Ciencias Marinas (2015), 41(4): 297–313
http://dx.doi.org/10.7773/cm.v41i4.2492

Review article/Artículo de revisión

The 16S rRNA gene in the study of marine microbial communities

El gen ARNr 16S en el estudio de comunidades microbianas marinas

Fabiola Valenzuela-González, Ramón Casillas-Hernández, Enrique Villalpando, Francisco Vargas-Albores*

Centro de Investigación en Alimentación y Desarrollo, PO Box 1735, Hermosillo CP 83000, Sonora, México

*Corresponding author: fvalbores@ciad.mx

ABSTRACT. New sequencing technologies and analytical capabilities have stimulated the study of microbial communities from specific environments, enabling researchers to understand the complexity of those systems. The 16S rRNA gene has proved very useful in describing the diversity and characterization of marine microbial communities, particularly of uncultivated organisms. The development of new sequencing techniques has contributed to the exponential increase in the number of reported 16S rRNA sequences as barcodes for microorganisms, forcing a review of concepts and methods for the taxonomic classification of these organisms. Manipulation and analysis of large amounts of genetic information have prompted the development of specific databases, specialized algorithms, and computational tools to compare thousands of such sequences and make a taxonomic assignment. Complete 16S rRNA sequences are thus needed for accurate and reproducible taxonomy assignment in the study of marine bacterial communities.

Key words: metagenomics, marine microbial communities, 16S rRNA databases.

RESUMEN. Las nuevas tecnologías de secuenciación y capacidades analíticas han estimulado el estudio de las comunidades microbianas de ambientes específicos, lo cual ha permitido conocer la complejidad de estos sistemas. El gen ARNr 16S ha demostrado ser de gran utilidad para describir y caracterizar las comunidades microbianas marinas, especialmente los organismos no cultivados. Las nuevas técnicas de secuenciación han contribuido al incremento exponencial de registro de secuencias, aunque parciales, del ARNr 16S como elemento de código de barras para microorganismos. Consecuentemente, ha sido necesaria una revisión de conceptos y métodos de clasificación taxonómica para estos organismos. El manejo y análisis de una gran cantidad de información génica han impulsado el desarrollo de bases de datos específicas, algoritmos y herramientas computacionales especializadas para comparar miles de secuencias semejantes y hacer la asignación taxonómica. Por lo tanto, las secuencias completas del ARNr 16S son necesarias para una asignación taxonómica certera y reproducible en los estudios de comunidades microbianas marinas.

Palabras clave: metagenómica, comunidades microbianas marinas, bases de datos ARNr 16S.

INTRODUCTION

The oceans, the world's largest ecological system, are inhabited by appproximately 3.6×10^{28} microorganisms at a mean density of 5×10^5 cells mL^{-1} (Cock *et al.* 2010). Having originated in a similar environment, marine microbial communities are highly abundant and diverse and have developed successful physiological adaptations (Hellweger *et al.* 2014).

The importance and impact of microbial communities in a specific environment have been demonstrated by studies of the microbiota that inhabit the human body. The human microbiome is a reservoir of approximately 10^9 microorganisms. There are ten times more microbes in the human body than human cells, and the microbiological community thus plays a major role in human health and disease (NRC 2007). The highest microbial densities are found in the human gastrointestinal tract (Ley *et al.* 2006), and the identification of 3.3 million different sequences in fecal samples from 124 individuals revealed that the microbiota of the human gut

INTRODUCCIÓN

En los océanos, que son el sistema ecológico más grande del mundo, habitan aproximadamente 3.6×10^{28} microorganismos a una densidad promedio de 5×10^5 células mL^{-1} (Cock *et al.* 2010). Además de abundantes, las comunidades microbianas marinas son altamente diversas debido a que los microorganismos se originaron en un mismo ambiente y han evolucionado adaptaciones fisiológicas exitosas (Hellweger *et al.* 2014).

La importancia y el impacto de las comunidades microbianas en un ambiente específico se han puesto de manifiesto con los estudios del microbioma del cuerpo humano, el cual es un reservorio de aproximadamente 10^9 microorganismos. Los seres humanos pueden contener hasta 10 veces más microbios que células propias, y la comunidad microbiana, por tanto, juega un papel primordial en la salud de las personas (NRC 2007). La mayor densidad microbiana se ubica en el tracto gastrointestinal humano (Ley *et al.* 2006), y la microbiota del intestino de un individuo contiene al menos

contains at least 160 of the 1000–1150 possible bacterial species (Olsen *et al.* 2012).

Oceans are the most extensive environments on Earth and microorganisms are the most predominant in both biomass and metabolic activity. The study of the composition and dynamics of marine microbial communities is therefore of the utmost importance (Fuhrman *et al.* 2015). In oceans, microorganisms capture and convert solar energy, catalyze key biogeochemical transformations of nutrients and trace elements that support productivity, participate in the regulation of greenhouse gases, and represent a vast reservoir of genetic diversity (Karl 2007, Karl and Church 2014). In the aquaculture industry, the microbial communities of bioflocs have become excellent tools for controlling water quality and improving production (Crab *et al.* 2012).

MARINE MICROBIAL COMMUNITIES

Far-reaching projects have been developed to analyze microbial groups and understand the marine environment. In the Sargasso Sea shotgun sequencing project, the random sequencing of microbial cell and virus samples allowed the identification of more than one million protein-coding genes (Venter *et al.* 2004, Sjöstedt *et al.* 2014). This project was followed by the Global Ocean Sampling Expedition, probably the largest sequencing project undertaken to build a metagenomic library, and more than 148 new phylotypes and 69,901 new genes have been identified (Rusch *et al.* 2007, Barberá *et al.* 2012). In the Hawaii Ocean Time-series program, observations carried out over a five-year period at Station ALOHA (10–4000 m depth) allowed the description of new genes (Karl and Church 2014). The International Census of Marine Microbes (ICoMM), one of 14 censuses of marine life supported by several research laboratories, aims to determine the range of genetic diversity and distribution of different microorganisms in the oceans. It comprises 356 datasets of sequences from the V6 region of the bacterial 16S rRNA gene obtained from studies of 40 different biomes, from pelagic and benthic regions to mangrove forests and sponge-associated bacteria (Amaral-Zettler *et al.* 2010, Karsenti *et al.* 2011, Zinger *et al.* 2012). Based on studies of samples collected at 210 diverse oceanic sites and depths of 2000 m, the multinational consortium Tara Oceans has generated the Ocean Microbial Reference Gene Catalog comprising 35,650 taxonomically different organisms (Karsenti *et al.* 2011, Logares *et al.* 2014, Sunagawa *et al.* 2015).

Other studies have searched for a correlation between the genomic properties of 16S of microbiomes and environmental conditions. The results, from different oceanic regions, indicate that temperature is the most important factor determining the composition of microbial communities in the epipelagic zone (Sunagawa *et al.* 2015).

The studies of marine microbial communities can be separated according to the characteristics of the environment they inhabit:

160 de las 1000–1150 especies de bacterias intestinales que pueden existir, según las 3.3 millones de secuencias identificadas en muestras fecales de 124 individuos (Olsen *et al.* 2012).

Los océanos son los ambientes de mayor extensión en el planeta y los microorganismos son la parte dominante tanto en biomasa como en actividad metabólica, por lo que el estudio de la composición y dinámica de las poblaciones microbianas marinas es muy importante (Fuhrman *et al.* 2015). En el océano, los microorganismos son los que capturan y traducen la energía solar, catalizan las transformaciones biogeoquímicas clave de los nutrientes y elementos traza que soportan la productividad oceánica, participan en la regulación de los gases de efecto invernadero y representan una gran reserva de variabilidad genética (Karl 2007, Karl y Church 2014). En la industria de la acuicultura, las comunidades bacterianas de los sistemas de bioflocs se han convertido en excelentes herramientas para el control de la calidad del agua y para mejorar la producción acuícola (Crab *et al.* 2012).

COMUNIDADES MICROBIANAS MARINAS

Para analizar comunidades microbianas marinas y así comprender el ambiente marino se han desarrollado proyectos con amplia cobertura. Por ejemplo, en el proyecto del mar de los Sargazos (*Sargasso Sea shotgun sequencing project*), la secuenciación aleatoria de muestras de virus y células microbianas permitió la identificación de más de un millón de genes codificantes de proteínas (Venter *et al.* 2004, Sjöstedt *et al.* 2014). Una continuación de este trabajo fue la expedición Global Ocean Sampling, posiblemente el proyecto de secuenciación más grande para la construcción de una genoteca metagenómica, que ha permitido la identificación de 148 nuevos filotipos y 69,901 nuevos genes (Rusch *et al.* 2007, Barberá *et al.* 2012). En el proyecto Series de tiempo del Océano en Hawái (Hawaii Ocean Time-series), las observaciones de cinco años desde la estación ALOHA (10–4000 m de profundidad) permitieron la descripción de nuevos genes (Karl y Church 2014). El International Census of Marine Microbes (ICoMM), uno de los 14 censos de la vida marina apoyado por diferentes laboratorios de investigación, tiene como finalidad determinar la distribución y los rangos de diversidad genética de diferentes microorganismos en los océanos. Comprende 356 bases de datos de secuencias de la región V6 del gen rRNA 16S bacteriano recabados de estudios de 40 biomas diferentes, desde regiones pelágicas y bentónicas hasta manglares y comunidades bacterianas asociadas a esponjas (Amaral-Zettler *et al.* 2010, Karsenti *et al.* 2011, Zinger *et al.* 2012). Con base en estudios de muestras de recolectadas de 210 sitios oceánicos distintos y a profundidades de hasta 2000 m, el consorcio multinacional Tara Oceans ha generado un catálogo de genes de referencia (Ocean Microbial Reference Gene Catalog) de 35,650 microbios oceánicos taxonómicamente diferentes (Karsenti *et al.* 2011, Logares *et al.* 2014, Sunagawa *et al.* 2015).

- *Pelagic planktonic microbial communities*. Studies have focused on determining the global patterns of microbial community composition. The findings indicate that the same taxa are present in all the oceans. That observed in one region at a given time are the changes in the relative abundance of community members (Gibbons *et al.* 2013). In addition to qualitative descriptions, studies have also addressed seasonal, interannual, and interseasonal variations (Hatosy *et al.* 2013, Karl and Church 2014), and considerable advances have been made in understanding the metabolic processes of the communities and their role in biogeochemical cycles (Hahnke *et al.* 2013, Brown *et al.* 2014, Karl and Church 2014, Klindworth *et al.* 2014, Logares *et al.* 2014).

- *Coastal planktonic microbial communities*. Regardless of the oceanic influence, changes in the community structure have been attributed to suspended particles from rivers, runoff, and upwelling (Boeuf *et al.* 2013, Ameryk *et al.* 2014, Satinsky *et al.* 2014, Aylward *et al.* 2015, Mueller *et al.* 2015). In these environments it is possible to distinguish communities attached to organic matter from free-living communities (Smith *et al.* 2013, Bižić-Ionescu *et al.* 2014, Mohit *et al.* 2014, Simon *et al.* 2014).

- *Benthic microbial communities*. These communities are very dynamic and the most diverse of the marine communities, most likely because of the high rates of population exchange promoted by the nutrients, environmental conditions, and substrate heterogeneity in the benthic zone (Ramette *et al.* 2009, Miller *et al.* 2013, Gobet *et al.* 2014). They are capable of nitrification and anaerobic ammonium oxidation (Prabavathi and Mathivanan 2012, Wang *et al.* 2012, Laverock *et al.* 2014, Lipsewers *et al.* 2014, Vigneron *et al.* 2014, Bowen *et al.* 2015), which allows the reuse of organic nitrogen.

