UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Reconstrucción de genomas a partir de metagenomas del Golfo de México

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

BIÓLOGO

P R E S E N T A:

Miguel Ángel González Arias

Director de tesis
Dr. Lorenzo Patrick Segovia Forcella
Ciudad Universitaria, CDMX, **2020**

Datos del jurado y sustentante

Alumno (Sustentante)
Miguel Ángel González Arias
Biología, Facultad de Ciencias, UNAM

Tutor (Secretario)
Dr. Lorenzo Patrick Segovia Forcella
Instituto de Biotecnología, UNAM

Sinodal 1 (Presidente)
Dr. Arturo Carlos II Becerra Bracho
Facultad de Ciencias, UNAM

Sinodal 2 (Vocal)
Dra. María Esperanza Martínez Romero
Centro de Ciencias Genómicas, UNAM

Sinodal 3 (Suplente 1)
Dr. Luis David Alcaraz Peraza
Facultad de Ciencias, UNAM

Sinodal 4 (Suplente 2)
Dr. José Luis Puente García
Instituto de Biotecnología, UNAM

Datos del trabajo escrito

González Arias M. A. (2020) *Reconstrucción de genomas a partir de metagenomas del Golfo de México*. Tesis de licenciatura. Facultad de Ciencias. Universidad Nacional Autónoma de México. México. 53p.

“How inappropriate to call this planet Earth, when it is clearly Ocean”

("Qué inapropiado llamar a este planeta Tierra, cuando claramente es Océano”)

— Arthur C. Clarke

Agradecimientos

Académicos

A la UNAM.

A Alejandra Escobar, Ernestina Godoy, Marisol Navarro, Jerome Verleyen, Julián Torres, Benjamin Tully y Alexandre Almeida por los scripts compartidos y su asesoría.

Al equipo bioinformático del laboratorio 12: Angélica Domínguez, Alejandro Garciarubio, Andrés De Sandozequi, Alfredo Rodríguez, Eira Aguirre, Lorenzo Segovia, Raquel Neri y especialmente a Rafael López, por su compañerismo, ayuda y asesorías.

Al Instituto de Biotecnología (UNAM) por el acceso al clúster Teopanzolco y FOS (Lab. 12) así como a José Luis Puente, Claudia Treviño, Enrique Rudiño y en general, a los investigadores del instituto que forman parte activa del taller “La biología a partir de las biomoléculas”.

Al Consorcio de Investigación del Golfo de México por el acceso al clúster BLAU así como por la colecta y procesamiento de las muestras.

A Lorenzo Segovia y Claudia Martínez por sus atenciones y apoyo económico (IBt/PAPIIT).

A los miembros de mi comité de sinodal por sus comentarios y correcciones.

A Lorenzo Segovia por introducirme en la bioinformática.

Personales

>GAMA00001
GCGATGATTAGTCCCGCCGACCGGGAGAGTTATCATGAAAGGATGGCCAACTAG

>GAMA00002
GCGATGATTAGCTTTGCCATGATTCTTATTGCCAGGGAATCTTATGCCATGATCGGGTAGAGC

>GAMA00003
GCCCTGTAGTCCCAATGAGAATGTTAGAATTAGAGTTGCTAGTATGATGAAAGTTGTTAGAACTAG
TCCTGTTAGGAATGAGAAATGGAACATGCCAATCCCGAAAGGATGATTACCATTGATTAGCTTCTT
GAAGGGGCCAGGCATGCCTCCACCGCCGACTAGAACGATGAAGAAAGTACCTAGTAT

>GAMA00004
GCCACCTAGGATTAGAGCATGTGATGTCATGCCAGTGGCAGAGCTTGTATTGCCTCC

Nota: Considere el primer marco de lectura y los codones UGA y UAG para la pirrolisina (O) y selenocisteína (U), respectivamente.

Índice

Agradecimientos académicos 4

Agradecimientos personales 4

1. Resumen 6

2. *Abstract* 7

3. Introducción 8

3.1 Antecedentes 12

3.2 Colecta y procesamiento de las muestras 13

4. Planteamiento 16

4.1 Justificación 16

4.2 Objetivo general 16

4.3 Objetivos particulares 16

5. Método 17

5.1 *Software* y *hardware* empleado 17

5.2 Análisis de los reads 18

5.3 Reconstrucción de los MAGs 20

5.4 Análisis de los MAGs 22

6. Resultados y discusión 24

6.1 Clasificación de los reads 24

6.2 MAGs reconstruidos 30

6.3 Inferencias del contexto ecológico de los MAGs 35

6.4 Limitantes 41

6.5 Perspectivas 43

7. Conclusiones 46

8. Referencias 47

9. Anexo 51

9.1. Líneas de comando 51

9.2. Sitios de interés 53

1. Resumen

En los océanos hay más que solo peces, mucho más. Esta enorme biodiversidad se encuentra principalmente representada por microorganismos eucariontes y procariontes, como son los protistas, bacterias, arqueas e incluso, los virus. En conjunto, dichos grupos regulan los ciclos biogeoquímicos del ecosistema, así como de la Tierra y forman la base de las cadenas tróficas que soportan a prácticamente cualquier otra forma de vida, incluidos nosotros y nuestras sociedades. En los últimos seis años, el Consorcio de Investigación del Golfo de México ha generado los primeros registros microbiológicos detallados de cuenca oceánica del Golfo de México correspondientes al territorio mexicano. Sin embargo, aún no han sido analizados los datos metagenómicos generados, los cuales permiten conocer detalles sobre la biodiversidad de microorganismos y los respectivos roles que desempeñan en sus comunidades sin la necesidad de cultivarlos en el laboratorio.

En esta tesis se analizaron datos metagenómicos con el objetivo reconstruir los genomas de los procariontes presentes en muestras metagenómicas de la columna de agua y en sedimentos que fueron colectadas en profundidades y regiones diferentes dentro de la cuenca oceánica del Golfo de México.

Se identificó la composición de las comunidades de procariontes aplicando métodos bioinformáticos permitiendo observar patrones similares de biodiversidad a los patrones globales previamente identificados en otros megaproyectos oceánicos. Además, se reconstruyeron un total de 116 genomas reconstruidos a partir de metagenomas, 28 de los cuales representan categorías taxonómicas a nivel de familia, género y especie sin previo registro en la *Genome Taxonomy Database*. Entre ellos, solo cinco genomas fueron lo suficientemente completos para realizar estudios metabólicos detallados que permitieron proponer hipótesis sobre el contexto ecológico de dichos microorganismos, abarcando posibles hábitos parasitarios y formas de vida libre con un metabolismo bien articulado y que además, representan un valioso recurso para las bases de datos genómicas.

2. *Abstract*

There are more than just fish in the oceans, much more. This huge biodiversity is mainly represented by eukaryotic and prokaryotic microorganisms like protists, bacteria, archaea and even viruses. Altogether, these groups regulates the biogeochemical cycles both ecosystem and the Earth and form the basis of the trophic chains that support practically any other life form, including us and our societies. In the last six years, the Consorcio de Investigación del Golfo de México *has generated the first detailed microbiological records of the ocean basin of Gulf of Mexico for the Mexican territory. However, the metagenomic data generated has not yet been analyzed, which allows us to know details about the biodiversity of microorganisms and the detailed roles that play in their communities without the need of culture them in the laboratory.*

In this thesis, metagenomic data was analyzed with the purpose of reconstructing the genomes of prokaryotes present in metagenomic samples of the water column and sediments that were collected in different depths and regions within the oceanic basin of the Gulf of Mexico.

The composition of the prokaryotic communities was identified by applying bioinformatic methods allowing to observe similar patterns of biodiversity to the global patterns previously identified in other oceanic megaprojects. Additionally, a total of 116 metagenome assembled genomes was reconstructed, 28 of which represent taxonomic categories at the family, genus and species level without prior registration in the Genome Taxonomy Database. Among them, only five genomes were complete enough to carry out detailed metabolic studies that made it possible to propose hypotheses about the ecological context of these microorganisms, covering possible parasitic habits and free life forms with a well-articulated metabolism and that also represent a valuable resource for genomic databases.

