboldt.github.io/varscan/
	GATK	https://software.broadinstitute.org/gatk/
	Snippy	https://github.com/tseemann/snippy
Plásmidos	PlasmidFinder	https://cge.cbs.dtu.dk//services/PlasmidFinder/
	PlasmidSPAdes	http://cab.spbu.ru/software/plasmid-spades/
	PLACNETw	https://castillo.dicom.unican.es/upload/
	Plasmid MLST	https://pubmlst.org/plasmid/
Tipado	BIGSdb	http://bigsdb.readthedocs.io
	EnteroBase	https://enterobase.warwick.ac.uk
	MLST	http://cge.cbs.dtu.dk/services/MLST/
Virulencia	VFDB	www.mgc.ac.cn/VFs
	VirulenceFinder	www.genomicepidemiology.org

variants (ASVs)—, que agrupan únicamente aquellas secuencias idénticas[6].

Dos herramientas bioinformáticas han popularizado el análisis de la microbiota presente en una muestra: Mothur[44] y QIIME[7], que recientemente han lanzado su segunda versión. QIIME dispone de tutoriales pormenorizados para ser ejecutados por línea de comando en la terminal y los archivos generados pueden ser visualizados en la web (https://view.qiime2.org/). La última versión introduce el uso del *software* DADA2, un sistema de filtrado de calidad de secuencia más estricto, que favorece la posterior identificación de ASVs. QIIME2 hace uso de diversos *scripts* para poder realizar alineamiento de las secuencias, construir gráficos de mapas calientes *(heatmaps)* de la taxonomía, inferir árboles filogenéticos y efectuar análisis de diversidad alfa (diversidad inherente en una muestra) y de beta diversidad (diversidad entre muestras). La alfa diversidad de una muestra se establece contando el número de especies presentes *(richness)*, la diversidad relativa de las distintas especies *(diversity)* y la heterogeneidad de la muestra *(evenness)*. Para ello se usan distintas métricas o índices clásicos de ecología: Chao1, Shannon y Simpson.