- *Microbial communities related to submarine volcanoes (fumaroles)*. These communities are representative of extreme environments and do not show a high diversity. The members of the main taxonic groups, however, show greater variability in the 16S sequences, a reflection of the great evolutionary pressure exerted by the environment (Biddle *et al.* 2006, Havelsrud *et al.* 2011, Roussel *et al.* 2011, Tang *et al.* 2013, Shao *et al.* 2014, Kato *et al.* 2015).

16S rRNA

In order to have a standardized tool for the identification of organisms inhabiting different environments, Hebert *et al.* (2003) proposed the so-called DNA barcode for use in the fields of ecology, systematics, and evolutionary biology. The idea was to create a fast, reliable, and reproducible method based on the amplification of a standardized DNA region by

Otros estudios se han dirigido a buscar una correlación entre las propiedades genómicas del 16S de los microbiomas y las condiciones ambientales. Los resultados, provenientes de diferentes regiones oceánicas, indican que la temperatura es el factor más importante en la determinación de la composición microbiana en las capas epipelágicas del océano (Sunagawa *et al.* 2015).

Los estudios de las comunidades microbianas marinas pueden agruparse por las características de los ambientes donde se establecen:

- *Comunidades microbianas planctónicas pelágicas*. Los estudios se han centrado en determinar el patrón global de la composición de la comunidad microbiana. Los resultados indican que en todos los océanos están presentes los mismos taxones, y lo que se observa en una región y tiempo dado son los cambios en las abundancias relativas de sus miembros (Gibbons *et al.* 2013). Más allá de una descripción cualitativa, los estudios han profundizado para determinar los niveles de fluctuación estacional, interanual e interestacional de estas comunidades (Hatosy *et al.* 2013, Karl y Church 2014). Actualmente, se están realizando grandes avances en el conocimiento de los procesos metabólicos que desarrollan las comunidades y su papel en los ciclos biogeoquímicos (Hahnke *et al.* 2013, Brown *et al.* 2014, Karl y Church 2014, Klindworth *et al.* 2014, Logares *et al.* 2014).

- *Comunidades microbianas planctónicas costeras*. Independientemente de la influencia oceánica, estas comunidades muestran cambios en su estructura atribuidos a las partículas en suspensión provenientes de ríos, escurrimientos o surgencias (Boeuf *et al.* 2013, Ameryk *et al.* 2014, Satinsky *et al.* 2014, Aylward *et al.* 2015, Mueller *et al.* 2015). En estos ambientes, es posible diferenciar una comunidad adherida a materia orgánica y otra de vida libre (Smith *et al.* 2013, Bižić-Ionescu *et al.* 2014, Mohit *et al.* 2014, Simon *et al.* 2014).

- *Comunidades microbianas bentónicas*. Estas comunidades son muy dinámicas con la más rica diversidad dentro de las comunidades marinas. Esto posiblemente sea influenciado por los nutrientes, las condiciones ambientales y la heterogeneidad del sustrato que se presentan en el bentos, lo que favorece las altas tasas de recambio de las poblaciones (Ramette *et al.* 2009, Miller *et al.* 2013, Gobet *et al.* 2014). Las comunidades microbianas bentónicas son las responsables de los procesos de nitrificación y oxidación anaeróbica de amonio (Prabavathi y Mathivanan 2012, Wang *et al.* 2012, Laverock *et al.* 2014, Lipsewers *et al.* 2014, Vigneron *et al.* 2014, Bowen *et al.* 2015), lo que permite la reutilización del nitrógeno orgánico.

- *Comunidades microbianas relacionadas con volcanes marinos (fumarolas)*. Estas comunidades representan un

polymerase chain reaction (PCR), and the proposed region was a 600 base-pair fragment of mitochondrial DNA encoding cytochrome *c* oxidase subunit I (COI). The COI region proved to be a valuable tool for the taxonomic classification of many animals and even species, but its use was limited in taxonomic and phylogenetic studies of plants, fungi, and microorganisms (Blaxter 2004, Lebonah *et al.* 2014). It thus became necessary to find other candidate genes or sequences that could be used as markers.

In general, to be considered a molecular marker for barcoding studies and other taxonomic or evolutionary studies, the DNA region must possess the following characteristics: (a) have significant genetic variability and divergence at species level; (b) have conserved sites for designing universal primers for PCR amplification; and (c) have an appropriate length that will allow, with current capabilities, easy, reproducible, and precise extraction and sequencing (Kress and Erickson 2012). Several genes or sequences were proposed as markers, but 16S ribosomal ribonucleic acid (16S rRNA), originally proposed by Pace *et al.* (1986), was considered a good option for the classification of bacteria. It was rapidly adopted by the scientific community and has been used to develop specialized databases. The 16S rRNA sequences have thus become an important tool for reconstructing phylogenetic relationships. Moreover, the use of 16S rRNA sequences led to the establishment of the All-Species Living Tree Project, which has become a reference of relationships among prokaryotes easily organized in dynamic databases that compile and treat the data of all the available 16S rRNA gene sequences (Yarza *et al.* 2008, 2010). Despite some controversies and technical difficulties, the 16S rRNA gene is still used as a molecular marker and, in view of the advantages of novel genomic techniques, new study strategies have been suggested (Savolainen *et al.* 2005, Tanabe and Toju 2013). Owing to the rapid generation of genomic information and the characterization of the 16S rRNA sequences, recent years have seen important changes in the identification of bacterial species and an accelerated assignation of species.

Characteristics of the 16S rRNA gene

The 16S rRNA gene is a polyribonucleotide of approximately 1500 nucleotides, encoded by the gene *rrs*, also called 16S ribosomal DNA. Like any single-chain nucleotide sequence, 16S rRNA folds and acquires a secondary structure, with alternating double-chain and single-chain segments forming approximately 50 helices. This molecule is considered a powerful universal marker because it is found in all known organisms. Its structure is apparently maintained over long periods and as its function has not changed, the changes in the sequence are likely random. In its eukaryotic counterpart, 18S rRNA, mutations are acquired slowly, so it is possible to obtain information about all organisms throughout the evolutionary scale. Nonetheless, rRNA sequences have sufficient variability to differentiate not only the most distant

ambiente extremo y no muestran gran diversidad. Sin embargo, como un reflejo de la gran presión evolutiva ejercida por el medio, los miembros de los principales grupos taxonómicos presentan mayor variación en sus secuencias del 16S (Biddle *et al.* 2006, Havelsrud *et al.* 2011, Roussel *et al.* 2011, Tang *et al.* 2013, Shao *et al.* 2014, Kato *et al.* 2015).

ARNr 16S

Con la intención de tener una herramienta estandarizada para la identificación de los organismos existentes en los diversos ambientes, Hebert *et al.* (2003) propusieron el llamado código de barras del ADN, que proyectaba su utilidad en estudios de sistemática, ecología y biología evolutiva. Estos autores pretendieron generar un método rápido, confiable y reproducible, basado en la amplificación de una región estandarizada del ADN por la reacción en cadena de la polimerasa (PCR, por sus siglas en inglés), y la región propuesta fue un fragmento de 600 pares de bases del ADN mitocondrial, que codifica para la subunidad I del citocromo *c* oxidasa (COI). El uso de la región COI fue excelente herramienta para la clasificación taxonómica de muchos animales, incluso para distinguir entre especies; sin embargo, su utilidad para estudios taxonómicos y/o filogenéticos en plantas, hongos y microorganismos estuvo limitada (Blaxter 2004, Lebonah *et al.* 2014) y fue necesario buscar otras secuencias o genes candidatos que pudieran usarse como marcadores.

En general, para que sea considerada como un marcador molecular para estudios de código de barra y/o en cualquier estudio taxonómico o de evolución, una región de ADN deberá cumplir con las siguientes características: (a) contener una variabilidad y una divergencia genética significativa a nivel de especie; (b) poseer sitios conservados adyacentes, que permitan el diseño de iniciadores universales, para su amplificación por PCR; y (c) tener una longitud adecuada que permita la extracción y secuenciación de forma fácil, reproducible y precisa (Kress y Erickson 2012). Aunque se sugirieron varias regiones o genes, el ácido ribonucléico ribosomal 16S (ARNr 16S), originalmente propuestó por Pace *et al.* (1986), fue presentado como una buena opción para la clasificación de bacterias. La idea fue rápidamente adoptada por la comunidad científica y la secuencia del ARNr 16S se ha utilizado para conformar bases de datos especializadas. Lo anterior ha permitido que las secuencias del ARNr 16S sean utilizadas como una herramienta importante en la reconstrucción de relaciones filogenéticas. Además, el uso de secuencias del ARNr 16S facilitó el establecimiento del proyecto árbol de la vida universal (All-Species Living Tree Project), el cual se ha constituido como una referencia de relación de procariotas fácilmente organizada en bases de datos dinámicas que compilan y curan los datos de todas las secuencias accesibles del gen ARNr 16S (Yarza *et al.* 2008, 2010). Pese a algunas controversias y dificultades técnicas, el ARNr 16S se sigue utilizando como un

organisms but also the nearest, and it is possible to differentiate species, strains, and varieties. Moreover, the relatively long size of 16S rRNA (1500 nucleotides) minimizes statistical fluctuations, while the conservation of its secondary structure allows precise alignment during the comparison of sequences (Rodicio and Mendoza 2004). The 16S rRNA gene contains nine less-conserved or hypervariable regions (V1–V9) (Baker *et al.* 2003), and it is these regions that contribute the most useful information for phylogenetic and taxonomic studies. The conserved regions, on the other hand, are very useful for designing universal primers that allow the amplification of the different hypervariable regions of the 16S rRNA genes of microorganisms found in a community.

While the use of universal primers has undoubtedly favored the detection and analysis of sequences, some authors indicate that they are deficient in detecting a considerable number of uncultured bacterial species from environmental samples (Baker *et al.* 2003, Huws *et al.* 2007). Several studies have reported the coverage of known universal primers and their combinations with sequences obtained from metagenomic studies. For example, an *in silico* analysis revealed that the V4/V5 region was the most accurate for the classification of bacterial sequences from intestinal microbiota, whereas the V7/V8 region was the least accurate (Liu *et al.* 2007, 2008).

Hypervariable regions

With the advent of massive sequencing methods there was an important increase in the number of reports on the characterization of bacterial communities using the 16S rRNA gene as marker. Since only partial sequences are obtained from different variable regions, however, the discrepancies in the results promoted comparative studies between some variable regions and the full-length gene (Nelson 2011, Sun *et al.* 2013).

Diversity indices are one way of evaluating bacterial communities, and the ability of the primer pairs and the region of the 16S rRNA gene that they amplify will strongly influence the description of the bacterial diversity of environmental samples. For example, in an analysis of microbial populations in sediments, Miller *et al.* (2013) observed that the estimated diversity was lower and that the number of operational taxonomic units (OTUs) that could not be classified, even at phylum level, increased from 8.6% to 34.6% when the V3 region was used as taxonomic tool instead of the complete sequence. In a characterization of microbial communities in wastewater, the information obtained with a fragment containing the V1 and V2 regions proved insufficient to detect organisms belonging to the phyla Verrucomicrobia, Planctomycetes, and Chlamydiae (Cai *et al.* 2013). In another study analyzing samples from the human gut and deep-sea vents, differences were observed in the diversity found for the 16S rRNA regions used: 42 taxa were detected in the V3 region and only 26 in the V6 region (Huse *et al.* 2008).

excelente marcador molecular y se han planteado nuevas estrategias de estudio, aprovechando las bondades de las nuevas técnicas genómicas (Savolainen *et al.* 2005, Tanabe y Toju 2013). Debido a la rápida generación de información genómica y a la caracterización de las secuencias del ARNr 16S, en los últimos años se ha observado un cambio significativo en los métodos para la identificación de especies bacterianas y una aceleración en la asignación de especies.