3. Introducción

Las actuales tecnologías de secuenciación permiten capturar información genómica de varios de los microorganismos en un hábitat, permitiendo conocer qué tipo de microorganismos están ahí y qué es lo que podrían estar haciendo[1,2]. Sin embargo, no todos los microorganismos son fáciles de cultivar, debido a sus complejas condiciones de crecimiento. Uno de los primeros enfoques que surgió para conocer la presencia los microorganismos sin necesidad de cultivarlos en el laboratorio fue la secuenciación del gen ribosomal 16S, que además se ser ampliamente usado actualmente, puede ser utilizado como un reloj molecular evolutivo[3]. Posteriormente, a inicios del siglo XXI surgió la metagenómica cuyo objetivo es identificar todo el contenido genómico de una comunidad microbiana presente en tiempo y espacio especificos[4]. Con la metagenómica es posible conocer parte de la biodiversidad de un ambiente en términos de sus taxa y/o sus respectivas capacidades metabólicas mediante la reconstrucción de los genes así como de los genomas de los microorganismos que ahí residen y a los que usualmente se refiere como genomas únicos amplificados o genomas reconstruidos a partir de metagenomas (SAGs y MAGs, por sus siglas en inglés, respectivamente) según la forma en que sean procesados[5] (Fig. 1; Fig. 2).

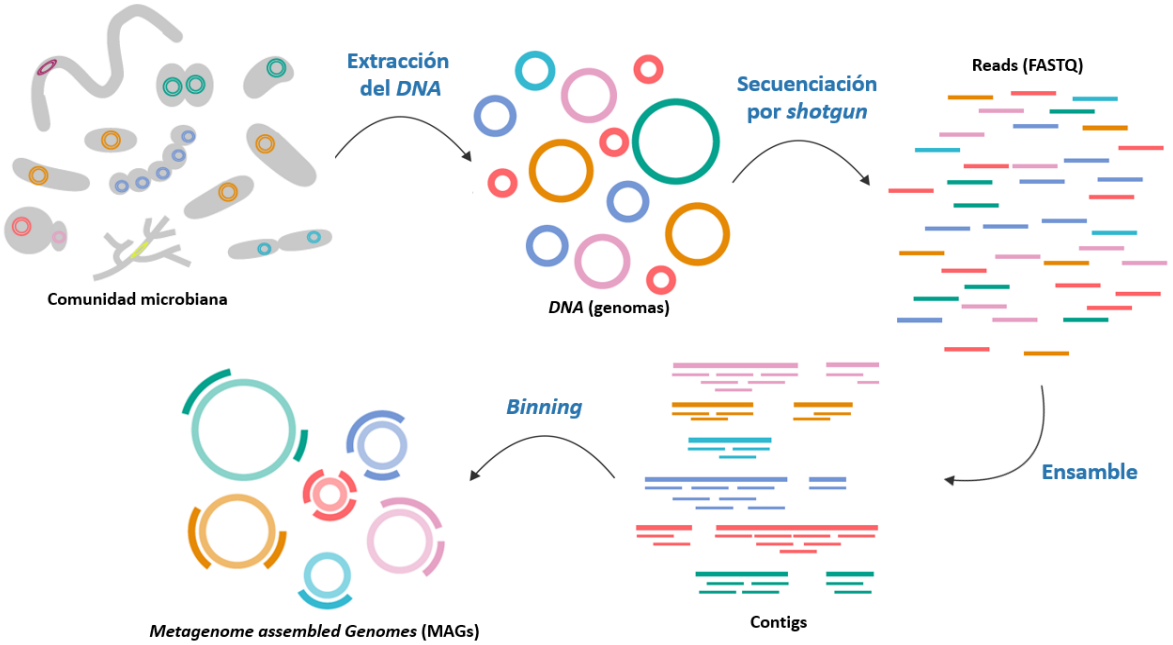


Figura 1. Descripción general del proceso para reconstruir genomas a partir de metagenomas (MAGs). Se muestra en el diagrama con letras azules los procesos y en negritas los tipos de datos que tienen que ser analizados (Reproducido de la ref. 6). El binning es un proceso exclusivo en la reconstrucción de MAGs, en los SAGs se excluye este paso debido a su proceder.

La idea es simple y se soporta por una serie de fundamentos bioquímicos, informáticos y matemáticos detrás de ellos[7]. Comienza con el muestreo y extracción del DNA a partir de las comunidades biológicas, el cual tiene que ser fragmentado aleatoriamente y a cada uno de los fragmentos, adicionar secuencias adaptadoras mediante ligación para facilitar su uso. Posteriormente, se preparan, amplifican y secuencian las librerías genómicas, generando millones de lecturas de secuencias (*reads*) de longitudes entre 75 a 300 pares de bases (pb) usualmente almacenadas en formato de archivo FASTQ, el cual incluye secuencias de nucleótidos y su respectiva calidad (en código ASCII)[8]. Los reads pueden ser analizados o bien, pueden ser ordenados en estructuras de mayor tamaño conocidas como *contigs* mediante el su ensamble[9,10]. Los contigs pueden ser analizados o pueden ser agrupados en grupos (*bins)* mediante el proceso de *binning*, el cual se basa en la composición de las secuencias (frecuencias de tetranucleótidos, dinucleótidos, GC%), la profundidad de las secuencias (cobertura de los reads en los contigs) y su perfil filogenético, entre otras características genómicas[11,12]. Un bin puede contener uno o varios contigs de diferentes tamaños. Cuando el tamaño y/o número de estos contigs es muy pequeño (por ejemplo, de uno a 10 contigs de 500pb) el bin representa un conjunto de contigs que se agruparon por su similitud; cuando un bin contiene uno o varios contigs de gran tamaño (por ejemplo, de uno a 10 contigs con longitudes entre 50kpb a 2Mpb) y además se identifica la presencia de genes de copia única dentro de él, representa un MAG[13,14]. Por su naturaleza, los MAGs presentan errores que hay que reconocer y que principalmente se generan durante el ensamble o el binning, sin embargo, también pueden representar el genoma de un linaje recuperado a partir de sus poblaciones naturales y además, brindan una oportunidad única para conocer la biodiversidad en términos taxonómicos y genómicos[15] (Fig. 2).

Respecto al avance científico en esta área, en el año 2004 se reportó el primer MAG reconstruido a partir de un metagenoma aislado de un *biofilm* con una comunidad biológica definida y poco diversa[16], y aunque el proceso de ensamble fue logrado mediante un ensamblador diseñado para genomas eucariontes, se logró demostrar que la reconstrucción de genomas era posible.