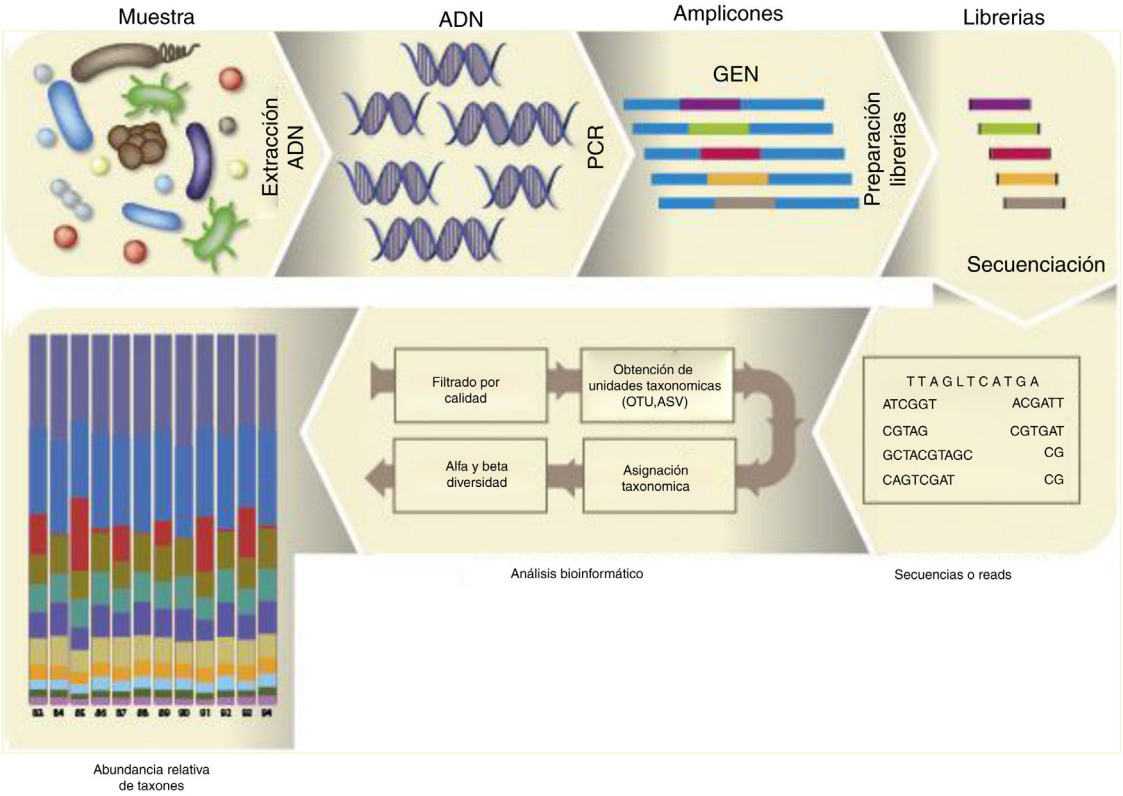

Figura 1 Flujograma del análisis de la microbiota desde las *reads* a la taxonomía.

Las curvas de rarefacción representan el número de especies, ASVs, OTUs, etc., *versus* el número de lecturas, y permiten visualizar de manera sencilla que se ha alcanzado la suficiente profundidad en la secuenciación para identificar aquellos taxones minoritarios en una muestra. La beta diversidad mide la distancia o disimilaridad entre un par de muestras; se pueden realizar distintos análisis estadísticos que representan las muestras en gráficos bi- o tridimensionales (análisis de coordenadas principales, PCoA). Se usan distintas métricas, entre ellas, el índice de Bray-Curtis (método no filogenético que tiene en cuenta la abundancia de OTUs), el *unweighted* UniFrac (método filogenético que tiene en cuenta la presencia o ausencia de una OTU) y el *weighted* UniFrac (método filogenético que tiene en cuenta la abundancia de OTUs). Existen muchas otras herramientas estadísticas, cabe destacar entre ellas el LEfSe, un programa gratuito que permite realizar un análisis discriminante lineal efecto-tamaño *(LDA effect size)* —para identificar las características que explican las diferencias entre clases o predecir marcadores responsables de un fenotipo— y los representa de forma gráfica[46] (https://huttenhower.sph.harvard.edu/galaxy/). Pequeñas modificaciones en algunos de los archivos de QIIME o Mothur permiten a LEfSe realizar el análisis estadístico diferencial entre varias muestras que reúnen distintas condiciones y que son objeto de estudio a todos los niveles taxonómicos.

Análisis del genoma bacteriano completo

Para el análisis de secuencias de genomas bacterianos completos, una vez purificado el ADN microbiano de la muestra es preciso someterlo a la fragmentación (física o bioquímica), seguido de la preparación de librerías y la secuenciación, tal como se describe en la figura 2. El archivo FASTQ generado por el secuenciador masivo a partir de un genoma bacteriano procedente de un cultivo axénico contiene las secuencias, lecturas o *reads*, las que, una vez que se han filtrado por su calidad, pueden ser ensambladas en unidades mayores o *contigs*. Existen varios ensambladores; Loman et al.[34] publicaron una revisión al respecto. Cabe distinguir entre los ensambladores que requieren un genoma de referencia (MIRA) y entre los ensambladores *de novo* (SPAdes). Con secuenciadores como los que comercializa Illumina, con fragmentos secuenciados muy pequeños y habiendo sometido el ADN a una fragmentación enzimática con transposasas (tagmentación) para la preparación de las librerías, resulta imposible cerrar el genoma; sí se podría cerrar usando los secuenciadores con tecnología de molécula única, aunque estos son menos precisos. Por tal razón, ambas tecnologías se consideran complementarias en la actualidad. El resultado, en general, es un *draft genome* en formato FASTA, formado por cierto número de fragmentos o *contigs*, que cubren entre el 95 y el 99% del genoma. El número de *contigs* que forman el genoma es un parámetro que se tiene en cuenta a la hora de evaluar la eficacia del ensamblado, junto con la longitud media y mínima de los *contigs*, el N50 (longitud del *contig* más pequeño de aquellos que representan el 50% del genoma), etc. Otro factor a tener en cuenta es la *depth* o profundidad de cobertura, es decir, el número de veces que una determinada posición nucleotídica fue secuenciada: se consideran genomas de calidad aquellos con una *depth* superior a 20-25 $\times$. Existe también *software* como QUAST[20], que permite analizar la eficiencia del ensamblado, y Qualimap2[38], que