Características del ARNr 16S

El ARNr 16S es un polirribonucleótido de aproximadamente 1500 nucleótidos codificado por el gen *rrs*, también denominado ADN ribosomal 16S. Como cualquier secuencia nucleotídica de cadena sencilla, el ARNr 16S se pliega y adquiere una estructura secundaria que se caracteriza por tener segmentos de doble cadena que permiten la formación de asas y hélices. Esta molécula ha sido reconocida como un poderoso marcador universal debido a que se encuentra en todos los organismos conocidos. Su estructura parece mantenerse por largos periodos de tiempo y, como su función no ha cambiado, los cambios en la secuencia probablemente son aleatorios. En su contraparte eucariota, el ARNr 18S, las mutaciones son adquiridas lentamente, y es posible obtener información acerca de todos los organismos en una escala evolutiva. Sin embargo, los ARNr poseen suficiente variabilidad para diferenciar no sólo los organismos más alejados, sino también los más próximos, y es posible diferenciar especies, cepas o variedades. Además, el tamaño relativamente largo de los ARNr 16S (1500 nucleótidos) minimiza las fluctuaciones estadísticas, y la conservación de su estructura secundaria favorece el alineamiento preciso durante la comparación de secuencias (Rodicio y Mendoza 2004). El ARNr 16S contiene nueve regiones (V1–V9) menos conservadas o hipervariables (Baker *et al.* 2003), que son las que aportan la mayor información útil para estudios de filogenética y taxonomía. Las regiones conservadas son de gran ayuda para diseñar iniciadores universales que permitan la amplificación de las diversas regiones hipervariables de la gran mayoría de los ARNr 16S de los microorganismos presentes en una comunidad.

El uso de los iniciadores universales ha favorecido la detección y análisis de secuencias; sin embargo, algunos autores señalan la deficiencia que tienen para detectar un número considerable de especies bacterianas no cultivadas provenientes de muestras medioambientales (Baker *et al.* 2003, Huws *et al.* 2007). Existen varios trabajos que reportan la cobertura de iniciadores universales y sus combinaciones, con secuencias obtenidas de estudios metagenómicos. Por ejemplo, mediante un análisis *in silico* se determinó que el segmento que incluye las regiones V4/V5 es el más eficiente para la clasificación de la microbiota intestinal, mientras que el segmento que abarca las regiones V7/V8 es el menos eficiente (Liu *et al.* 2007, 2008).

Despite these inconveniences, the 16S rRNA gene continues to be the most commonly used marker to understand bacterial communities in all environments. In view of the differences in primer specificities, sometimes highly specific for a spectrum of bacterial species, some authors recommend the combined use of different primer sets, different DNA extraction methods, and deep sequencing of the bacterial community (Tringe and Hugenholtz 2008, Hong *et al.* 2009, Wang and Qian 2009).

Metagenomics, one of the new "omics" research fields, arose out of the interest of studying bacterial communities of specific environments in an integrated manner. Diverse approaches and methodologies are used to understand the functions of a microbial community. The term *meta* means to transcend and in metagenomics it represents a strategic concept that includes research at three interrelated levels (sample processing, DNA sequencing, and functional analysis) in order to understand the function and importance of genes in a community and how they influence each other's activities in a collective function (NRC 2007). The general study process (shown in fig. 1) and some metagenomic techniques have also proved useful for studying the genomic information of unculturable species. Some authors define these studies as the methodological route for the genomic characterization of microbial communities using culture-independent approaches (Chen and Pachter 2005).

These studies are important because approximately 99% of microorganisms present in a natural environment are not culturable (Amann *et al.* 1995, Curtis *et al.* 2002, Handelsman 2004, Cock *et al.* 2010). The percentage of unculturable organisms varies depending on the environment from which they come and how intensely they have been studied. For example, in the human microbiome, the percentage of unculturable organisms is approximately 70–80% (Nelson 2011), whereas in marine biomes it is higher than 97% (Rappé and Giovannoni 2003). Studies of marine microbial communities would benefit from the application of genomic techniques that prevent the isolation and growth of microbes while generating useful information for their taxonomic identification and classification.

TAXONOMIC ASSIGNMENT

When 16S rRNA is used to study a microbial community, the extraction of DNA and the amplification and subsequent sequencing of the gene or a segment do not represent technical challenges since the general protocols used are well established and highly reproducible. The taxonomic assignment and/or identification of the members, on the other hand, can be challenging not only because of the differences that can occur depending on the primers and regions selected for the study but also because of the conceptual aspects inherent to taxonomic work, the bioinformatics tools, and the availability of reference sequences.

The biological species concept is, to a certain extent, difficult to apply to microorganisms. For many years, the

Las regiones hipervariables

Con la introducción de técnicas de secuenciación masiva, hubo un incremento importante en el número de reportes sobre la caracterización de comunidades bacterianas con el gen ARNr 16S como marcador. Sin embargo, las secuencias son parciales y corresponden a distintas regiones variables. Las discrepancias en los hallazgos fomentó estudios comparativos entre algunas regiones variables y el gen completo (Nelson 2011, Sun *et al.* 2013).

Los índices de diversidad son una forma de evaluar las comunidades bacterianas, por lo que la capacidad del par de iniciadores, y la región del gen ARNr 16S que amplifican, tiene un efecto determinante en la descripción de la diversidad bacteriana de muestras ambientales. Por ejemplo, en un análisis de poblaciones microbianas de sedimentos, Miller *et al.* (2013) utilizaron como herramienta taxonómica la región V3 en lugar de la secuencia completa y observaron que la diversidad estimada fue menor y el número de unidades taxonómicas operativas (OTU, por sus siglas en inglés) que no pudieron ser clasificadas, ni siquiera a nivel de filo, se incrementó de 8.6% a 34.6%. En la caracterización de comunidades microbianas de aguas residuales, la información obtenida con un fragmento que contenía a las regiones V1 y V2 no fue suficiente para detectar organismos de los filos Verrucomicrobia, Planctomycetes y Chlamydiae (Cai *et al.* 2013). En otro estudio, Huse *et al.* (2008) analizaron muestras tan distintas como de intestino humano y chimeneas submarinas y demostraron que cada región del ARNr 16S proporciona diferentes valores de diversidad microbiana: mientras que con el uso de la región V3 registraron 42 taxones, con la V6 solamente encontraron 26. Pese a estos inconvenientes, el uso del ARNr 16S como marcador sigue siendo la herramienta más fuerte para el entendimiento de las comunidades bacterianas de todos los ambientes estudiados. Ante las diferencias de especificidades entre los iniciadores conocidos, a veces altamente específicos para un grupo de bacterias, algunos autores recomiendan el uso combinado de diferentes juegos de iniciadores, diferentes técnicas de extracción de ADN y una secuenciación profunda del material genómico obtenido de la comunidad bacteriana (Tringe y Hugenholtz 2008, Hong *et al.* 2009, Wang Yong y Qian 2009).

En el contexto de los nuevos campos de investigación "ómicas" y del interés de estudiar de forma integral las comunidades bacterianas de ambientes específicos, surgió la metagenómica, que, a través de diversos enfoques y metodologías, permite la comprensión de las funciones de una comunidad microbiana. El término *meta* significa trascender y en el concepto *metagenómica* es estratégico porque este campo de la ciencia incluye investigación a tres niveles interrelacionados (procesamiento de muestras, secuenciación de ADN y análisis funcional) para comprender la función e importancia de los genes en una comunidad y analizar su influencia en las actividades de otros genes al realizar una función colectiva (NRC 2007). Además, el proceso general de estudio (ver

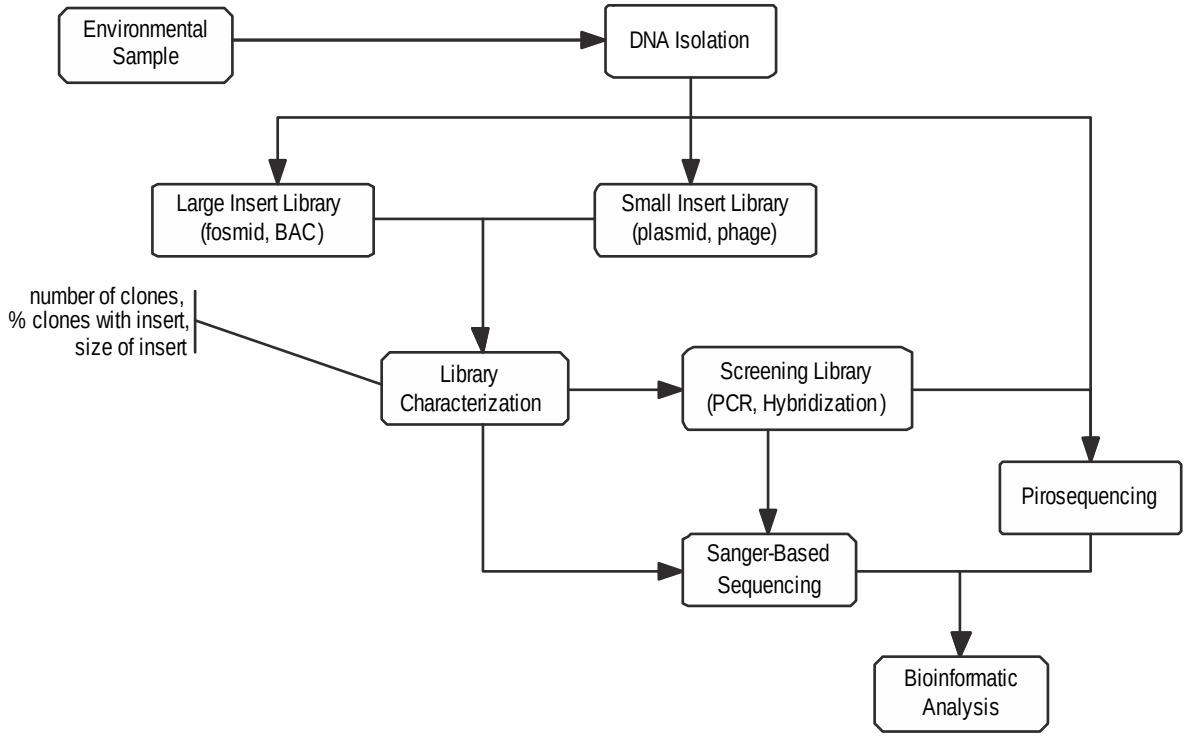

Figure 1. Studies of the 16S rRNA gene follow similar steps to those in metagenomic studies, as shown in the figure modified from Cock *et al.* (2010), except that the cloning process includes only small-insert libraries.
Figura 1. Los estudios del gen ARNr 16S comprenden pasos semejantes a los estudios de metagenómica, como se puede visualizar en la figura modificada de Cock *et al.* (2010), con la diferencia de que el proceso de clonación sólo comprende genotecas de insertos cortos.

assignment of taxonomic categories (species and/or strain) was based on distinct biochemical and/or antigenic characteristics, and at best, on physiological features detectable by chemical or biochemical analyses. With the advent of genomic technologies, some chemical and physical properties of nucleic acids could be considered and relatively short fragment sequences or a pair of genes could subsequently be compared. New high-resolution techniques that allow massive analyses propose a species concept based on more than the similarity of a nucleotide segment or of some genes. While it is true that these new tools enhance the accuracy and contribution of large amounts of information, they also introduce greater complexity and the boundaries between species and strains are still not accurately defined. A prokaryotic species is described as "a monophyletic and genomically coherent cluster of individual organisms that show a high degree of overall similarity in many independent characteristics" (Rosselló-Mora and Amann 2001).

The first genotypic method systematically used in taxonomic or identification studies consisted in determining the guanine and cytosine (G + C) content. It is currently relatively easy to estimate the composition of a whole genome and it has been observed that, in nearly all cases, the difference in G + C content between strains does not exceed 1%; higher values are indicative of different species (Rosselló-Mora and Amann 2001, Meier-Kolthoff *et al.* 2014). Over the fig. 1), y algunas técnicas metagenómicas han sido útiles para estudiar y conocer la información genómica de especies que no son cultivables. Por ello, algunos autores definen estos estudios como la ruta metodológica para la caracterización genómica de comunidades microbianas independientes de cultivos (Chen y Pachter 2005).