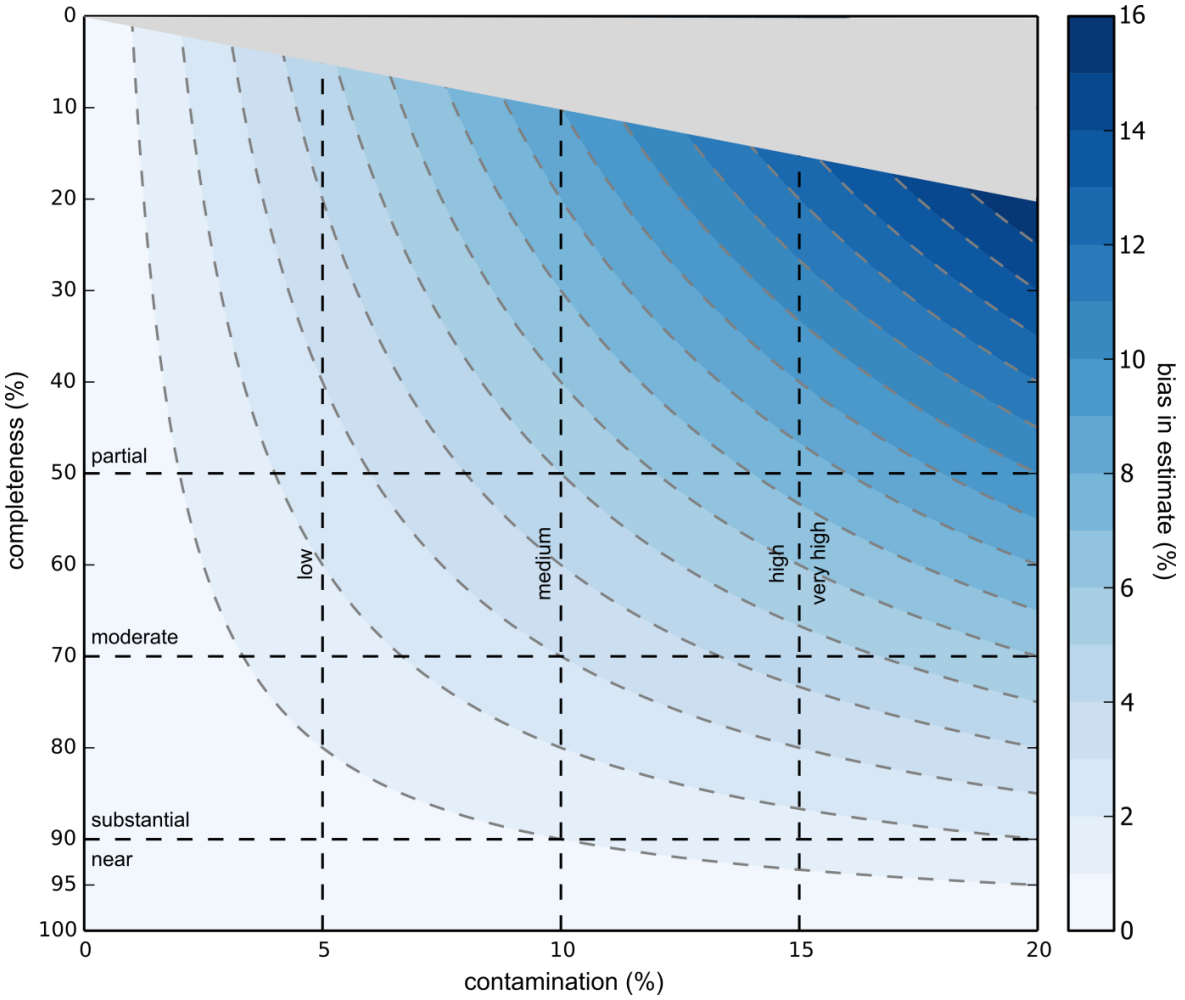


Figura 2. Modelo binomial empleado en CheckM del nivel de sesgo de las estimaciones del *completeness* y contaminación de un genoma (Reproducido de la ref. 14). Estas dos métricas influyen en la interpretación de la información y se basan en la presencia y repeticiones de genes de copia única en un genoma, los cuales pueden ser conjuntos universales para procariontes como los propuestos en la *Enveomics collection*[17] e implementados en MiGA[18], conjuntos para los tres dominios como los identificados por Hug *et al.*[19] o conjuntos taxón-específicos como es implementado en CheckM[14]. En cualquiera de los casos, el estándar actual establecido para reportar MAGs (MIMAG) y SAGs (MISAG) es que no cuenten con >10% de contaminación, clasificándolos en categorías de baja, media y alta calidad en función la presencia del aparato ribosomal (genes 5S, 16S y 23S), numero de tRNAs y de sus porcentajes de completeness y contaminación[13].

En 2011 se publica Genovo, el primer meta-ensamblador para comunidades microbianas[9,20]; en 2013, se publica uno de los primeros enfoques de binning metagenómico que daría paso a los algoritmos automatizados de la actualidad[21]. La metagenómica es un área en continuo desarrollo y actualmente se están integrando algoritmos de inteligencia artificial así como enfoques de la epigenética e inmunología, que junto con las nuevas tecnologías de secuenciación como PacBio® y Nanopore® permitirán conocer con mayor exactitud y precisión el genoma de los microorganismos[22–25]. Con dichos recursos, se han podido

caracterizar comunidades microbianas simples y complejas a partir de ecosistemas de fácil y difícil acceso, como los son los ambientes extremos[26,27], productos lacteos[28], el microbioma humano[29–31], suelos agrícolas[32] e incluso algunos trabajos abarcan biomas tan grandes como mares[33] y océanos[34,35] que buscan dilucidar la diversidad del microbioma de la Tierra[36]. En conjunto, todos estos trabajos han permitido proponer detalles sobre la topología del árbol de la vida y sus grandes dominios[19,37–39], llevando a proponer a las bacterias como el grupo más diverso; y aunque se han podido registrar grandes grupos nuevos e interesantes de bacterias[40] y arqueas[41], la mayor parte de la biodiversidad de los procariontes aún no ha sido registrada[42,43] (Fig. 3).

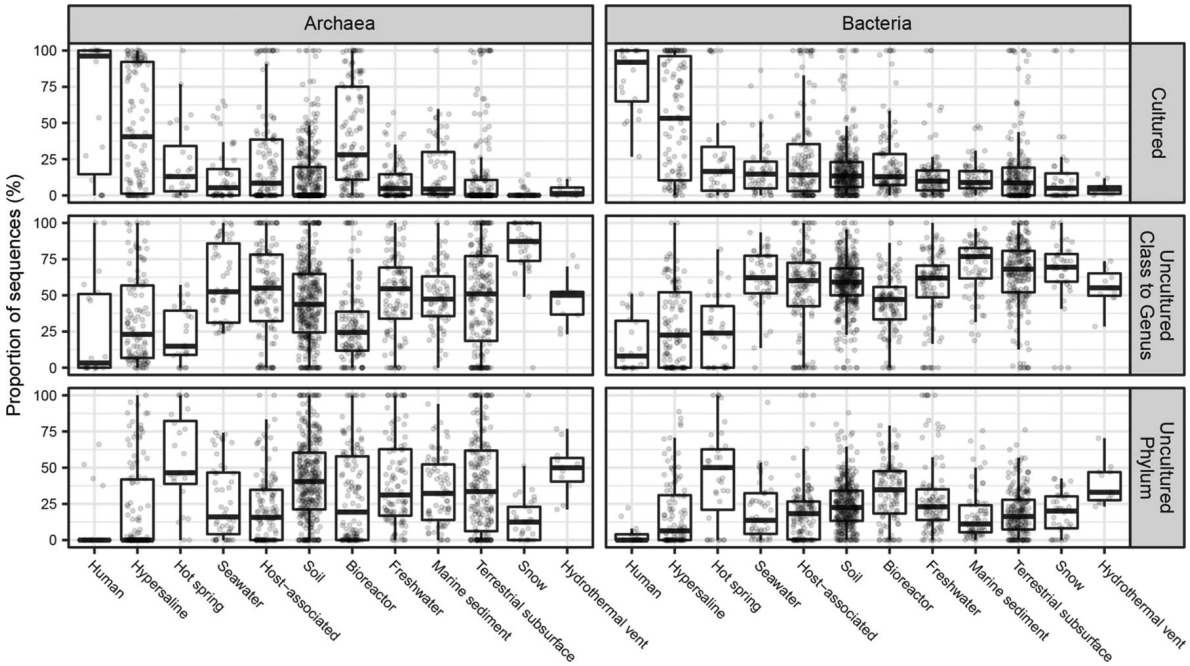


Figura 3. Registros y estimaciones de los procariontes cultivados y no cultivados en diferentes ambientes (Reproducido de la ref. 43). El análisis se basa en 1,135,799 secuencias del gen ribosomal 16S de procariontes presentes en bases de datos de amplicónes, metagenómicas y metatranscriptómicas para un total de 8,260 estudios. Nótese los registros y estimaciones realizadas para los ambientes de aguas y sedimentos marinos, de los cuales se desconoce una gran parte en comparación con otros ambientes como el microbioma humano del cual se tiene un registro relativamente bueno.

Actualmente, la dimensión y complejidad de los proyectos metagenómicos es muy variable, sin embargo, uno de los grandes hitos y su consecuente popularización surgió con el primer metagenoma masivo, el cual se obtuvo en 2004 con la secuenciación del mar del Sargazo[33]. Desde entonces la metagenómica marina ha sido una de las áreas con mayor interés debido

al potencial de las enzimas y proteínas con nuevas y/o con mejor rendimiento en procesos biotecnológicos, industriales o en procesos como la biorremediación, permitido conocer paralelamente los roles que juegan las comunidades microbianas en los ciclos biogeoquímicos de la Tierra así como en la regulación y resiliencia de los ecosistemas marinos[2,33–35,44–47] (Fig. 4).