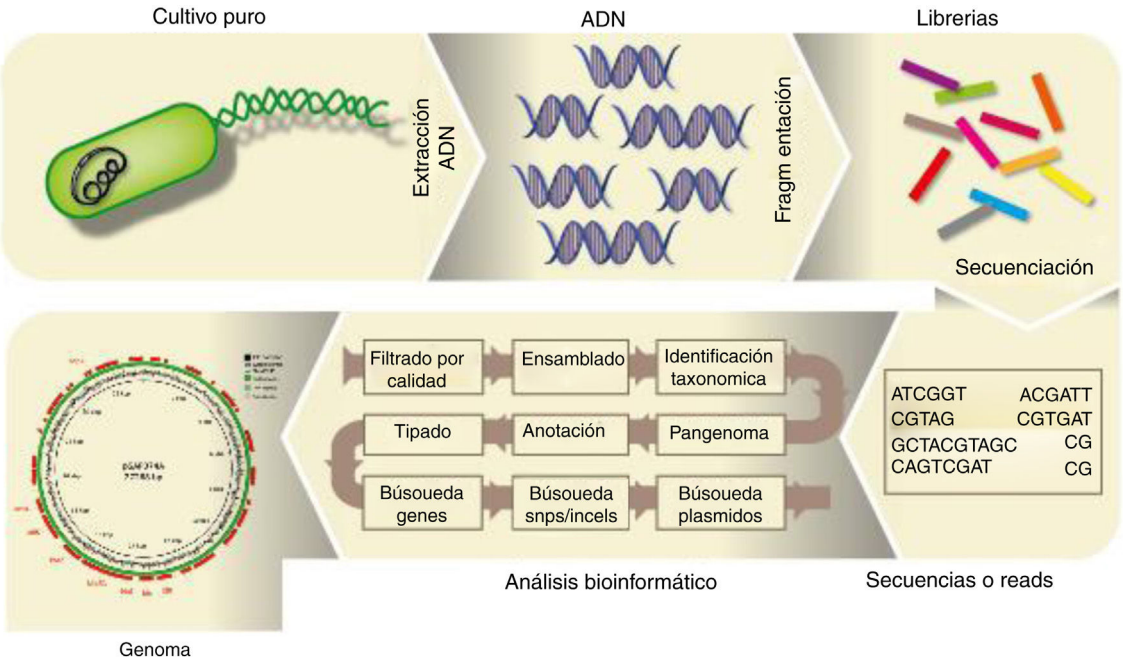

Figura 2 Flujograma del proceso analítico para analizar el genoma bacteriano.

permite analizar la *depth* de cada posición nucleotídica, más allá de una estimación media.

Posteriormente, los *contigs* pueden ser visualizados utilizando *software* como Mauve[14] o Artemis, y ser ordenados frente a un genoma de referencia utilizando, por ejemplo, Mauve o ACT, que a su vez permiten visualizar los genomas utilizados. La figura 2 ilustra el flujograma del proceso para analizar el genoma bacteriano. El ensamblado es uno de los «cuellos de botella» del proceso, computacionalmente hablando, por lo que existen distintos programas que pueden funcionar directamente desde las *reads*. Tal es el caso de Kraken, que permite identificar taxonómicamente una especie o género de forma rápida sin ensamblar las *reads*, utilizando alineamientos exactos de *k-mers* contra su propia base de datos, por lo que es muy útil para detectar contaminaciones[49], o *snippy* (https://github.com/tseemann/snippy), que mapea las *reads* contra un genoma de referencia con el objeto de detectar polimorfismos. También el *software* ARIBA[25] permite detectar genes de resistencia a antibióticos, virulencia o replicones plasmídicos a partir de las *reads*. Existen visualizadores de genomas propiamente dichos, como BLAST Ring Image Generator (BRIG)[2], que permiten la comparación visual de varios genomas frente a uno de referencia y localiza los genes anotados.