La importancia de estos estudios recae en que aproximadamente el 99% de los microorganismos en un ambiente natural no son cultivables (Amann *et al.* 1995, Curtis *et al.* 2002, Handelsman 2004, Cock *et al.* 2010). El porcentaje de organismos no cultivables varía dependiendo del ambiente de donde provienen y de la intensidad con que se han estudiado. Por ejemplo, el porcentaje de organismos no cultivables del microbioma humano es del ~70–80% (Nelson 2011), mientras que el porcentaje para los biomas marinos es mayor que el 97% (Rappé y Giovannoni 2003). El estudio de las comunidades microbianas marinas puede beneficiarse de la aplicación de técnicas genómicas que eviten el aislamiento y crecimiento de los microbios y que generen información útil para la clasificación e identificación taxonómica.

ASIGNACIÓN DE TAXONES

Cuando se utiliza el ARNr 16S para esudiar una comunidad microbiana, la extracción del ADN, la amplificación del gen o un segmento y la correspondiente secuenciación no

past 50 years, the DNA-DNA hybridization (DDH) technique proved to be a valuable tool for comparing prokaryotes and measuring the degree of similarity between two genomes. It is still used to show the degree of hybridization and similarity between probable members in those cases where one taxon has more than one strain (Tindall *et al.* 2010). The DDH technique is also recommended for determining closeness when strains have more than 97% 16S rRNA sequence similarity. Initially, a 16S rRNA sequence similarity equal to or greater than 97% indicated that the strains belonged to the same species (Stackebrandt and Ebers 2006, Vos 2011). A more recent study, supported by cross-validation tests, revealed that 98.65% 16S rRNA gene sequence similarity can be used as the threshold for differentiating two species (Kim *et al.* 2014). It has also been established that when comparing two 16S rRNA, a similarity equal to or less than 94.5%, 86.5%, 82.0%, 78.5% and 75% distinguishes genus, family, order, class, and phylum, respectively (Yarza *et al.* 2014).

The use of sequences of other genes has also been proposed for taxonomic assignment; however, less than 100 genes have the distribution and variability that make them suitable for use as markers (Koonin 2003). The average nucleotide identity (ANI) represents a mean of identity/similarity values between homologous genomic regions (Konstantinidis and Tiedje 2005). ANI values of 95–96% correspond to a DDH value of 70% and they are the minimum values for two organisms to belong to the same species (Kim *et al.* 2014). This measure can be supported by multilocus sequence analysis (MLSA), though it is not applicable to uncultured organisms (Schleifer 2009). Ribosomal multilocus sequence typing (rMLST), a variation of MLSA, analyzes 53 genes encoding ribosomal proteins, even though some of these genes have not been detected in some bacterial genomes (Larsen *et al.* 2014). On the other hand, the Genome BLAST Distance Phylogeny (GBDP) strategy considers the whole genome sequence and a global comparison is made using alignment tools such as BLAT or BLAST (Basic Local Alignment Search Tool) (Chun and Rainey 2014). Another genomic approach for taxonomic assignment is the characterization of DNA fragments and their variations, which can be restriction sites, insertions and deletions, DNA sequence repeats or microsatellites, single nucleotide polymorphisms or other sequence variations (Moore *et al.* 2010). Amplified fragment length polymorphism (AFLP) has also proven useful for the delineation of new species (Nemec *et al.* 2001). In order to apply as much information as possible to obtain the most accurate characterization possible, taxonomic assignment protocols using other molecules and a polyphasic approach, including MALD-TOF mass spectrometry (Ramasamy *et al.* 2014), have also been proposed (Moore *et al.* 2010). The most commonly used techniques and the level of taxonomic resolution are shown in figure 2.

representan retos técnicos, pues los protocolos generales empleados están muy bien establecidos y son altamente reproducibles. La parte central del proceso es la asignación de taxones y/o identificación de los miembros, no solamente por las diferencias que se pueden obtener dependiendo de los iniciadores seleccionados o de las regiones comprendidas en el estudio, sino también por los aspectos conceptuales inherentes al trabajo taxonómico, a las herramientas bioinformáticas y a la disponibilidad de secuencias de referencia.

El concepto biológico de especie es, hasta cierto punto, problemático de aplicar a los microorganismos. Durante muchos años, la asignación de las categorías taxonómicas (especie y/o cepa) de microorganismos estuvo basada en la distinción de características bioquímicas y/o antigénicas y, en el mejor de los casos, en aspectos fisiológicos detectables por análisis químicos o bioquímicos. Con el desarrollo de las técnicas genómicas, primero se tomaron en cuenta algunas características químicas y fisicoquímicas de los ácidos nucleicos y, posteriormente, se llegó a la comparación de secuencias de fragmentos relativamente cortos o un par de genes. Las nuevas tecnologías de secuenciación con alta resolución y que permiten análisis masivos proponen un concepto de especie que va más allá de la similitud de un segmento nucleotídico o de algunos genes. Si bien es cierto que estas nuevas herramientas incrementan la precisión y el aporte de información, también introducen mayor complejidad, y la frontera delimitante entre especie y cepa continúa sin definirse con precisión. Actualmente, se considera que una especie procariota es un grupo genómicamente coherente de individuos/cepas, que comparten un alto grado de similitud en características independientes (Rosselló-Mora y Amann 2001).

La primera metodología genotípica empleada de manera sistemática para estudios taxonómicos o de identificación fue la determinación del contenido de los nucleótidos guanina y citosina (G+C). Hoy en día es relativamente simple estimar la composición de un genoma completo y se ha observado que, en casi todos los casos, la diferencia de contenido G+C entre cepas no rebasa el 1%; los valores superiores al 1% son indicativos de diferentes especies (Rosselló-Mora y Amann 2001, Meier-Kolthoff *et al.* 2013). Durante los últimos 50 años, la hibridación ADN–ADN (DDH, por sus siglas en inglés) ha sido considerada una herramienta de oro para la comparación de procariotas, y parece ser muy útil para medir directamente el grado de similitud entre dos genomas. En la actualidad, en aquellos casos donde un taxón contiene más de una cepa, se recomienda la aplicación de esta técnica para mostrar el grado de hibridación y la similitud entre los probables miembros (Tindall *et al.* 2010). El uso de DDH también se recomienda para demostrar cercanía cuando las cepas tienen más del 97% de similitud en sus secuencias de ARNr 16S. Inicialmente, se consideraba que una similitud igual o mayor que el 97% entre las secuencias de ARNr 16S indicaba pertenencia a la misma especie (Stackebrandt y Ebers 2006, Vos 2011). Un estudio más reciente, apoyado por pruebas

DATABASES AND ANALYSIS STRATEGIES

Taxonomic assignments are made by comparing sequences, and the databases that store them, the search tools, and the comparison strategies used are key to achieving this goal. Several databases and programs are available. The programs have different analysis strategies and the databases they operate show some variations. In addition to the general databases, given the progress of 16S rRNA gene sequencing methods, specific databases have been established that are fundamental tools for the taxonomic classification of microorganisms (Santamaria *et al.* 2012, Kim *et al.* 2013, Selama *et al.* 2013, Chun and Rainey 2014).

General databases

The National Center for Biotechnology Information (NCBI) in the United States is the institution that has the highest number of sequences (more than 182 million) deposited in its GenBank database (fig. 3). Sequences of the 16S rRNA gene are stored in two nucleotide sequence databases: the non-redundant nucleotide (nr/nt) database containing more than 30 million sequences and the 16S database, which is dedicated to 16S rRNA sequences of identified bacteria and archaea and currently contains more than 17,600 sequences (for the most part whole). While both databases can serve as reference for searching and comparing sequences using the BLAST algorithm, the nr/nt database is broader and contains the GenBank, European Bioinformatics Institute (EMBL-EBI), DNA Databank of Japan (DDBJ), and Protein Data Bank (PDB) sequence collections, as well as the NCBI Reference Sequence (RefSeq) database. It is a non-redundant database because in some cases identical sequences have been merged into one entry, while preserving the accession number, GenBank identifier, title, and taxonomy information for each sequence (http://www.ncbi.nlm.nih.gov/guide/all/#databases).

Through the International Nucleotide Sequence Database Collaboration (INSDC), all the GenBank information can be found in the EMBL-EBI and DDBJ databases. INSDC is a long-standing initiative between DDBJ, EMBL-EPI, and NCBI for the maintenance of sequence databases and covers the spectrum of data, including raw reads, assemblies and alignments to functional annotations and contextual information relating to the samples (Karsch-Mizrachi *et al.* 2012).

Specific databases

While the 16S rRNA gene was originally used to identify bacteria, today it is used as a standard for the classification and identification of microorganisms and the sequences are readily available in public databases. As these sequences often lack validation, databases that collect only 16S sequences were created. Some databases, such as the Greengenes database (http://greengenes.lbl.gov/), have not

estadísticas de validación cruzada, reveló que un valor del 98.65% de similitud del gen ARNr 16S puede ser utilizado como umbral para la diferenciación de dos especies (Kim *et al.* 2014). Además, se ha establecido que una similitud entre dos ARNr 16S igual o menor que el 94.5%, 86.5%, 82.0%, 78.5% ó 75.0% establece la distinción de género, familia, orden, clase y filo, respectivamente (Yarza *et al.* 2014).

También se ha propuesto el uso de secuencias de otros genes para la asignación taxonómica, pero hay menos de 100 genes con la distribución y la variabilidad adecuada para ser utilizados como marcadores (Koonin 2003). El índice de nucleótidos promedio (ANI) representa una media de valores de identidad o similitud entre regiones genómicas homólogas (Konstantinidis y Tiedje 2005). Un valor ANI del 95–96% es comparado con un valor DDH del 70% y son los valores mínimos para considerar dos organismos dentro de la misma especie (Kim *et al.* 2014). La medición del ANI puede ser apoyada por el análisis de secuencias multilocus (MLSA: *multilocus sequence analysis*), aunque no es aplicable para organismos no cultivados (Schleifer 2009). Una variación del MLSA es la tipificación por secuencias multilocus ribosomales (rMLST: *ribosomal multilocus sequence typing*) que analiza 53 genes codificantes para proteínas ribosomales, aunque algunos de estos genes no se han detectado en algunos genomas bacterianos (Larsen *et al.* 2014). Por otro lado, la determinación de la distancia filogénica por comparación de genomas (GBDP: Genome BLAST Distance Phylogeny) es una estrategia que considera la secuencia genómica completa y se realiza una comparación global mediante un alineamiento local con herramientas como BLAST o BLAT (Chun y Rainey 2014). Otra estrategia genómica para la asignación taxonómica es la caracterización de fragmentos de ADN y sus variaciones, que pueden ser sitios de restricción, inserciones, supresiones, secuencias repetidas o microsatélites, polimorfismos de una sola base u otras diferencias de las secuencias (Moore *et al.* 2010). La detección de polimorfismo del tamaño de fragmentos amplificados (AFLP: *amplified fragment length polymorphism*) ha mostrado ser útil para la delineación de nuevas especies (Nemec *et al.* 2001). Con la intención de utilizar la mayor información posible e incrementar la precisión en la caracterización, también se han propuesto protocolos de asignación de taxones con otras moléculas y con un enfoque polifásico (Moore *et al.* 2010) mediante espectrometría de masas (Ramasamy *et al.* 2014). Las técnicas más utilizadas para la asignación taxonómica y el nivel de resolución que permiten se presentan en la figura 2.