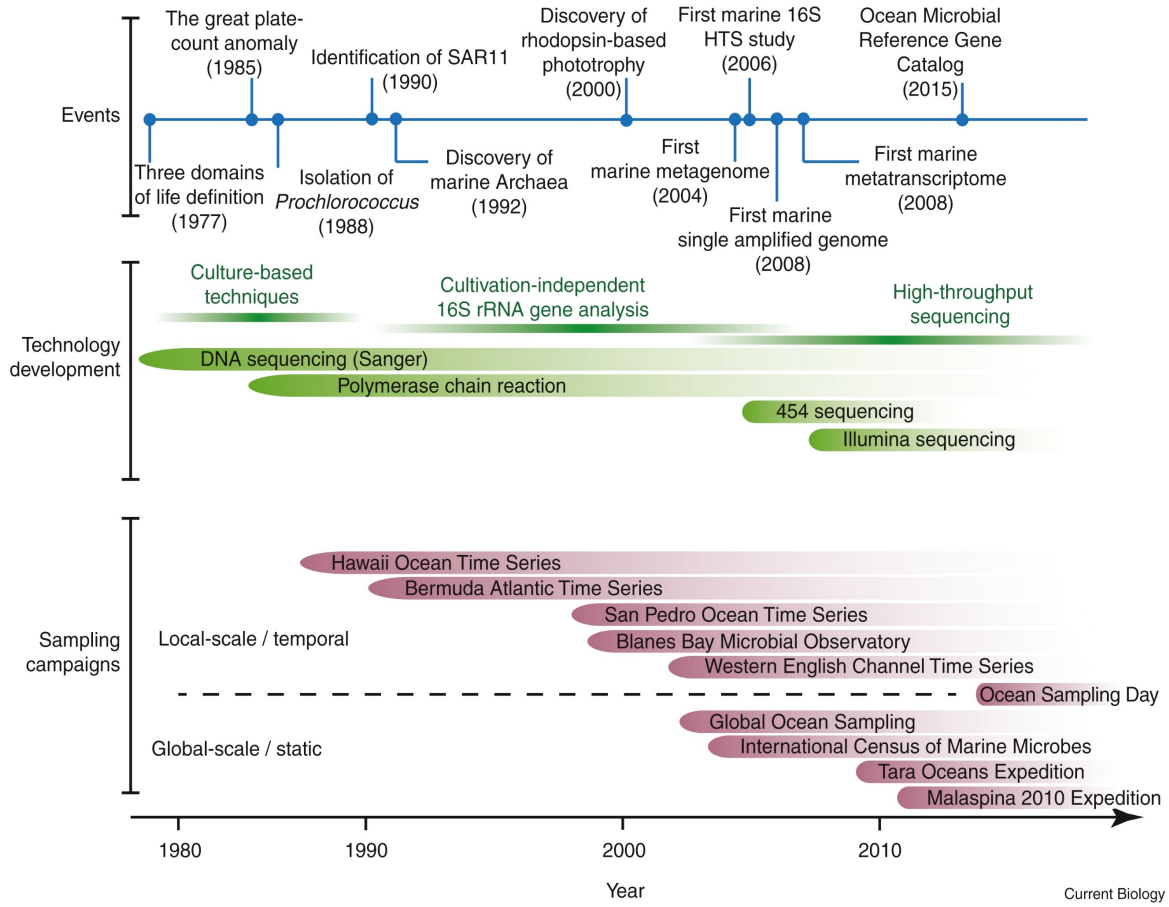


Figura 4. Progreso histórico de la metagenómica marina en relación con los principales descubrimientos, desarrollo de tecnologías de secuenciación y proyectos a escala local y global (Reproducido de la ref.44).

3.1 Antecedentes

Uno de los ambientes marinos que ha resultado de particular interés en la última década ha sido la cuenca oceánica del Golfo de México, debido a que en ella se encuentra una considerable presencia de hidrocarburos provenientes de fuentes naturales y antropogénicas, entre otros factores[48]. Debido a ello y demás motivos, en el 2014 se fundó el Consorcio de Investigación del Golfo de México (CIGoM) cuyo objetivo general es conocer

cuáles son las condiciones físicas, químicas, geológicas y biológicas del de la región Golfo de México que corresponde a los Estados Unidos Mexicanos, sentar una línea base de investigación y recabar información para la prevención de escenarios de contingencia ambiental[49]. Mediante colectas de muestras de sedimentos profundos, someros y de la columna de agua, se ha generado un perfil taxonómico base de las bacterias presentes en la región sur-oeste del Golfo de México mediante el análisis de un fragmento de las secuencias ribosomales 16S, así como su relación con varios factores abioticos[50]. Con dicho enfoque, fue posible determinar que la concentración de hidrocarburos aromáticos, contenido de materia orgánica y profundidad, son los principales factores abióticos que afectan al distribución de los microorganismos[50]. Sin embargo, para conocer más al respecto de sus capacidades metabólicas, se colectó información de muestras de agua y de sedimentos donde se aplicó el enfoque metagenómico con lo cual se espera poder conocer más acerca de la parte biótica del Golfo de México a través de sus microorganismos[51].

3.2 Colecta y procesamiento de las muestras

En esta tesis se analizaron los seis metagenomas de Raggi *et al.*[51] de alto rendimiento de extracción de DNA en los que las cantidades fueron suficientes para para realizar metagenómica. Todos los procedimientos de extracción, preparación y secuenciación de las muestras fueron realizados por la Unidad Universitaria de Secuenciación Masiva y Bioinformática en el Instituto de Biotecnología de la UNAM y fueron colectadas en dos expediciones marinas realizadas en el buque oceanográfico Justo Sierra de la UNAM en los años 2015 y 2017 por miembros del CIGoM. Las muestras provienen de la región “Perdido” frente a la costa del estado de Tamaulipas y una región del sur frente al estado de Campeche. La colecta se realizó usando rosetas con botellas de Niskin y un perfilador CTD(SBE-9) acoplados para evaluar las condiciones fisicoquímicas (Fig. 5 y 6).

Los metagenomas de la columna de agua corresponden a zonas de máxima fluorescencia (MAX), mínima oxigenación (MIN) y 1,000m de profundidad (MIL). Una vez colectadas, se concentraron y se reservaron 100mL para pasarlos a través de un filtro Sterivex y otros

100mL a través de una membrana de policarbonato de 0.22µm Merck Millipore. Los filtros se almacenaron en congelación hasta que se procesaron. Las muestras de sedimentos fueron colectadas a profundidades entre 1,200m a 3,000m, se congelaron y almacenaron hasta su procesamiento (Fig. 5). Se usó el kit de aislamiento de DNA Power Water (MO BIO-QIAGEN) para las muestras de agua, mientras que para las muestras de sedimentos se usó kit DNA PowerSoil (MO BIO-QIAGEN) a partir de 0.5g de sedimento superficial. Se prepararon las bibliotecas con el kit TruSeq DNA PCR-Free HT Library Prep de Illumina® y la secuenciación de los reads paired-end se realizó con la plataforma Illumina® NextSeq500 con una configuración de 75 o 150 ciclos[51]. En los siguientes análisis únicamente se evalúa a los procariontes presentes en los metagenomas, descartando a los eucariontes, virus o viroides que pudieran están presentes, sin embargo, no hay que olvidar su importancia en los ecosistemas marinos[52].