Una vez que los *contigs* han sido ordenados en un borrador de genoma más o menos cerrado, se procede a la asignación de las funciones de un genoma, es decir, a su anotación, lo que conlleva una serie de procesos como la búsqueda de posibles genes (*coding sequences* o CDS) mediante programas como Prodigal[26], su traducción a proteínas y la identificación de estas proteínas contra bases de datos (diamond y blastp). Existen herramientas para automatizar el ensamblado en plataformas web, como RAST[5], o por línea de comandos en la terminal usando Prokka[45]. El resultado son distintos archivos con información de anotación del tipo GFF, GBK, GBL, etc., los cuales son siempre dependientes de la base de datos que se haya utilizado para anotar; la del NCBI es una de las más completas. Una de las ventajas de la anotación por programas de línea de comandos radica en la capacidad del usuario de generar bases de datos de genes propias, surgidas a partir de sus investigaciones particulares, ya que, en última instancia, estas mejorarán la calidad de la anotación.

Un objetivo probable cuando se intenta obtener el genoma de una bacteria es poder buscar genes de interés que permitan el genotipado. Por ejemplo, puede ser necesario localizar los genes *housekeeping*, que permiten obtener el MLST. PubMLST contiene todos los esquemas y, específicamente para *Listeria* spp., existe el esquema del Instituto Pasteur (http://bigsdb.pasteur.fr/listeria/listeria.html). Asimismo, EnteroBase tiene esquemas para las principales enterobacterias y actualmente también para *Clostridioides*. Existen plataformas como las MLST[32,35] o el Plasmid MLST Databases[28] y el servidor del *Center for Genomic Epidemiology* (CGE) (http://www.genomicepidemiology.org/), que permiten obtener el secuenciotipo de un aislado. También en el servidor del CGE se puede serotipificar usando la herramienta SerotypeFinder 2.0, hacer spaTyping en el caso de *Staphylococcus aureus*, tipificar *Escherichia coli* por fimbrias con fimTyper, etc. Por otro lado, se pueden buscar genes que codifiquen resistencia o virulencia con *software* como ARIBA (ya mencionado) o Abricate (https://github.com/tseemann/abricate), que usa la herramienta BLAST del NCBI (http://blast.ncbi.nlm.nih.gov) para buscar en el genoma genes de resistencia a antibióticos que se encuentren en bases de datos como ResFinder (acquired antimicrobial resistance gene finder)[50], Comprehensive Antimicrobial Resistance Database (CARD)[27],

Antibiotic Resistance Genes Database (ARDB) o Antibiotic Resistance Gene-ANNOTation (ARG-ANNOT)[19]. En cuanto a los genes de virulencia, existe una base de datos denominada Virulence Factor Database (VFDB)[10], que contiene los principales determinantes de virulencia bacterianos y puede usarse con la herramienta BLAST.

Una de las aplicaciones más interesantes de la secuenciación masiva es la posibilidad de detectar variantes (SNPs, Indels). Para ello se mapean las *reads* directamente contra un genoma de referencia utilizando *software* como SMALT (https://www.sanger.ac.uk/science/tools/smalt-0), Bowtie[30] o BWA[33]. El archivo resultante es *Sequence Aligmant Map* (SAM) y debe ser manipulado con Samtools y VarScan o GATK para generar un archivo *Variant Calling File* (VCF), que pueda ser considerado como suficiente para definir SNP/InDel. Con esta aproximación se buscan todas aquellas *reads* que tengan un nucleótido que difiera con el genoma de referencia, es decir, permite obtener todas aquellas *reads* que soportan un SNP o InDel. No podemos definir un número concreto de *reads* que pueda ser considerado como suficiente, ya que hay muchos aspectos por considerar (como la calidad de esa posición y otros). Los archivos SAM/BAM y/o VCF pueden visualizarse con Artemis o IGV. Algunos tipos de *software* como Parsnp permiten la comparación de aislados mediante *core-genome*-SNP, esto es, mediante la comparación de variantes nucleotídicas en cada una de las posiciones. Teniendo en cuenta la información de los SNP, es posible construir dendrogramas, que pueden ser visualizados con programas como Gingr, FigTree, PhyML o RAxML. El programa FastTree utiliza métodos heurísticos de máxima verosimilitud para generar un árbol a partir de alineamientos de nucleótidos o secuencias proteicas. Los árboles se pueden visualizar y manipular, por ejemplo, con iTOL, con el paquete de R ggtree y con el ya mencionado Figtree.