BASES DE DATOS Y ESTRATEGIAS DE ANÁLISIS

Debido a que la asignación taxonómica se realiza por comparación de secuencias, las bases de datos donde se almacenan estas secuencias, las herramientas de búsqueda y las estrategias de comparación son primordiales. Existen distintas bases de datos y programas que se pueden utilizar

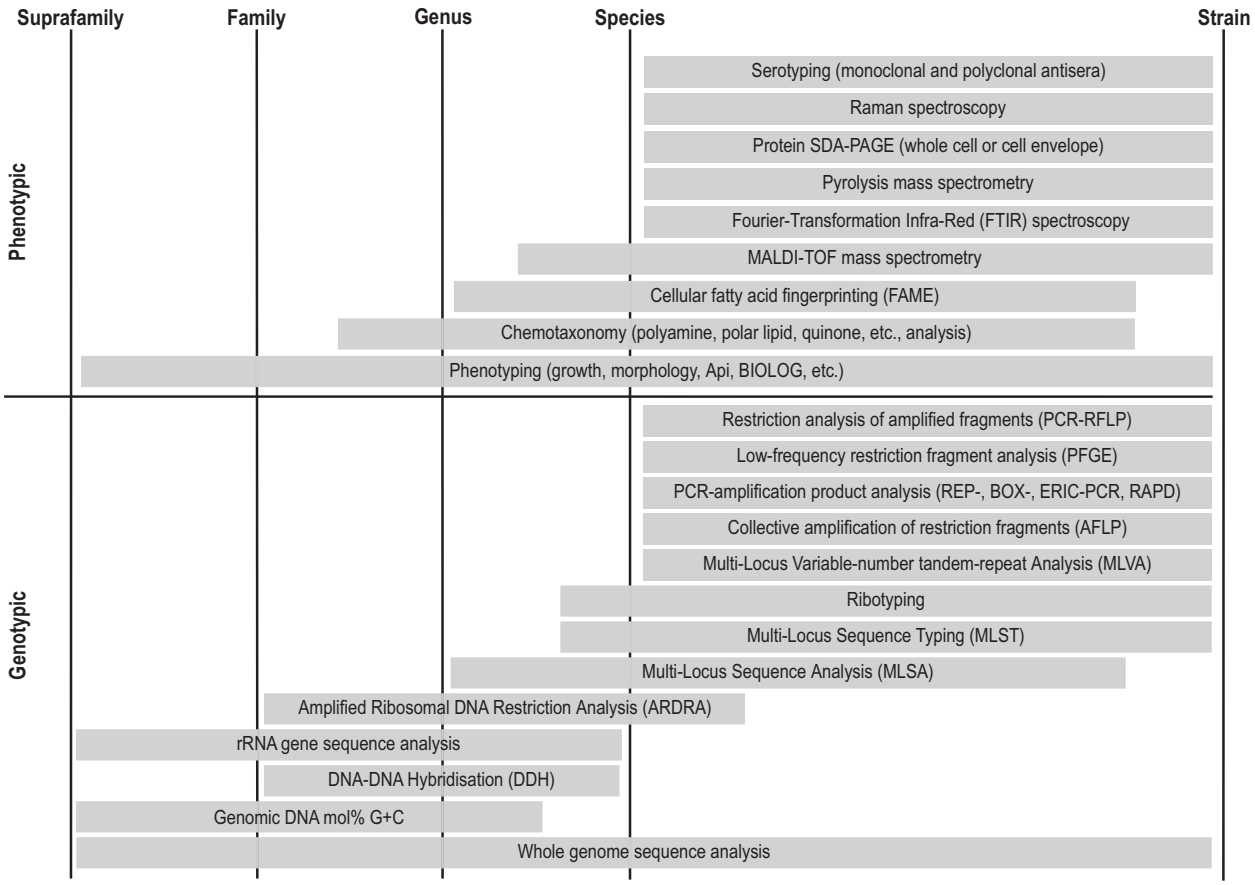

Figure 2. Methodology for the characterization of prokaryotes and the different levels of taxonomic resolution (modified from Moore *et al.* 2010).

Figura 2. Metodología para la caracterización de procariotes y los distintos niveles de resolución taxonómica (modificada de Moore *et al.* 2010).

been updated but can still be accessed. Below we mention the most important databases.

SILVA: High quality ribosomal RNA databases

The SILVA website (http://www.arb-silva.de) provides databases of aligned small subunit (16S/18S, SSU) and large subunit (23S/28S, LSU) rRNA sequences for the Bacteria, Archaea, and Eukarya domains. In the case of SSU, the information is found in three datasets that differ by the size and quality of the sequences: (1) SSU Parc contains more than 4.3 million sequences larger than 300 nucleotides; (2) SSU Ref contains more than 1.5 million high-quality, nearly full-length sequences (1450 nucleotides); and (3) SSU Ref NR contains more than 534,000 non-redundant sequences, which are basically the same as in SSU Ref but have <99% similarity. All sequences are associated with their GenBank accession number, taxonomy, multiple sequence alignment, type of strain, and last valid nomenclature (Yilmaz *et al.* 2014).

In addition to the datasets, SILVA also developed an alignment tool called SINA (SILVA Incremental Aligner), para la asignación taxonómica; cada programa sigue diferentes estrategias de análisis y las bases de datos con las que operan muestran algunas variaciones. El nivel de desarrollo que han alcanzado los estudios sobre el ARNr 16S y sus aplicaciones, ha favorecido el establecimiento de bases de datos específicas, que son herramientas fundamentales para la clasificación taxonómica microbiana (Santamaria *et al.* 2012, Kim *et al.* 2013, Selama *et al.* 2013, Chun y Rainey 2014).

Bases de datos generales

El Centro Nacional para la Información Biotecnológica (NCBI) de los Estados Unidos es la institución con mayor número de secuencias depositadas en sus bases de datos, el GenBank, con más de 182 millones de registros (fig. 3). Las secuencias del ARNr 16S están localizadas en dos bases de datos con secuencias de nucleótidos: la base de datos no redundante, que contiene más de 30 millones de registros, y la base de datos 16S, que contiene solamente secuencias (17,600, completas en su mayoría) del ARNr 16S de bacterias y arqueas identificadas. Si bien ambas bases de datos

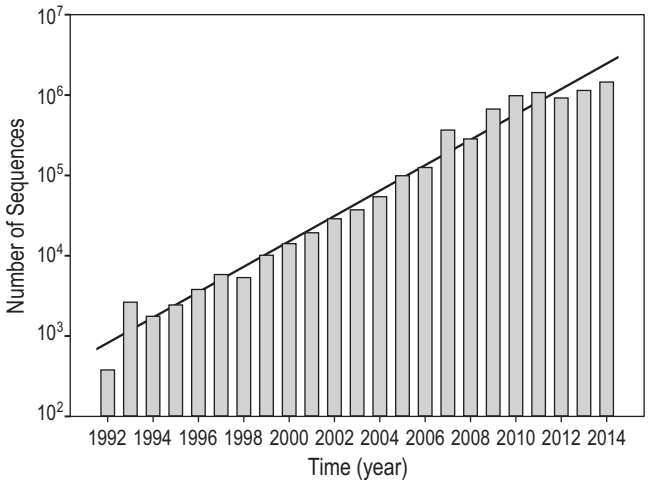

Figure 3. Exponential growth of the number of 16S rRNA sequences registered in the GenBank database managed by the National Center for Biotechnology Information (NCBI, USA).

Figura 3. Crecimiento exponencial del número de secuencias ARNr 16S registradas en el GenBank del Centro Nacional para la Información Biotecnológica (NCBI) de los Estados Unidos.

which can align hundred thousands of sequences based on a curated seed alignment. SINA uses a combination of *k*-mer searching and partial order alignment to maintain high alignment accuracy (Pruesse *et al.* 2012). After the alignment, taxonomic classification is possible using the lowest common ancestor (LCA) method (Clemente *et al.* 2011), based on different taxonomies provided by SILVA (SILVA, LTP, Greengenes, RDP, EMBL).

Ribosomal Database Project

The Ribosomal Database Project (RDP) at Michigan State University provides ribosome-related data services, including online data analysis, rRNA-derived phylogenetic trees, and aligned and annotated rRNA sequences (https://rdp.cme.msu.edu/index.jsp). This database contains more than three million sequences of both culturable and unculturable bacteria and archaea (Cole *et al.* 2005, 2014). The sequences are aligned by an alignment program, called INFERNAL (INFERence of RNA ALignment), based on stochastic context that incorporates secondary structure information (Nawrocki *et al.* 2009, Nawrocki and Eddy 2013). For the taxonomic identification, this database uses the RDP Classifier tool, based on a simple Bayesian algorithm (Vilo and Dong 2012).

EzTaxon

The EzTaxon database (http://www.ezbiocloud.net/eztaxon) originally included only 16S sequences of type strains of identified bacterial species (Chun *et al.* 2007). It was later expanded to include sequences of uncultured pueden servir como referencia para la búsqueda y comparación de secuencias nucleotídicas, mediante el algoritmo BLAST, la base de datos no redundante es más amplia y contiene las colecciones de secuencias del GenBank, del Instituto Europeo de Bioinformática (EMBL-EBI), del Banco de Datos de ADN de Japón (DDBJ), y del Banco de Datos de Proteínas (PDB), además de las secuencias de referencia del NCBI (RefSeq). Se le considera no redundante porque en algunos casos las secuencias idénticas han sido fusionadas en una sola entrada, que conserva el número de acceso, el identificador del GenBank, título e información taxonómica para cada secuencia (http://www.ncbi.nlm.nih.gov/guide/all/#databases).

A través del International Nucleotide Sequence Database Collaboration (INSDC), la información del GenBank también se puede encontrar en las bases de datos de EMBL-EBI y DDBJ. La INSDC es una iniciativa de larga duración entre DDBJ, EMBL-EBI y NCBI para mantener actualizadas y accesibles las bases de datos de secuencias y la lectura de datos crudos, así como los alineamientos y ensambles para anotaciones funcionales y con información contextual relacionada a las muestras (Karsch-Mizrachi *et al.* 2012).

Bases de datos específicas

Aunque el gen ARNr 16S originalmente se utilizó para la identificación de bacterias, hoy en día es utilizado como un estándar en la clasificación e identificación de microorganismos y las secuencias están disponibles en muchas bases de datos públicas. Sin embargo, con frecuencia esas secuencias no han sido validadas y por ello se han creado varias bases de datos que recolectan únicamente secuencias 16S. Algunas bases de datos, como la de Greengenes (http://greengenes.lbl.gov/), no se encuentran actualizadas pero permiten el acceso a y descarga de las secuencias registradas. A continuación, se mencionan las bases de datos más importantes.

SILVA: Bases de datos de ARN ribosomal de alta calidad

SILVA es un portal (http://www.arb-silva.de) que alberga secuencias del ARNr tanto de 16S/18S (subunidad pequeña, SSU) como de 23S/28S (subunidad grande) para los dominios Bacteria, Archaea y Eukarya. Para la SSU, la información se encuentra en tres bases de datos que se diferencian por el tamaño y la calidad de las secuencias: (1) SSU Parc contiene más de 4.3 millones de secuencias con tamaño mayor que 300 nucleótidos; (2) SSU Ref contiene más de 1.5 millones de secuencias de buena calidad y con longitud cercana a la del gen completo (1450 nucleótidos); y (3) SSU Ref NR contiene más de 534,000 secuencias no redundantes, que básicamente son las mismas de la base de datos SSU Ref, pero con similitud por debajo del 99%. Todas las secuencias están asociadas a su registro en GenBank, taxonomía, alineamiento de secuencias múltiples, información del tipo de cepa y última nomenclatura válida (Yilmaz *et al.* 2014).

species, which are very frequent in environmental studies (Kim *et al.* 2012). This database currently contains more than 62,600 species and phylotypes. BLAST and MEGABLAST searches are followed by rigorous global pair-wise sequence alignment (Myers and Miller 1988).