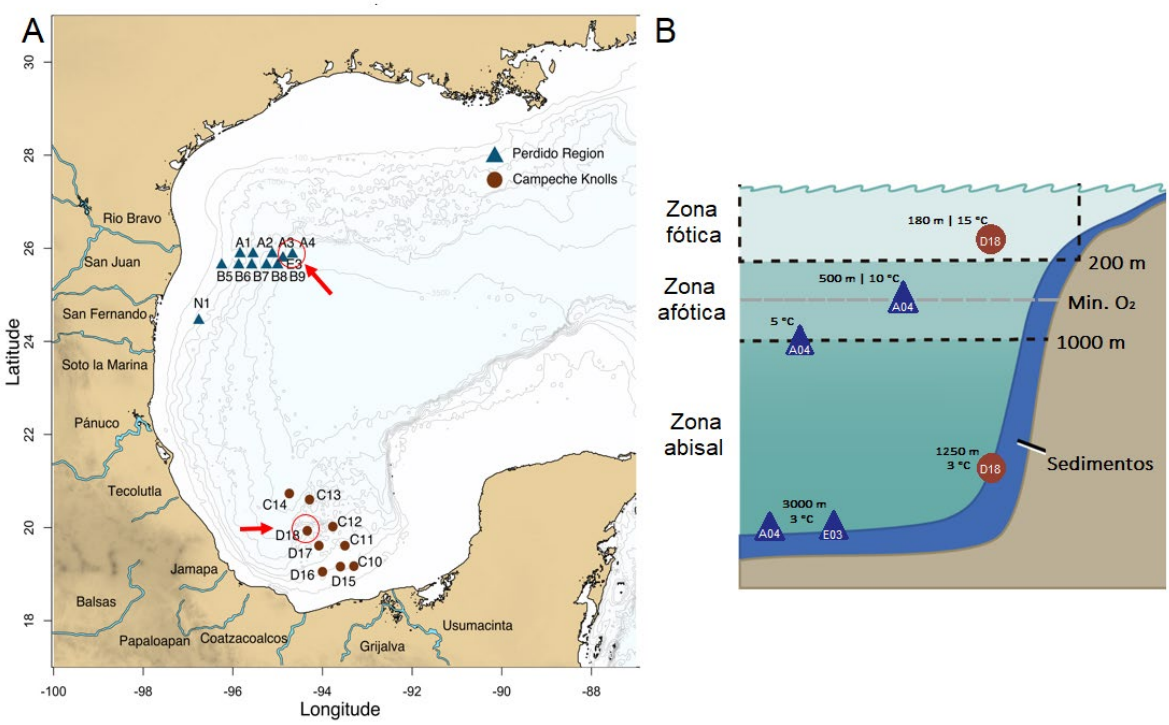


Figura 5. Ubicación de los sitios de muestreo en el Golfo de México. (A) Regiones y plataformas donde fueron colectadas las muestras. En un círculo rojo se señala las zonas específicas de las muestras que se usaron en este trabajo (Raggi *et al.*[51]). (B) Diagrama de la ubicación espacial de las muestras en la columna de agua de las tres diferentes plataformas: D18MAX en la zona de máxima florescencia, D18SED en el lecho oceánico de Campeche (círculos marrones), A04MIN en la zona de mínima oxigenación, A04MIL en la zona de 1000m de profundidad, A04SED y E03SED en el lecho oceánico de la región de Perdido (triángulos azules). Se indica en cada figura las aproximaciónes de profundidad y temperatura registradas con el perfilador (Fig. 6).

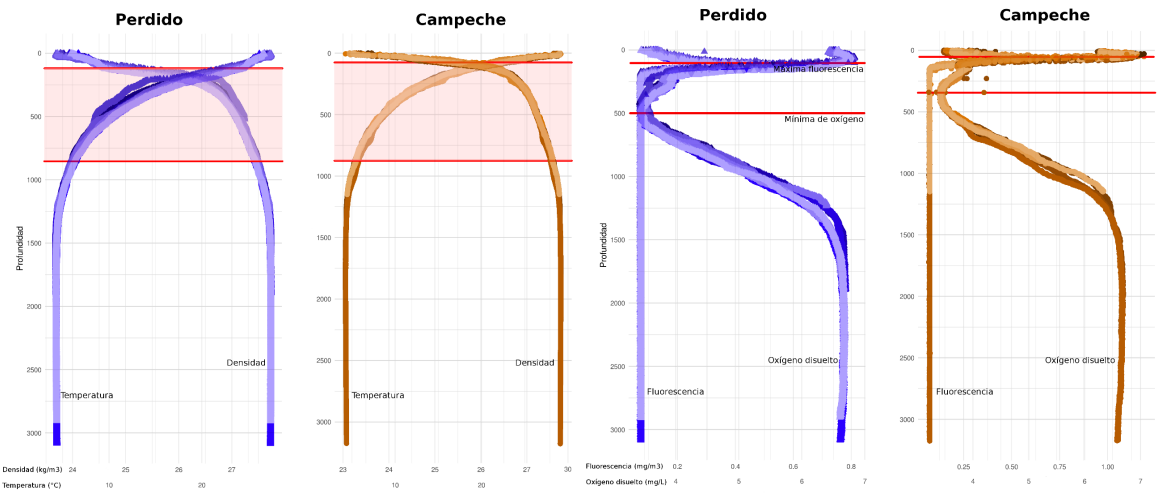


Figura 6. Clinas de las condiciones fisicoquímicas registradas por el perfilador CTD(SBE-9) durante la colecta de las muestras biológicas en el buque Justo Sierra (A. Escobar, comunicación personal).

Con la finalidad de identificar qué microorganismos están presentes en las diferentes muestras de fracciones de la columna de agua, así como de sedimentos del Golfo de México, en esta tesis se reconstruyeron MAGs utilizando una variedad de herramientas bioinformáticas. Con ello, se espera conocer más acerca de la biodiversidad del microbioma del Golfo de México y que características metabólicas presentan, un aspecto que es necesario para la comprensión del funcionamiento del ecosistema y que además representa un valioso recurso para las bases de datos genómicas.

4. Planteamiento

4.1 Justificación

Conocer la biodiversidad del microbioma del Golfo de México es necesario para la comprensión de cómo funciona dicho ecosistema. Los procariontes están estrechamente vinculados al sustento de los ecosistemas marinos y ciclos biogeoquímicos que en ellos tienen lugar. Aunado a que ya se tiene una línea base[50], contribuir con genomas reconstruidos de procariontes del Golfo de México es un valioso recurso que permitirá conocer sus capacidades metabólicas e interpretar su posible relación con el ecosistema. Así mismo, la identificación y reconstrucción de taxa nuevos o ya registrados, proporciona un recurso útil con que trabajar en proyectos de genómica comparativa.

4.2 Objetivo general

Reconstruir genomas a partir de seis metagenomas obtenidos por *whole-genome shotgun sequencing* de aislamientos de muestras marinas de agua y sedimentos del Golfo de México.

4.3 Objetivos particulares

1. Analizar la calidad las lecturas (reads) y determinar su asignación taxonómica.
2. Ensamblar las lecturas para reconstruir fragmentos (contigs).
3. Clasificar por propiedades de composición de nucleótidos y optimización de los ensambles para reconstruir los MAGs.
4. Realizar la identificación taxonómica de los MAGs y construir un árbol filogenético.
5. Realizar la anotación funcional de los MAGs y evaluar su posible papel ecológico en los metagenomas.

5. Método

5.1 *Software* y *hardware* empleado

Los procedimientos aquí descritos se realizaron en el servidor de cómputo del laboratorio 12 (FOS) del Instituto de Biotecnología de la UNAM que cuenta con 56 núcleos Intel Xeon® y 500 GB de memoria RAM. La descripción de los programas, sus versiones y respectivas referencias se presentan en la Tabla 1. Los comandos empleados para todos los análisis se resumen en el apartado 9.1 del Anexo y se detallan para quienes no cuentan con previa experiencia en el tema en el repositorio en GitHub asociado a la referencia 53.

Tabla 1. Software empleado para el procesamiento, análisis y visualización de la información.

Software (Versión)	Breve descripción	Ref.
FastQC (0.11.5)	Análisis y control de calidad de secuencias nucleotídicas almacenadas en diferentes formatos basado principalmente en el parámetro de calidad Phred que permite evaluar la cantidad de información que cumple con estándares de calidad y cual otra podría generar sesgos en posteriores análisis.	54
Kraken2 (2.0.8)	Sistema de clasificación taxonómica de reads cortos basado en la identificación de k-meros en los reads y cuya asignación ocurre mediante el algoritmo *Lowest Common Ancestor.* La clasificación de los reads se realiza a partir de bases de datos estructuradas en k-meros y comprimidas en hashes creadas a partir de genomas; por ejemplo, usando genomas de la RefSeq o cualquier otro conjunto determinado por el usuario, lo cual permite realizar la clasificación con una alta precisión y velocidad.	55
Pavian (1.0)	Plataforma interactiva para el análisis, comparación y visualización de resultados de clasificación taxonómica de reads generados por herramientas como Kraken, Centrifuge o MetaPhlAn.	56
IDBA-UD (1.1.1)	Ensamble metagenómico *de novo* e iterativo basado en gráficos de Brujin diseñado para el ensamble de reads paired-end cortos. Realiza un preprocesamiento de la calidad de los reads y se considera la cobertura/profundidad de estos para la construcción de los grafos, los cuales se construyen con diferentes longitudes de k-meros, se procesan y optimizan para generar los contigs finales.	57
Megahit (1.1.1.2)	Ensamble metagenómico *de novo* con un proceder similar a IDBA pero que implementa gráficos de Brujin sucintos en la resolución de los grafos, permitiendo un procesamiento de las secuencias ultrarrápido y eficiente en la utilización de memoria.	58
Quast (5.0.0)	Paquetería de utilidades para la evaluación y comparación de ensambles genómicos mediante el cálculo de métricas relacionadas como N50, L50, numero de contigs y pares de bases contenidas, así como longitud de los ensambles y que puede realizarse con o sin referencia genómica.	59
Minimus2 (3.1.0)	Ensamblador modular basado en Minimus diseñado para la para la unión de ensambles mediante el paradigma de solapamiento y grado de identidad de secuencias que hace uso de nucmer (NUCleotide MUMmer) como herramienta para determinar los alineamientos.	60
Bowtie2 (2.3.4.1)	Herramienta de alineamiento de secuencias cortas contra secuencias de referencia de mayor tamaño que implementa las estrategias de creación de índices a partir de estas y su compresión mediante el algoritmo de Burrows-Wheeler para el análisis de conjuntos masivos de datos y permitiendo la optimización del proceso en términos de memoria y tiempo.	61
SamTools (1.9)	Paquetería de utilidades para el post procesamiento de alineamientos de secuencias almacenados en formatos SAM, BAM y CRAM que permite la visualización, edición, creación de índices y conversión entre estos formatos.	62
Binsanity (0.2.7)	Paquetería de scripts diseñados para el procesamiento, análisis y binning automatizado de contigs generados a parir del ensamble de reads de metagenomas que implementa el algoritmo de propagación por afinidad y un enfoque bifásico. Este último considera la cobertura de los contigs como parámetro principal de agrupamiento seguido de un refinamiento basado en el contenido de GC así como frecuencias de k-meros.	63
CheckM (1.0.13)	Herramientas para la evaluación de la calidad de genomas procariontes reconstruidos a partir de aislamientos, células individuales (SAG) o metagenomas (MAGs). La calidad de determina principalmente a partir el posicionamiento de la consulta dentro de un árbol genómico de referencia, permitiendo evaluar la presencia, ausencia o duplicación de conjuntos de genes de copia única propios de un determinado linaje filogenético.	14
Rfam (14.1)	Base de datos de 2687 familias de RNA representadas por alineamientos múltiples de secuencias, estructuras secundarias consenso y modelos de covarianza. Estos últimos son perfiles probabilísticos de secuencias de RNA y sus respectivas estructuras secundarias que en conjunto permiten la identificación de secuencias homólogas en las consultas de manera precisa.	64