Otra aplicación de la secuenciación de genomas completos es comparar las regiones del genoma obtenido contra un genoma de referencia. Se pueden comparar ambos genomas utilizando BLAST y visualizarlos mediante Artemis Comparison Tool. Con la información de todos los genes obtenemos el denominado pangenoma (todos los genes presentes en un genoma), el genoma *core* (genes que están presentes en cualquier genoma dentro de un filotipo) y los genes accesorios (genes que están presentes en un conjunto de aislados dentro del conjunto de datos). Se puede comparar el pangenoma de distintos aislados anotados mediante programas como Roary[39], que nos permite identificar a partir del archivo GFF los *core genes* (genes compartidos en el 99-100% de los genomas comparados), *softcore* (95-99%), *shell* (15-95%) y los *cloud genes* (0-15%), y construir una matriz de distancias para generar un árbol basándose en la presencia/ausencia de genes en las distintas muestras. También existen esquemas de tipificado basados en *core genome* MLST (cgMLST) y *whole genome* MLST (wgMLST), es decir, basados en un conjunto fijo de *loci* conservados en todos los genomas y se usan esquemas específicos de especie[32].

Por último, la secuenciación permite obtener tanto la información genética cromosómica como la contenida en plásmidos como material transferible y de replicación autónoma. Así, por ejemplo, el *software* PlasmidFinder permite la identificación de motivos plasmídicos usando una base de datos no redundante de replicones (secuencias que activan o controlan la replicación del plásmido y son identificativas de un determinado tipo de plásmido) y los asigna a un determinado grupo Inc (grupo de incompatibilidad plasmídica) o Rep[9]. Posteriormente se siguen distintas aproximaciones para identificar las *reads* plasmídicas. PlasmidSPAdes es otro *software* basado en el concepto de que, durante la extracción de ADN, a igual concentración molar de ADN habrá más número de plásmidos que de cromosomas; esta mayor concentración de plásmidos se traduce en un mayor número de *reads* plasmídicas tras la secuenciación y el programa utiliza un algoritmo que es capaz de separar y tratar de ensamblar estas *reads* que están en un mayor porcentaje que el supuesto cromosoma. Similarmente, pueden mapearse las *reads* contra un cromosoma de referencia mediante Bowtie o BWA y efectuar el descarte de aquellas que mapeen (y, por tanto, con posibilidad de ser cromosómicas) y el ensamblado posterior de aquellas que no mapearon con SPAdes. Lanza et al.[31] desarrollaron una herramienta denominada PLACNET para la reconstrucción gráfica de plásmidos, cuya versión web[48] apareció en 2017.

Cualquiera sea el *software*, siempre es más interesante realizar los análisis por línea de comandos que en la web, porque se pueden analizar múltiples genomas a la vez y lanzar procesos en paralelo, lo cual reduce el tiempo de procesado (siempre dependiendo de la capacidad de procesamiento del *hardware*) e independiza al usuario de la disponibilidad de servidores web por lo general más lentos.

Conclusiones y perspectivas futuras

La secuenciación masiva ha transformado la microbiología, sobre todo desde que se han reducido los costes y los tiempos de análisis, gracias al desarrollo de la bioinformática, que va generando nuevos programas de análisis. La masiva generación de datos requerirá cada vez más inversiones en grandes centros de supercomputación, y el desarrollo de programas de código abierto será cada vez más profuso, aunque, paralelamente, se comercializarán paquetes de uso bioinformático que requerirán mínimos conocimientos técnicos.

La secuenciación masiva se impondrá en el análisis microbiano, ya que permite estudiar la epidemiología y trazar microorganismos individuales en pacientes, en los hospitales, en la comunidad y en el planeta en general, además de estudiar su evolución. Por ejemplo, Comas et al.[12] han revisado las aportaciones de la secuenciación masiva al diagnóstico y la epidemiología de la tuberculosis y los esfuerzos que se están haciendo para dar el salto a su aplicación en el contexto clínico. Pero la aplicación de la secuenciación masiva va más allá de la medicina humana. Teniendo en cuenta que 7 de cada 10 enfermedades infecciosas son de origen animal, estas tecnologías pueden favorecer la cooperación en el estudio de las enfermedades en humanos y animales, y su interacción en el medio ambiente bajo el concepto «una sola salud»[51]. Asimismo, también se puede conocer la situación de la comunidad microbiana en el nicho ecológico estudiado (en lugar de obtener información de los microorganismos aislados) y, por tanto, establecer asociaciones microbianas que pueden explicar la etiología o la evolución de algunas enfermedades o síndromes de causa desconocida o no bien definida. La secuenciación masiva, a