Computer programs

MEGAN allows the analysis of metagenomic sequences using the BLAST comparison tool and the LCA algorithm to assign taxons and the NCBI taxonomy as reference (Huson *et al.* 2007).

PhymmBL was designed for the taxonomic classification of short metagenomic reads. Basically, this program combines composition-based (Phymm) and homology-based (BLAST) taxonomic predictions (Brady and Salzberg 2009).

RITA searches for an agreement between homology-based and composition-based approaches, which considerably increases the accuracy of taxonomic assignments. This program uses a naive Bayesian classifier and optimized homology detection algorithms, and is thus faster than PhymmBL and other programs (MacDonald *et al.* 2012).

Kraken is an accurate program for assigning taxonomic lables to metagenomic sequences based on the exact alignment of *k*-mers, making it a faster tool than those that use inexact alignment of sequences. It also provides greater taxonomic richness because it uses *k*-mers of 31 base pairs, which allows high sensitivity among sequences with a high degree of similarity. The sensitivity and precision achieved depend on the database used by the program to make the comparisons (Wood and Salzberg 2014)**.**

The RDP Classifier, a naive Bayesian classifier, provides taxonomic assignments of bacteria from domain to genus. It uses eight-base subsequences (words) to achieve a balance between the sensitivity and speed of analysis or computer requirements. The position of the word is ignored and only the words found contribute to the classification. This program is useful for analyzing long sequences but does not discriminate well when using short sequences (Wang *et al.* 2007).

TRENDS

Multiple strategies exist for the use of 16S rRNA as a tool in microbial identification; however, the accurate classification of microorganisms is a complex task. The identification of organisms from a community, using either the complete sequence or a region of the 16S rRNA gene, is made easier if several of its members have previously been placed in a taxonomic arrangement. In environments such as the marine environment, where the microbial life has not yet been well documented, the task of identification is hampered by the absence of reference sequences. The advent of these novel sequencing techniques allowed many studies aiming to characterize little-known microbial communities to generate

Además de las bases de datos, SILVA dispone del programa SILVA Incremental Aligner (SINA), que permite alinear cientos de secuencias sobre la base de alineamiento semilla utilizando una combinación de búsqueda por *k*-meros y un alineamiento de orden parcial para mantener alta exactitud de alineación (Pruesse *et al.* 2012). Posterior al alineamiento, es posible clasificar taxonómicamente a través del método del ancestro común más bajo (LCA) (Clemente *et al.* 2011), basado en diferentes taxonomías alojadas por SILVA (SILVA, LTP, Greengenes, RDP, EMBL).

Ribosomal Database Project

El Ribosomal Database Project (RDP) es un portal (https://rdp.cme.msu.edu/index.jsp) de la Universidad Estatal de Michigan (EUA) que, además de contener secuencias de ARNr 16S alineadas y con anotaciones, proporciona servicios de análisis y un marco de referencia taxonómico y filogenético. En esta base de datos se encuentran más de tres millones de secuencias de bacterias y arqueas, tanto cultivables como no cultivables (Cole *et al.* 2005, Cole *et al.* 2014). Las secuencias son alineadas por medio de un programa de alineamientos basados en el contexto estocástico (INFERNAL: INFERence of RNA ALignment) que incorpora información sobre la estructura secundaria del ARNr 16S (Nawrocki *et al.* 2009, Nawrocki y Eddy 2013). Para la identificación taxonómica, esta base de datos utiliza la herramienta RDP Classifier, la cual se basa en un algoritmo Bayesiano sencillo (Vilo y Dong 2012).

EzTaxon

La base de datos EzTaxon (http://www.ezbiocloud.net/eztaxon) originalmente incluía solamente las secuencias de 16S de cepas bacterianas tipo de especies identificadas (Chun *et al.* 2007). Posteriormente se amplió al incluir el registro de secuencias provenientes de especies no cultivadas, las cuales son muy frecuentes en los estudios medioambientales (Kim *et al.* 2012). Actualmente, esta base de datos contiene más de 64,000 especies y filotipos, y para la búsqueda de secuencias similares se utiliza BLAST y MEGABLAST, además de un riguroso alineamiento de secuencias globales por pares (Myers y Miller 1988).

Programas computacionales

MEGAN es un programa que permite el análisis de secuencias metagenómicas mediante la herramienta de comparación BLAST y el algoritmo LCA para la asignación de taxones, y utiliza como referencia la taxonomía del NCBI (Huson *et al.* 2007).

PhymmBL es un programa diseñado para la clasificación taxonómica de lecturas cortas de secuencias metagenómicas. Básicamente, el programa combina las predicciones taxonómicas basadas en composición (Phymm) y aquellas basadas en la homología (BLAST) (Brady y Salzberg 2009).

genetic information, but other limitations then arose associated with the storage, processing, analysis, and interpretation of the large amount of raw data produced. Fortunately, all these limitations have been solved with the development of computational tools and of algorithms and programs for accurate microbial identification.

Massive sequencing has produced a large amount of partial sequences but efforts to obtain a complete sequence have not yet produced the desired results (Miller *et al.* 2011, 2013). Many studies that use partial sequences to identify microorganisms do not find a reference sequence and the taxonomic certainty is reduced. In some cases, the result can be an overestimation of the diversity. In an analysis of a simulated community, built by the 16S rRNA sequences of soil bacteria belonging to 20 OTUs and using fragments that included the V4/V5 and V2/V3 regions, after appropriate selection and removal of chimera, from 72 to 333 OTUs were found, depending on the experimental conditions during the sequencing process (Jeon *et al.* 2015). Hence, to maintain parallelism in the development of "omics" techniques and the exponential growth of 16S rRNA databases, and particularly in the study of marine bacterial communities (Giovannoni *et al.* 2005, Pedros-Alio 2006, Kämpfer and Glaeser 2012), complete sequences of this gene are needed in order to solve with greater precision and reproducibility the different taxonomic assignments.

English translation by Christine Harris.

REFERENCES

Amann, RI, Ludwig, W, Schleifer, KH 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Mol. Biol. Rev. 59: 143–169.

Amaral-Zettler L, Artigas LF, Baross J, Bharathi L, Boetius A, Chandramohan D, Herndl G, Kogure K, Neal P, Pedros-Alio C, Ramette A, Schouten S, Stal L, Thessen A, de Leeuw J, Sogin M. 2010. A global census of marine microbes. In: McIntyre A (ed.), Life in the World's Oceans: Diversity, Distribution and Abundance. Blackwell Publishing, Oxford.

Ameryk A, Hahnke RL, Gromisz S, Kownacka J, Zalewski M, Szymanek L, Całkiewicz J, Dunalska J, Harder J. 2014. Bacterial community structure influenced by *Coscinodiscus* sp. in the Vistula river plume. Oceanologia 56: 825–856. http://dx.doi.org/10.5697/oc.56-4.825

Aylward FO, Eppley JM, Smith JM, Chavez FP, Scholin CA, DeLong EF. 2015. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. Proc. Natl. Acad. Sci. USA 112: 5443–5448. http://dx.doi.org/10.1073/pnas.1502883112

Baker GC, Smith JJ, Cowan DA. 2003. Review and re-analysis of domain-specific 16S primers. J. Microbiol. Methods 55: 541–555. http://dx.doi.org/10.1016/j.mimet.2003.08.009

Barberá A, Fernádez-Guerra A, Bohannan BJM, Casamayor EO. 2012. Exploration of community traits as ecological markers in microbial metagenomes. Mol. Ecol. 21: 1909–1917. http://dx.doi.org/10.1111/j.1365-294X.2011.05383.x

RITA es un programa que realiza un consenso entre los enfoques basados en homología y aquellos basados en composición, lo cual incrementa de forma importante la certeza en las asignaciones taxonómicas. RITA utiliza un clasificador Bayesiano ingenuo e involucra a algoritmos optimizados de detección de homología, por lo que es más rápido que PhymmBL y otros programas (MacDonald *et al.* 2012).

Kraken es un programa para la asignación de niveles taxonómicos de secuencias metagenómicas que utiliza el alineamiento exacto de la secuencia problema con la base de datos de *k*-meros, lo cual lo convierte en una herramienta rápida en comparación con las que realizan el alineamiento inexacto. Además, provee una mayor riqueza taxonómica, ya que el utilizar *k*-meros de 31 pares de bases le permite una alta sensibilidad entre secuencias con un alto grado de similitud. La sensibilidad y precisión se modifican dependiendo de la base de datos que utiliza el programa para hacer las comparaciones (Wood y Salzberg 2014).

RDP Classifier es un clasificador Bayesiano ingenuo que permite la rápida clasificación de bacterias con asignaciones taxonómicas de dominio a género. Utiliza "palabras" de ocho bases subsecuentes, logrando un equilibrio entre la sensibilidad y la velocidad de análisis o requerimientos computacionales. La posición de la palabra es ignorada y solamente las palabras que se encuentran contribuyen en la calificación. Este programa es útil para analizar secuencias largas y no discrimina bien al utilizar secuencias cortas (Wang *et al.* 2007).

TENDENCIAS

Existen múltiples estrategias para el uso del ARNr 16S como herramienta de identificación de los organismos presentes en comunidades microbianas; sin embargo, se hace patente que la clasificación precisa de los microorganismos es una tarea compleja. Más aún, la identificación de organismos de una comunidad, ya sea que se utilice la secuencia completa o una región del ARNr 16S, se facilita si una buena proporción de sus miembros ha sido previamente localizada dentro de un arreglo taxonómico. En ambientes como el marino, donde la vida microbiológica ha sido poco revelada, la tarea de identificación se ve dificultada por la ausencia de secuencias de referencia. Muchas investigaciones centradas en la caracterización de comunidades microbianas de ambientes poco conocidos encontraron en las nuevas técnicas de secuenciación una opción para generar información génica, pero se enfrentaron con otras limitantes inherentes a la gran extensión de datos crudos producidos en la secuenciación, como almacenamiento, procesamiento, análisis e interpretación. Afortunadamente, todas las limitantes se han ido resolviendo gracias al desarrollo de la infraestructura computacional y el desarrollo de algoritmos y programas con la precisión suficiente para la identificación microbiana.

Con la secuenciación masiva se ha generado una gran cantidad de secuencias parciales y los esfuerzos para obtener una secuencia completa aún no dan los resultados deseados

(Miller *et al.* 2011, 2013). Muchos estudios que utilizan secuencias parciales para la identificación microbiológica no encuentran una secuencia de referencia, y reducen la certeza de las determinaciones taxonómicas. En otros casos, el resultado puede ser una sobreestimación de la diversidad. En el análisis de una comunidad simulada, construida por las secuencias del ARNr 16S de bacterias del suelo pertenecientes a 20 OTUs y usando fragmentos que incluían a las regiones V4/V5 y V2/V3, después de un proceso adecuado de selección de calidad y eliminación de quimeras, se encontraron de 72 a 333 OTUs, dependiendo de las condiciones experimentales durante la secuenciación (Jeon *et al.* 2015). Por ello, para mantener el paralelismo con el desarrollo de las técnicas "ómicas" y el incremento exponencial de las bases de datos del ARNr 16S y, particularmente en el estudio de las comunidades bacterianas marinas (Giovannoni *et al.* 2005, Pedros-Alio 2006, Kämpfer y Glaeser 2012), es necesario contar con un mayor número de secuencias completas de este gen para resolver con mayor precisión y reproducibilidad las distintas asignaciones taxonómicas.