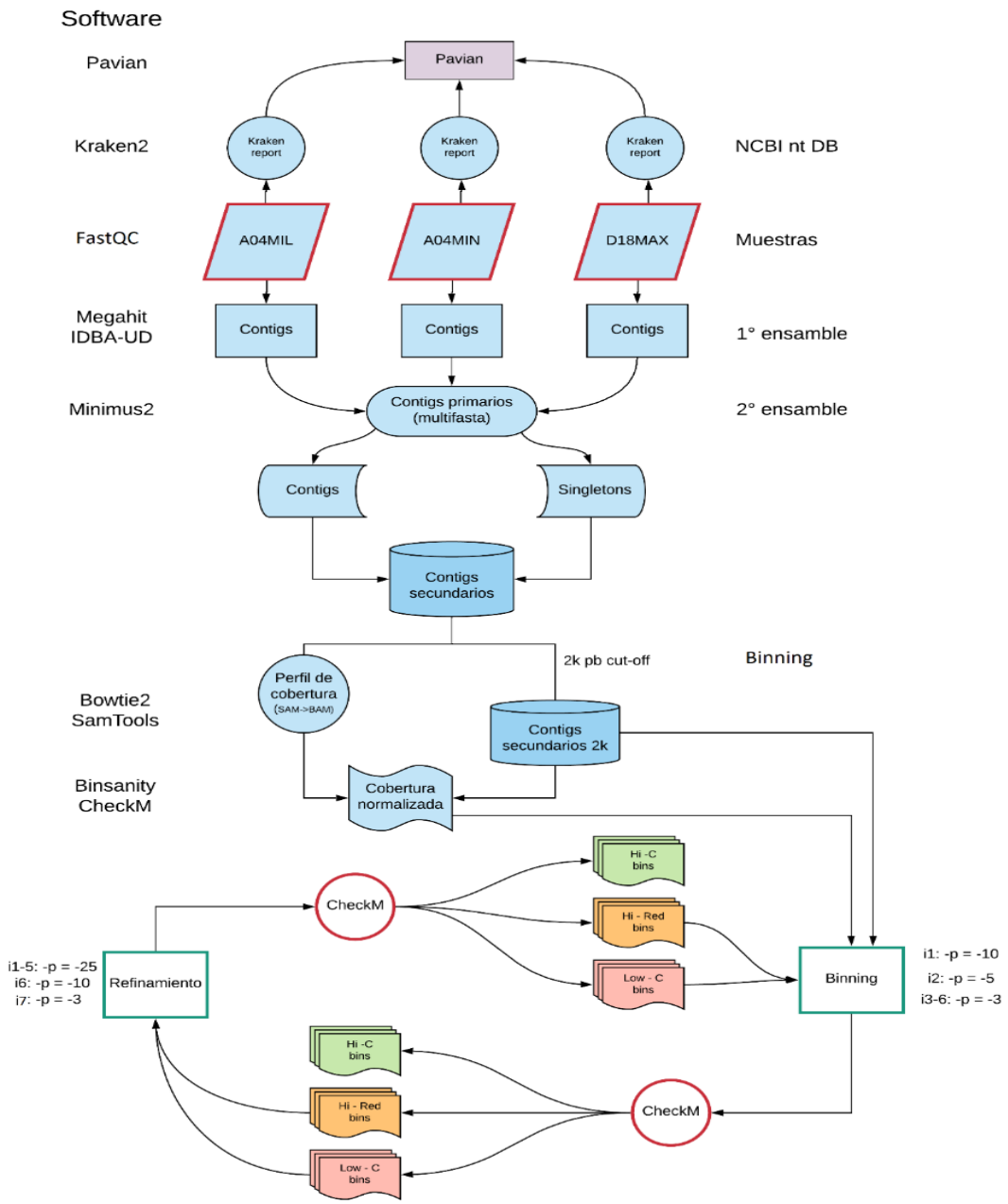
Tabla 1. Continuación.

Software (Versión)	Breve descripción	Ref.
Infernal (1.1.2)	Herramienta de búsqueda e identificación de secuencias de RNA no codificantes, sus estructuras secundarias y similitudes mediante la implementación de modelos de covarianza de Rfam y modelos ocultos de Márkov que sirven como set de entrenamiento al algoritmo y permiten el análisis, alineamiento y anotación de RNAs homólogos en secuencias de consulta.	65
tRNA-Scan-SE (2.0.2)	Herramienta especializada en la detección y anotación de tRNA en genomas de los tres dominios. Implementa el algoritmo de Infernal v1.1 y algoritmos de post procesamiento, así como modelos de covarianza específicos de tRNAs construidos a partir de 155, 4016 y 182 genomas de referencia de 110, 647 y 75 géneros de eucariontes, bacterias y arqueas, respectivamente.	66
GtoTree (1.4.2)	Pipeline automatizado para análisis filogenómicos. Por defecto, realiza la predicción de genes con Prodigal, identifica conjuntos de genes de copia única de diversos linajes filogenéticos de acuerdo con la consulta con HMMER, filtra las secuencias y genomas de mala calidad para reducir sesgos, realiza el alineamiento y trimming de las secuencias con Muscle y Trimal, respectivamente, añade información taxonómica para visualizarse en los árboles con TaxonKit y finalmente construye el árbol filogenético con FastTree.	67
GTDB-Tk (0.3.2)	Herramienta para la clasificación taxogenómica de procariontes basada en la *Genome Taxonomy Database* (GTDB) que utiliza Prodigal, HMMER, pplacer, FastTree, FastANI y Mash para procesar las consultas. En su actual versión (R04-RS89), la GTDB se compone de 145,904 genomas representativos de 1248 y 23,458 especies de arqueas y bacterias, respectivamente. Para la clasificación, utiliza una combinación de valores ANI, divergencia evolutiva relativa y la asignación de la consulta en la topología de un árbol filogenómico de referencia inferido por el alineamiento de 120 o 122 genes de copia única para bacterias y arqueas, respectivamente.	68
iTOL (4.0)	Herramienta online para la visualización, manipulación y anotación de árboles filogenéticos.	69
RStudio (3.5)	Entorno de desarrollo integrado diseñado bajo lenguaje de programación R.	70
ggplot2 (3.2)	Paquete de R para la visualización de datos basado en la gramática de los gráficos.	71
Anvi'o (5.5)	Plataforma interactiva de análisis, visualización, exploración, manipulación y reporte de datos ómicos para la caracterización de genomas microbianos que implementa diferentes pipelines que integran herramientas bioinformáticas consideradas el estado del arte actual.	72
GhostKOALA (2.2)	Servidor online de anotación automática de genomas y metagenomas que permite la caracterización de las funciones individuales de los genes mediante el análisis de las relaciones de ortología utilizando el algoritmo de búsqueda de homología GHOSTX a partir de la base de datos *Kyoto Encyclopedia of Genes and Genomes Orthology* (KEGG Orthology). Esta última representa funciones moleculares en términos de ortólogos funcionales.	73
KEGG Decoder (1.0)	Herramienta que integra y analiza las anotaciones funcionales de los genes realizadas por los algoritmos BlastKOALA, GhostKOALA o KOFAMSCAN para evaluar la integridad de vías metabólicas de acuerdo con los elementos que las conforman de acuerdo con los mapas metabólicos de la base de datos KEGG.	74
antiSMASH (5.0)	Pipeline automatizado para la Identificación, anotación y análisis de clústers de genes biosintéticos (BGC) del metabolismo secundario a partir de genomas. Infiere la presencia e identidad de entre 52 tipos deferentes de BGCs al identificar patrones de coocurrencia de enzimas conservadas clave de estos, así como su vecindad genómica utilizando perfiles curados creados con modelos ocultos de Márkov.	75