través de su aplicación a la exploración de la microbiota y del metagenoma, ha cambiado el modo en el que el microbiólogo investiga, y su potencial con las nuevas herramientas bioinformáticas es enorme. El descubrimiento de nuevos genes que codifican resistencias a antibióticos, factores de virulencia, enzimas; en definitiva, de nuevas funciones de la célula procariota, abre un enorme campo para investigar, muy dependiente del avance de los métodos de secuenciación y también del desarrollo de algoritmos informáticos de análisis.

Tanto las bacterias patógenas como las no patógenas pueden representar una amenaza, pero también una oportunidad para la salud. Describir la epidemiología y los mecanismos de virulencia, así como las interacciones con el organismo hospedador, proporciona un conocimiento indispensable para establecer la relación salud-enfermedad y obtener verdaderos triunfos en enfermedades infecciosas y grandes avances en el tratamiento de muchas de ellas. No obstante, los métodos aquí descritos requieren una estandarización y validación para que los resultados sean comparables, como garantía de que la extensa cantidad de datos que se están produciendo son fiables y representan un progreso real en la generación de conocimiento. Además, la secuenciación masiva y su análisis permite la integración de la microbiología humana con la veterinaria, con el objetivo de vigilar los microorganismos zoonóticos conocidos y su transmisión, y de estar atentos a la posible aparición de nuevas amenazas. La comprensión de la salud en términos globales, atendiendo a los microrganismos circulantes entre la población, pero también entre los animales y el medio ambiente, constituye la clave del éxito en el control de las patologías infecciosas presentes y futuras.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Bibliografía

1. Adserias-Garriga J, Quijada NM, Hernandez M, Rodríguez Lázaro D, Steadman D, Garcia-Gil LJ. Dynamics of the oral microbiota as a tool to estimate time since death. Mol Oral Microbiol. 2017;32:511–6.
2. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. BMC Genomics. 2011;12:402.
3. Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of *Salmonella*. PLoS Genet. 2018;14:e1007261.
4. Argimón S, Abudahab K, Goater R, Fedosejev A, Bhai J, Glasner C, Feil E, Holden M, Yeats C, Grundmann H, Spratt B, Aanensen D. Microreact: Visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom. 2016;2:e000093, http://dx.doi.org/10.1099/mgen.0.000093.
5. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. The RAST Server: Rapid annotations using subsystems technology. BMC Genomics. 2008;9:75, http://dx.doi.org/10.1186/1471-2164-9-75.
6. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11:2639–43.
7. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.
8. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. 2011;108 Suppl 1:4516–22.
9. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58:3895–903.
10. Chen LH, Zheng DD, Liu B, Yang J, Jin Q. VFDB 2016 hierarchical and refined dataset for big data analysis-10 years on. Nucleic Acids Res. 2016;44:D694–7.
11. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42:D633–42.
12. Comas I, Gil A. Secuenciación masiva para el diagnóstico y la epidemiología de tuberculosis. Enferm Infecc Microbiol Clin. 2016;34 Supl 3:32–9.
13. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ, Richardson E, Ismail M, Thompson SE, Kitchen C, Guest M, Bakke M, Sheppard SK, Pallen MJ. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): An online resource for the medical microbiology community. Microb Genom. 2016;2:e000086.
14. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.
15. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26:2460–1.
16. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 2013;4:1111–9.
17. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012;40:D136–43.
18. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;269:496–512.
19. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain JM. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014;58:212–20.
20. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
21. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu F, Marchler GH, Song JS, Thanki N, Yamashita RA, Zheng C, Thibaud-Nissen F, Geer LY, Marchler-Bauer A, Pruitt KD. RefSeq: An update on prokaryotic genome annotation and curation. Nucleic Acids Research. 2018;46:D851–60.
22. Hernández M, de Frutos M, Rodríguez-Lázaro D, López-Urrutia Lorente L, Quijada NM, Eiros JM. Fecal microbiota of toxigenic *Clostridioides difficile*-associated diarrea. Front Microbiol. 2018;9:3331.