Biddle JF, Lipp JS, Lever MA, Lloyd KG, Sørensen KB, Anderson R, Fredricks HF, Elvert M, Kelly TJ, Schrag DP, Sogin ML, Brenchley JE, Teske A, House CH, Hinrichs KU. 2006. Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. Proc. Natl. Acad. Sci. USA 103: 3846–3851. http://dx.doi.org/10.1073/pnas.0600035103

Bižić-Ionescu M, Zeder M, Ionescu D, Orlić S, Fuchs BM, Grossart, H-P, Amann R. 2014. Comparison of bacterial communities on limnic versus coastal marine particles reveals profound differences in colonization. Environ. Microbiol. http://dx.doi.org/10.1111/1462-2920.12466.

Blaxter ML. 2004. The promise of a DNA taxonomy. Philos. Trans. R. Soc. Lond. (B Biol. Sci.) 359: 669–679. http://dx.doi.org/10.1098/rstb.2003.1447

Boeuf D, Cottrell MT, Kirchman DL, Lebaron P, Jeanthon C. 2013. Summer community structure of aerobic anoxygenic phototrophic bacteria in the western Arctic Ocean. FEMS Microbiol. Ecol. 85: 417–432. http://dx.doi.org/10.1111/1574-6941.12130

Bowen J, Weisman D, Yasuda M, Jayakumar A, Morrison H, Ward B. 2015. Marine oxygen-deficient zones harbor depauperate denitrifying communities compared to novel genetic diversity in coastal sediments. Microb. Ecol.: 1–11. http://dx.doi.org/10.1007/s00248-015-0582-y

Brady A, Salzberg SL. 2009. Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. Nature Methods 6: 673–676. http://dx.doi.org/10.1038/nmeth.1358

Brown MV, Ostrowski M, Grzymski JJ, Lauro FM. 2014. A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. Mar. Genomics 15: 17–28. http://dx.doi.org/10.1016/j.margen.2014.03.002

Cai L, Ye L, Tong AHY, Lok S, Zhang T. 2013. Biased diversity metrics revealed by bacterial 16S pyrotags derived from different primer sets. PLOS ONE 8: e53649. http://dx.doi.org/10.1371/journal.pone.0053649

Chen K, Pachter L. 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLOS Comput. Biol. 1: e24. http://dx.doi.org/10.1371/journal.pcbi.0010024

Chun J, Rainey FA. 2014. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int. J. Syst. Evol. Microbiol. 64: 316–324. http://dx.doi.org/10.1099/ijs.0.054171-0

Chun J, Lee JH, Jung Y, Kim M, Kim S, Kim BK, Lim YW. 2007. EzTaxon: A web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. Int. J. Syst. Evol. Microbiol. 57: 2259–2261. http://dx.doi.org/10.1099/ijs.0.64915-0

Clemente JC, Jansson J, Valiente G. 2011. Flexible taxonomic assignment of ambiguous sequencing reads. BMC Bioinformatics 12: 8. http://dx.doi.org/10.1186/1471-2105-12-8

Cock JM, Tessmar-Raible K, Boyen C, Viard F. 2010. Introduction to Marine Genomics, Advances in Marine Genomics 1. Springer, Dordrecht, 399 pp.

Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM. 2005. The Ribosomal Database Project (RDP-II): Sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. 33: D294–D296. http://dx.doi.org/10.1093/nar/gki038

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42: D633–642. http://dx.doi.org/10.1093/nar/gkt1244

Crab R, Defoirdt T, Bossier P, Verstraete W. 2012. Biofloc technology in aquaculture: Beneficial effects and future challenges. Aquaculture 356–357: 351–356. http://dx.doi.org/10.1016/j.aquaculture.2012.04.046

Curtis TP, Sloan WT, Scannell JW. 2002. Estimating prokaryotic diversity and its limits. Proc. Natl. Acad. Sci. USA 99: 10494–10499. http://dx.doi.org/10.1073/pnas.142680199

Fuhrman JA, Cram JA, Needham DM. 2015. Marine microbial community dynamics and their ecological interpretation. Nature Rev. Microbiol. 13: 133–146. http://dx.doi.org/10.1038/nrmicro3417

Gibbons SM, Caporaso JG, Pirrung M, Field D, Knight R, Gilbert JA. 2013. Evidence for a persistent microbial seed bank throughout the global ocean. Proc. Natl. Acad. Sci. USA 110: 4651–4655. http://dx.doi.org/10.1073/pnas.1217767110

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science 309: 1242–1245. http://dx.doi.org/10.1126/science.1114057

Gobet A, Boetius A, Ramette A. 2014. Ecological coherence of diversity patterns derived from classical fingerprinting and Next Generation Sequencing techniques. Environ. Microbiol. 16: 2672–2681. http://dx.doi.org/10.1111/1462-2920.12308

Hahnke RL, Probian C, Fuchs BM, Harder J. 2013. Variations in pelagic bacterial communities in the North Atlantic Ocean coincide with water bodies. Aquat. Microb. Ecol. 71: 131–140. http://dx.doi.org/10.3354/ame01668

Handelsman J. 2004. Metagenomics: Application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev. 68: 669–685. http://dx.doi.org/10.1128/mmbr.68.4.669-685.2004

Hatosy SM, Martiny JBH, Sachdeva R, Steele J, Fuhrman JA, Martiny AC. 2013. Beta diversity of marine bacteria depends on temporal scale. Ecology 94: 1898–1904. http://dx.doi.org/10.1890/12-2125.1

Havelsrud OE, Haverkamp TH, Kristensen T, Jakobsen KS, Rike AG. 2011. A metagenomic study of methanotrophic microorganisms in Coal Oil Point seep sediments. BMC Microbiology 11: 221. http://dx.doi.org/10.1186/1471-2180-11-221

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. (B) 270: 313–321. http://dx.doi.org/10.1098/rspb.2002.2218

Hellweger FL, van Sebille E, Fredrick ND. 2014. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. Science 345: 1346–1349. http://dx.doi.org/10.1126/science.1254421

Hong S, Bunge J, Leslin C, Jeon S, Epstein SS. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME J. 3: 1365–1373. http://dx.doi.org/10.1038/ismej.2009.89

Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable Tag sequencing. PLOS Genet. 4: e1000255. http://dx.doi.org/10.1371/journal.pgen.1000255

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. Genome Res. 17: 377–386. http://dx.doi.org/10.1101/gr.5969107

Huws SA, Edwards JE, Kim EJ, Scollan ND. 2007. Specificity and sensitivity of eubacterial primers utilized for molecular profiling of bacteria within complex microbial ecosystems. J. Microbiol. Meth. 70: 565–569. http://dx.doi.org/10.1016/j.mimet.2007.06.013

Jeon YS, Park SC, Lim J, Chun J, Kim BS. 2015. Improved pipeline for reducing erroneous identification by 16S rRNA sequences using the Illumina MiSeq platform. J. Microbiol. 53: 60–69. http://dx.doi.org/10.1007/s12275-015-4601-y

Kämpfer P, Glaeser SP. 2012. Prokaryotic taxonomy in the sequencing era: The polyphasic approach revisited. Environ. Microbiol. 14: 291–317. http://dx.doi.org/10.1111/j.1462-2920.2011.02615.x

Karl DM. 2007. Microbial oceanography: Paradigms, processes and promise. Nature Rev. Microbiol. 5: 759–769. http://dx.doi.org/10.1038/nrmicro1749

Karl DM, Church MJ. 2014. Microbial oceanography and the Hawaii Ocean time-series programme. Nature Rev. Microbiol. 12: 699–713. http://dx.doi.org/10.1038/nrmicro3333

Karsch-Mizrachi I, Nakamura Y, Cochrane G. 2012. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 40: D33–D37. http://dx.doi.org/10.1093/nar/gkr1006

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie JM, Follows M, Gorsky G, Hingamp P, Iudicone D, Jaillon O, Kandels-Lewis S, Krzic U, Not F, Ogata H, Pesant S, Reynaud EG, Sardet C, Sieracki ME, Speich S, Velayoudon D, Weissenbach J, Wincker P, the Tara Oceans Consortium. 2011. A holistic approach to marine eco-systems biology. PLOS Biol. 9: e1001177.http://dx.doi.org/10.1371/journal.pbio.1001177

Kato S, Ikehata K, Shibuya T, Urabe T, Ohkuma M, Yamagishi A. 2015. Potential for biogeochemical cycling of sulfur, iron and carbon within massive sulfide deposits below the seafloor. Environ. Microbiol. 17: 1817–1835. http://dx.doi.org/10.1111/1462-2920.12648

Kim M, Lee KH, Yoon SW, Kim BS, Chun J, Yi H. 2013. Analytical tools and databases for metagenomics in the next-generation sequencing era. Genomics Inform. 11: 102–113. http://dx.doi.org/10.5808/GI.2013.11.3.102

Kim M, Oh HS, Park SC, Chun J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int. J. Syst. Evol. Microbiol. 64: 346–351. http://dx.doi.org/10.1099/ijs.0.059774-0

Kim OS, Cho YJ, Lee K, Yoon SH, Kim M, Na H, Park SC, Jeon YS, Lee JH, Yi H, Won S, Chun J. 2012. Introducing EzTaxon-e: A prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. Int. J. Syst. Evol. Microbiol. 62: 716–721. http://dx.doi.org/10.1099/ijs.0.038075-0

Klindworth A, Mann AJ, Huang S, Wichels A, Quast C, Waldmann J, Teeling H, Glöckner FO. 2014. Diversity and activity of marine bacterioplankton during a diatom bloom in the North Sea assessed by total RNA and pyrotag sequencing. Mar. Genomics 18B: 185–192. http://dx.doi.org/10.1016/j.margen.2014.08.007

Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. USA 102: 2567–2572. http://dx.doi.org/10.1073/pnas.0409727102

Koonin EV. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nature Rev. Microbiol. 1: 127–136. http://dx.doi.org/10.1038/nrmicro751

Kress WJ, Erickson DL. 2012. DNA Barcodes: Methods and Protocols. Methods in Molecular Biology 858. Humana Press/ Springer, New York, 470 pp.

Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, Sicheritz-Pontén T, Aarestrup FM, Ussery DW, Lund O. 2014. Benchmarking of methods for genomic taxonomy. J. Clin. Microbiol. 52: 1529–1539. http://dx.doi.org/10.1128/jcm.02981-13

Laverock B, Tait K, Gilbert JA, Osborn AM, Widdicombe S. 2014. Impacts of bioturbation on temporal variation in bacterial and archaeal nitrogen-cycling gene abundance in coastal sediments. Environ. Microbiol. Rep. 6: 113–121. http://dx.doi.org/10.1111/1758-2229.12115

Lebonah DE, Dileep A, Chandrasekhar K, Sreevani S, Sreedevi B y Kumari PJ. 2014. DNA barcoding on bacteria: A review. Adv. Biol. 2014: 541787. http://dx.doi.org/10.1155/2014/541787

Ley RE, Peterson DA, Gordon JI. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124: 837–848. http://dx.doi.org/10.1016/j.cell.2006.02.017

Lipsewers YA, Bale NJ, Hopmans EC, Schouten S, Sinninghe Damste JS, Villanueva L. 2014. Seasonality and depth distribution of the abundance and activity of ammonia oxidizing microorganisms in marine coastal sediments (North Sea). Front. Microbiol. 5: 472. http://dx.doi.org/10.3389/fmicb.2014.00472

Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 35: e120. http://dx.doi.org/10.1093/nar/gkm541

Liu Z, DeSantis TZ, Andersen GL, Knight R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 36: e120–e120. http://dx.doi.org/10.1093/nar/gkn491

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G, Raes J, Poulain J, Jaillon O, Wincker P, Kandels-Lewis S, Karsenti E, Bork P, Acinas SG. 2014. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ. Microbiol. 16: 2659–2671. http://dx.doi.org/10.1111/1462-2920.12250

MacDonald NJ, Parks DH, Beiko RG. 2012. Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Res. 40: e111. http://dx.doi.org/10.1093/nar/gks335

Meier-Kolthoff JP, Klenk HP, Göker M. 2014. Taxonomic use of DNA G + C content and DNA–DNA hybridization in the genomic age. Int. J. Syst. Evol. Microbiol. 64: 352–356.