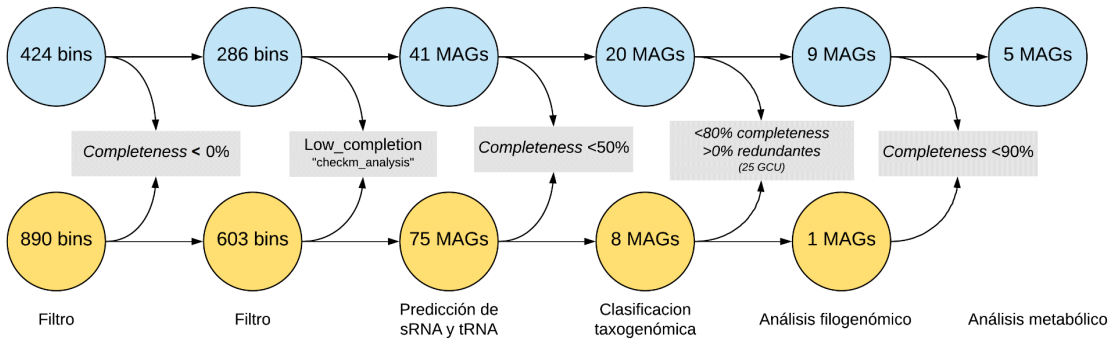
5.2 Análisis de los reads

En la Figura 7 se resumen los procedimientos realizados en el análisis de los metagenomas, desde los reads hasta los MAGs. Se evaluó la calidad de los reads para determinar si era necesario o no, un filtro de calidad en función de conservar la mayor cantidad posible de reads con valores de calidad Phred buenos (≥28). Para ello se utilizó el script *basic_stats.pl* (A. Escobar, comunicación personal) y FastQC[54] (Tab. 2; Fig. 9). Posteriormente, se analizó la distribución y asignación taxonómica de los reads para identificar la composición taxonómica general de los metagenomas y tener un atisbo de los taxa que podrían ser recuperados como MAGs. Se usó Kraken2[55] y Pavian[76] para evaluar la abundancia e identidad taxonómica de los reads usando la *non-redundant nucleotide database* (nt) del NCBI y se obtuvieron las abundancias relativas de los reads bacterianos por muestra (Figs. 10-12).

A



B



Leyenda en la siguiente pagina →

Figura 7. Diagramas del pipeline implementado para la reconstrucción y análisis de los MAGs. (A) Pipeline empleado para la clasificación, ensamble y binning de los metagenomas de la fracción de agua. Se indican en los laterales con letras los programas utilizados y procedimientos. El mismo proceder fue realizado con los respectivos contigs de la fracción de sedimentos, pero con un corte de 3kpb en la longitud mínima de los contigs secundarios debido a los requerimientos de memoria dado el tamaño del ensamble para realizar el binning. A los lados de los recuadros 'Binning' y 'Refinamiento' se indica el número de iteración (i) y el valor del parámetro de preferencia (-p) que se utilizó; a mayor valor de -p, se hace más estricta la similitud y se obtienen más clústers mientras que a menores valores de -p se hace más flexible la similitud y se obtienen menos clústers. Los valores empleados son los mismos que los autores sugieren para procesar metagenomas marinos[46]. (B) Criterios de descarte de los bins generados con Binsanity hasta los MAGs. Se señala el número de bins/MAGs según corresponda provenientes de la fracción de agua y sedimentos en esferas de color azul y amarillo, respectivamente. En los recuadros grises se señala el criterio de descarte y en la parte inferior se señala el procedimiento realizado para el conjunto de datos. Abreviaciones: GCU (Genes de Copia Única), Hi-C (*High completeness*), Low-C (*Low completeness*), Hi-R (*High redundancy*), i (iteración). Consultar la Tabla 3 para más detalles del número de contigs y porcentajes de mapeo a los metagenomas.

5.3 Reconstrucción de los MAGs

Para realizar la reconstrucción de los MAGs, se adaptó y modifico el protocolo desarrollado por Tully *et al.*[46] para el procesamiento de grandes cantidades de datos (Fig. 7A). Se realizo el ensamble *de novo* individualmente de los seis metagenomas con IDBA-UD[57] usando valores por defecto y con Megahit[58] usando los ajustes *meta, meta-large* y *meta-sensitive*; los contigs resultantes se usaron para crear un archivo multifasta según correspondieran a la fracción de agua o sedimentos y se evaluó el desempeño de los ensambladores con Quast[59], conservando los contigs del ensamblador con mejor desempeño para los siguientes análisis y a los que se refiere de aquí en adelante como contigs primarios (Fig. 13; Tab. 3). Se coensamblarón los contigs primarios de la fracción de agua y de sedimentos de forma independiente mediante ensamble por solapamiento usando Minimus2[60] con un solapamiento mínimo de 100pb e identidad mínima de 95% para cada fracción, se combinaron los contigs y singletones (contigs no coensamblados) resultantes del coensamble para formar un conjunto de contigs de mayor tamaño a los que se refiere de ahora en adelante como contigs secundarios (Tab. 3). Se utilizo Bowtie2[61] para mapear los reads de los metagenomas de la fracción de agua o sedimentos contra los contigs secundarios de las respectivas fracciones, comprimiendo y ordenando los archivos SAM resultantes con SamTools[62].

Se descartaron los contigs secundarios menores a 2,000 y 3,000 pb de la fracción de agua y sedimentos, respectivamente, con la opción *megahit_toolkit filterbylen* y se sometieron a seis ciclos de binning, evaluación y refinamiento. En breve, se realizó el binning con Binsanity[63] y se evaluó la calidad de los bins con CheckM[14], se clasificaron como *hi_completion*, *hi_redundancy* y *low_completion* con el script *checkm_analysis*[63], se conservaron los bins clasificados como *hi_completion* y se combinaron los bins de las dos categorías restantes en un solo archivo multifasta que se refino con Binsanity-refine. Este ciclo de binning, evaluación y refinamiento se realizó durante seis iteraciones y dos últimos refinamientos, cambiando los valores del parámetro de preferencia (-p) de Binsanity el cual refleja el grado en que el algoritmo de propagación por afinidad agrupa o divide los contigs en bins. Se generaron un total de 424 y 890 bins a partir de los contigs secundarios de la fracción de agua y sedimentos, respectivamente; se descartaron los bins de mala calidad y se conservaron únicamente 41 bins de la fracción de agua y 75 bins de la fracción de sedimentos referidos de ahora en adelante como a MAGs (Fig. 7B).

Se adaptó la metodología desarrollada por de Almeida *et al*.[29] para la identificación de la presencia de los genes ribosomales 5S, 16S y 23S en los MAGs. En breve, se ordenaron los MAGs en nuevos directorios según correspondieran a bacterias o arqueas, se descargaron los respectivos modelos de covarianza desde la Rfam[64] de los genes codificantes de las subunidades ribosomales 5S (RF00001), 16S (Bacteria: RF02541; Arquea: RF02540), 23S (Bacteria: RF00177; Arquea: RF01959) y se realizó la predicción con Infernal[65,77], considerándolos presentes en los MAGs siempre y cuando presentaran ≥80% de la longitud de la secuencia esperada en el MAG (Longitudes totales esperadas: Arquea 5S = 119pb, 16S = 1477pb, 23S = 2990pb; Bacteria 5S = 119pb, 16S = 1533pb, 23S = 2925pb). La predicción de los tRNAs canónicos se realizó con tRNAscan-SE[66] usando los respectivos modelos de covarianza para bacterias (-B) y arqueas (-A) y con parámetros por defecto (A. Almeida, comunicación personal)[29]. Finalmente, se clasificaron los MAGs en categorías de alta, mediana y baja calidad de acuerdo a los estándares de MIMAG[13] e incluyendo las categorías

"casi completos" y sub categorías del grupo de calidad media basadas en valores de *Quality Score* sugeridas por Parks *et al.*[47] (QS = *Completeness* – (5 X contaminación)) (Fig. 14).