23. Hernández M, Iglesias MR, Rodríguez-Lázaro D, Gallardo A, Quijada N, Miguela-Villoldo P, Campos MJ, Píriz S, López-Orozco G, de Frutos C, Sáez JL, Ugarte-Ruiz M, Domínguez L, Quesada A. Co-occurrence of colistin-resistance genes *mcr-1* and *mcr-3* among multidrug-resistant *Escherichia coli* isolated from cattle, Spain September 2015. Euro Surveill. 2017;22, pii:30586.
24. Hernández M, Quijada NM, Lorente LL, de Frutos M, Rodríguez-Lázaro D, Eiros JM. Infrequent isolation of extensively drug-resistant (XDR) *Klebsiella pneumoniae* resistant to colistin in Spain. Int J Antimicrob Agents. 2018;51:531–3.
25. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, Harris SR. ARIBA: Rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genom. 2017;3:e000131, http://dx.doi.org/10.1099/mgen.0.000131.
26. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
27. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN, Doshi S, Courtot M, Lo R, Williams LE, Frye JG, Elsayegh T, Sardar D, Westman EL, Pawlowski AC, Johnson TA, Brinkman FS, Wright GD, McArthur AG. CARD 2017 expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 2017;45:D566–73, http://dx.doi.org/10.1093/nar/gkw1004.
28. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124, http://dx.doi.org/10.12688/wellcomeopenres.14826.1.
29. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013;41:e1, http://dx.doi.org/10.1093/nar/gks808.
30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9, http://dx.doi.org/10.1038/nmeth.1923.
31. Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J, Coque TM, de la Cruz F. Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. PLoS Genet. 2014;10:e1004766, http://dx.doi.org/10.1371/journal.pgen.1004766.
32. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol. 2012;50:1355–61, http://dx.doi.org/10.1128/JCM. 06094-11.
33. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95, http://dx.doi.org/10.1093/bioinformatics/btp698.
34. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ. High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. Nat Rev Microbiol. 2012;10:599–606, http://dx.doi.org/10.1038/nrmicro2850.
35. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. MLST revisited: The gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013;11:728–36, http://dx.doi.org/10.1038/nrmicro3093.
36. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376–80, http://dx.doi.org/10.1038/nature03959.
37. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6:610–8.
38. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32:292–4, http://dx.doi.org/10.1093/bioinformatics/btv566.
39. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3, http://dx.doi.org/10.1093/bioinformatics/btv421.
40. Quijada NM, Rodríguez-Lázaro D, Eiros JM, Hernández M. TORMES: An automated pipeline for whole bacterial genome analysis. Bioinformatics. 2019, http://dx.doi.org/10.1093/bioinformatics/btz220.
41. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35:833–44, http://dx.doi.org/10.1038/nbt.3935.
42. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM, Smith M. Nucleotide sequence of bacteriophage ΦX174 DNA. Nature. 1977;265:687–95, http://dx.doi.org/10.1038/265687a0.
43. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94:441–8, http://dx.doi.org/10.1016/0022-2836(75)90213-2.
44. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: Open-Source, Platform-Independent Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol. 2009;75:7537–41, http://dx.doi.org/10.1128/AEM. 01541-09.
45. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9, http://dx.doi.org/10.1093/bioinformatics/btu153.
46. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12:R60, http://dx.doi.org/10.1186/gb-2011-12-6-r60.
47. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet. 2018;34:666–81, http://dx.doi.org/10.1016/j.tig.2018.05.008.
48. Vielva L, de Toro M, Lanza VF, de la Cruz F. PLACNETw: A web-based tool for plasmid reconstruction from bacterial genomes. Bioinformatics. 2017;33:3796–8, http://dx.doi.org/10.1093/bioinformatics/btx462.
49. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46, http://dx.doi.org/10.1186/gb-2014-15-3-r46.
50. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67:2640–4, http://dx.doi.org/10.1093/jac/dks261.
51. Zinsstag J, Schelling E, Waltner-Toews D, Tanner M. From «one medicine» to «one health» and systemic approaches to health and well-being. Prev Vet Med. 2011;101:148–56, http://dx.doi.org/10.1016/j.prevetmed.2010.07.003.