Miller C, Baker B, Thomas B, Singer S, Banfield J. 2011. EMIRGE: Reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. 12: R44. http://dx.doi.org/10.1186/gb-2011-12-5-r44

Miller CS, Handley KM, Wrighton KC, Frischkorn KR, Thomas BC, Banfield JF. 2013. Short-read assembly of full-length 16S amplicons reveals bacterial diversity in subsurface sediments. PLOS ONE 8: e56018. http://dx.doi.org/10.1371/journal.pone.0056018

Mohit V, Archambault P, Toupoint N, Lovejoy C. 2014. Phylogenetic differences in attached and free-living bacterial communities in a temperate coastal lagoon during summer, revealed via high-throughput 16S rRNA gene sequencing. Appl. Environ. Microbiol. 80: 2071–2083. http://dx.doi.org/10.1128/aem.02916-13

Moore ER, Mihaylova SA, Vandamme P, Krichevsky MI, Dijkshoorn L. 2010. Microbial systematics and taxonomy: Relevance for a microbial commons. Res. Microbiol. 161: 430–438. http://dx.doi.org/10.1016/j.resmic.2010.05.007

Mueller RS, Bryson S, Kieft B, Li Z, Pett-Ridge J, Chavez F, Hettich RL, Pan C, Mayali X. 2015. Metagenome sequencing of a coastal marine microbial community from Monterey Bay, California. Genome Announcements 3: e00341–00315. http://dx.doi.org/10.1128/genomeA.00341-15

Myers EW, Miller W. 1988. Optimal alignments in linear space. Comput. Appl. Biosci. 4: 11–17. http://dx.doi.org/10.1093/bioinformatics/4.1.11

[NRC] National Research Council (US). 2007. The New Science of Metagenomics: Revealing the secrets of our microbial planet. National Academies Press, Washington, DC, 158 pp.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 2933–2935. http://dx.doi.org/10.1093/bioinformatics/btt509

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: Inference of RNA alignments. Bioinformatics 25: 1335–1337. http://dx.doi.org/10.1093/bioinformatics/btp157

Nelson KE. 2011. Metagenomics of the Human Body. Springer Science + Business Media, LLC, New York, 351 pp.

Nemec A, De Baere T, Tjernberg I, Vaneechoutte M, van der Reijden TJ, Dijkshoorn L. 2001. *Acinetobacter ursingii* sp. nov. and *Acinetobacter schindleri* sp. nov., isolated from human clinical specimens. Int. J. Syst. Evol. Microbiol. 51: 1891–1899. http://dx.doi.org/10.1099/00207713-51-5-1891

Olsen LA, Choffnes ER, Mack A. 2012. The Social Biology of Microbial Communities. Workshop Summary. National Academies Press, Washington, D.C., xxviii, 603 pp.

Pace NR, Stahl DA, Lane DJ, Olsen GJ. 1986. The analysis of natural microbial populations by ribosomal RNA sequences. In: Marshall KC (ed.), Advances in Microbial Ecology. Springer, US, pp. 1–55.

Pedros-Alio C. 2006. Genomics and marine microbial ecology. Int. Microbiol. 9: 191–197.

Prabavathi R, Mathivanan V. 2012. Isolation of genomic DNA and construction of metagenomic library from marine soil sediments. Int. J. Pharma Bio Sci. (IJPBS) 3: B1–B9.

Pruesse E, Peplies J, Glöckner FO. 2012. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28: 1823–1829. http://dx.doi.org/10.1093/bioinformatics/bts252

Ramasamy D, Mishra AK, Lagier J-C, Padhmanabhan R, Rossi M, Sentausa E, Raoult D, Fournier P-E. 2014. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int. J. Syst. Evol. Microbiol. 64: 384–391. http://dx.doi.org/10.1099/ijs.0.057091-0

Ramette A, Tiedje JM, Boetius A. 2009. Impact of space, time and complex environments on microbial communities. Clin. Microbiol. Infect. 15: 60–62. http://dx.doi.org/10.1111/j.1469-0691.2008.02692.x

Rappé MS, Giovannoni SJ. 2003. The Uncultured Microbial Majority. Annu. Rev. Microbiol. 57: 369-394. http://dx.doi.org/10.1146/annurev.micro.57.030502.090759

Rodicio MDR, Mendoza MDC. 2004. Identificación bacteriana mediante secuenciación del ARNr 16S: Fundamento, metodología y aplicaciones en microbiología clínica. Enfermedades Infecciosas y Microbiología Clínica 22: 238–245.

Rosselló-Mora R, Amann R. 2001. The species concept for prokaryotes. FEMS Microbiol. Rev. 25: 39–67. http://dx.doi.org/10.1111/j.1574-6976.2001.tb00571.x

Roussel EG, Konn C, Charlou JL, Donval JP, Fouquet Y, Querellou J, Prieur D, Cambon Bonavita MA. 2011. Comparison of microbial communities associated with three Atlantic ultramafic hydrothermal systems. FEMS Microbiol. Ecol. 77: 647–665. http://dx.doi.org/10.1111/j.1574-6941.2011.01161.x

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y-H, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC. 2007. The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLOS Biol 5: e77. http://dx.doi.org/10.1371/journal.pbio.0050077

Santamaria M, Fosso B, Consiglio A, De Caro G, Grillo G, Licciulli F, Liuni S, Marzano M, Alonso-Alemany D, Valiente G, Pesole G. 2012. Reference databases for taxonomic assignment in metagenomics. Brief. Bioinformatics 13: 682–695. http://dx.doi.org/10.1093/bib/bbs036

Satinsky BM, Crump BC, Smith CB, Sharma S, Zielinski BL, Doherty M, Meng J, Sun S, Medeiros PM, Paul JH, Coles VJ, Yager PL, Moran MA. 2014. Microspatial gene expression patterns in the Amazon River Plume. Proc. Natl. Acad. Sci. USA 111: 11085–11090. http://dx.doi.org/10.1073/pnas.1402782111

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R. 2005. Towards writing the encyclopaedia of life: An introduction to DNA barcoding. Philos. Trans. R. Soc. Lond. (B Biol. Sci.) 360: 1805–1011. http://dx.doi.org/10.1098/rstb.2005.1730

Schleifer KH. 2009. Classification of Bacteria and Archaea: Past, present and future. Syst. Appl. Microbiol. 32: 533–542. http://dx.doi.org/10.1016/j.syapm.2009.09.002

Selama O, James P, Nateche F, Wellington EMH, Hacene H. 2013. The world bacterial biogeography and biodiversity through databases: A case study of NCBI nucleotide database and GBIF database. BioMed Research International 2013, 11 pp. http://dx.doi.org/10.1155/2013/240175

Shao S, Luan X, Dang H, Zhou H, Zhao Y, Liu H, Zhang Y, Dai L, Ye Y, Klotz MG. 2014. Deep-sea methane seep sediments in the Okhotsk Sea sustain diverse and abundant anammox bacteria. FEMS Microbiol. Ecol. 87: 503–516. http://dx.doi.org/10.1111/1574-6941.12241

Simon HM, Smith MW, Herfort L. 2014. Metagenomic insights into particles and their associated microbiota in a coastal margin ecosystem. Front. Microbiol. 5: 466. http://dx.doi.org/10.3389/fmicb.2014.00466

Sjöstedt J, Martiny JBH, Munk P, Riemann L. 2014. Abundance of broad bacterial taxa in the Sargasso Sea explained by environmental conditions but not water mass. Appl. Environ. Microbiol. 80: 2786–2795. http://dx.doi.org/10.1128/aem.00099-14

Smith MW, Allen LZ, Allen AE, Herfort L, Simon HM. 2013. Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. Front. Microbiol. 4: 120. http://dx.doi.org/10.3389/fmicb.2013.00120

Stackebrandt E, Ebers J. 2006. Taxonomic parameters revisited: Tarnished gold standards. Microbiol. Today 33: 152–155.

Sun DL, Jiang X, Wu QL, Zhou NY. 2013. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Appl. Environ. Microbiol. 79: 5962–5969. http://dx.doi.org/10.1128/AEM.01282-13

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, coordinators TO, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. 2015. Structure and function of the global ocean microbiome. Science 348. http://dx.doi.org/10.1126/science.1261359

Tanabe AS, Toju H. 2013. Two new computational methods for universal DNA barcoding: A benchmark using barcode sequences of bacteria, archaea, animals, fungi, and land plants. PLOS ONE 8: e76910. http://dx.doi.org/10.1371/journal.pone.0076910

Tang K, Liu K, Jiao N, Zhang Y, Chen CT. 2013. Functional metagenomic investigations of microbial communities in a shallow-sea hydrothermal system. PLOS ONE 8: e72958. http://dx.doi.org/10.1371/journal.pone.0072958

Tindall BJ, Rosselló-Mora R, Busse HJ, Ludwig W, Kampfer P. 2010. Notes on the characterization of prokaryote strains for taxonomic purposes. Int. J. Syst. Evol. Microbiol. 60: 249–266. http://dx.doi.org/10.1099/ijs.0.016949-0

Tringe SG, Hugenholtz P. 2008. A renaissance for the pioneering 16S rRNA gene. Curr. Opin. Microbiol. 11: 442–446. http://dx.doi.org/10.1016/j.mib.2008.09.011

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO. 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science 304: 66–74. http://dx.doi.org/10.1126/science.1093857

Vigneron A, Cruaud P, Roussel EG, Pignet P, Caprais J-C, Callac N, Ciobanu M-C, Godfroy A, Cragg BA, Parkes JR, Van Nostrand JD, He Z, Zhou J, Toffin L. 2014. Phylogenetic and functional diversity of microbial communities associated with subsurface sediments of the Sonora Margin, Guaymas Basin. PLOS ONE 9: e104427. http://dx.doi.org/10.1371/journal.pone.0104427

Vilo C, Dong Q. 2012. Evaluation of the RDP classifier accuracy using 16S rRNA gene variable regions. Metagenomics 1: 1–5. http://dx.doi.org/10.4303/mg/235551

Vos M. 2011. A species concept for bacteria based on adaptive divergence. Trends Microbiol. 19: 1–7. http://dx.doi.org/10.1016/j.tim.2010.10.003

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73: 5261–5267. http://dx.doi.org/10.1128/aem.00062-07

Wang Y, Qian PY. 2009. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLOS ONE 4: e7401. http://dx.doi.org/10.1371/journal.pone.0007401

Wang Y, Sheng HF, He Y, Wu JY, Jiang YX, Tam NF, Zhou HW. 2012. Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. Appl. Environ. Microbiol. 78: 8264–8271. http://dx.doi.org/10.1128/aem.01821-12

Wood D, Salzberg S. 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15: R46. http://dx.doi.org/10.1186/gb-2014-15-3-r46

Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH, Ludwig W, Glockner FO, Rossello-Mora R. 2008. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. 31: 241–250. http://dx.doi.org/10.1016/j.syapm.2008.07.001

Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, Glockner FO, Rossello-Mora R. 2010. Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. Syst. Appl. Microbiol. 33: 291–299. http://dx.doi.org/10.1016/j.syapm.2010.08.001

Yarza P, Yilmaz P, Pruesse E, Glockner FO, Ludwig W, Schleifer KH, Whitman WB, Euzeby J, Amann R, Rossello-Mora R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12: 635–645. http://dx.doi.org/10.1038/nrmicro3330

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 42: D643–D648. http://dx.doi.org/10.1093/nar/gkt1209

Zinger L, Gobet A, Pommier T. 2012. Two decades of describing the unseen majority of aquatic microbial diversity. Mol. Ecol. 21: 1878–1896. http://dx.doi.org/10.1111/j.1365-294X.2011.05362.x

Received November 2014,
accepted September 2015.