5.4 Análisis de los MAGs

Se descartaron todos los MAGs con completeness <50% de ambas fracciones para realizar la asignación taxonómica, conservando un total de 20 y ocho MAGs de la fracción de agua y sedimentos, respectivamente (Fig. 7B). Aunque CheckM proporciona una asignación taxonómica, no es un software desarrollado para tales fines. Por ello, se utilizó GTDB-Tk[68], una herramienta automatizada de clasificación taxogenómica de procariontes basada en la *Genome Taxonomy Data Base* (GTDB)[78]. Para evaluar las asignaciones taxogenómicas de los MAGs inferidas por GTDB-Tk, se utilizó GToTree[67,79] usando los parámetros por defecto para construir un árbol filogenómico basado en la concatenación un conjunto de 25 genes de copia única de bacterias y arqueas con los 28 MAGs reconstruidos a partir de los seis metagenomas y los genomas de referencia inferidos por GTDB-tk. Finalmente, el árbol resultante se editó y visualizo en iTOL[69] (Fig. 15).

De todos los MAGs reconstruidos, solo cinco provenientes la columna de agua tienen un completeness y contaminación inferidos por CheckM de >90% y <10%, respectivamente, (Tab. 4); solo con dichos MAGs se realizaron los siguientes análisis. Para evaluar la abundancia de los MAGs en los metagenomas, se realizó un análisis del mapeo de reads de los contigs de los MAGs con el pipeline metagenómico de Anvi'o[72] implementado en SnakeMaKe[80] (Fig. 16; Ver Anexo apartado 9.2). Para identificar el potencial metabólico de los MAGs seleccionados, se obtuvieron sus respectivas secuencias de aminoácidos con Anvi'o, se anotaron los genes usando GhostKOALA[73] y se organizaron y visualizaron en 162 rutas metabólicas diferentes con KEGG-Decoder[74] (Fig. 17). Para obtener más información de las anotaciones de las secuencias codificantes de los MAGs, además de las anotaciones realizadas usando la base de datos *Clusters of Orthologous Groups* (COG) con el pipeline de Anvi'o, se anotaron las secuencias usando la base de datos *Pfam* (v32.0) y se importaron las anotaciones de los genes realizadas con GhostKOALA (basadas en la base de datos *KEGG*) a las respectivas bases de datos de los MAGs realizadas con Anvi'o (Fig. 8).

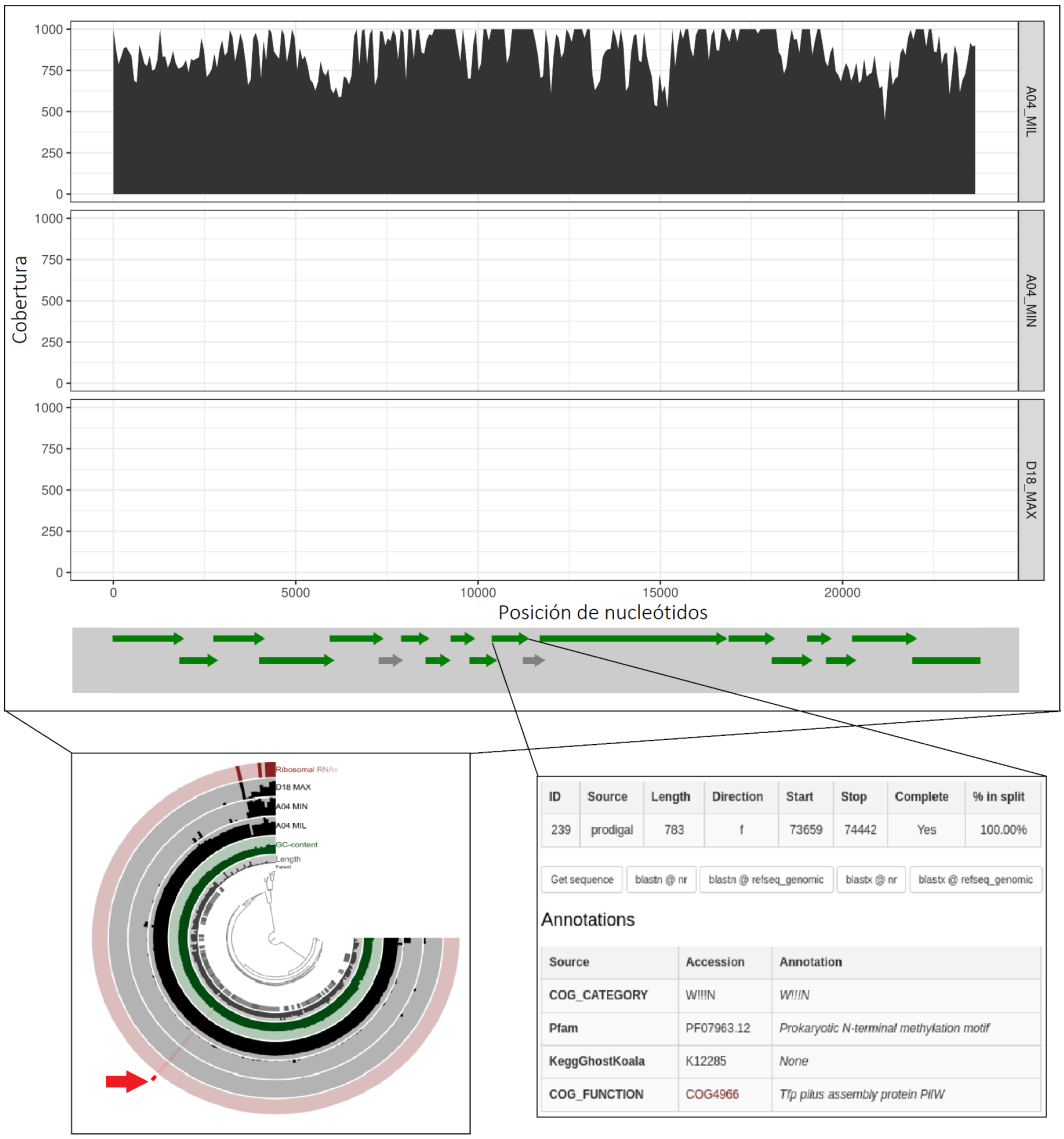


Figura 8. Ejemplo de la visualización de las secuencias codificantes anotadas en Anvi'o. Se indica con una flecha roja en el diagrama circular (desarrollados en la Figura 16A) la división en la que se encuentra el gen con el identificador 239 (ID) con sus respectivas anotaciones con las bases de datos COG, KEGG y Pfam en el panel a la derecha. En el panel superior se muestra la cobertura en los tres metagenomas de la división que contiene al gen 239 así como otros genes anotados (flechas verdes) y no anotados (flechas grises) con ninguna de las tres bases de datos. Esta estrategia es útil para poder capturar una mayor información respecto función de los genes, nótese como las anotaciones realizadas con KEGG no con conclusivas, pero con COG y Pfam sí lo son.

6. Resultados y discusión

6.1 Clasificación de los reads

En esta tesis se evaluó la composición taxonómica de seis metagenomas provenientes de la columna de agua y de sedimentos marinos colectados en el Golfo de México. Además, se realizaron análisis metabólicos y filogenómicos en un pequeño grupo de MAGs reconstruidos a partir de los metagenomas, sin embargo, hay que considerar que los MAGs no representan la composición taxonómica completa de los metagenomas, pero si reflejan a los microorganismos con mayor prevalencia dentro de la comunidad microbiana, un aspecto que es aún más remarcado con el MAG PASS1-11 discutido adelante.

Tabla 2. Estadísticas generales de los reads de los metagenomas (determinadas con *basic_stats.pl*).

Metagenoma	Reads totales	Bases totales	Tamaño de los reads (pb)	GC (%)	Calidad promedio (Phred score (Q))	Q20 (%)	Q30 (%)
A04MIL	144 635 708	10 847 678 100	75	51.3	33.98	99.88	91.43
A04MIN	83 327 042	6 249 528 150	75	44.15	34.28	99.9	92.64
D18MAX	77 396 766	5 804 757 450	75	49.59	33.86	99.89	90.97
A04SED	147 151 128	22 072 669 200	150	63.8	30.38	98.02	62.17
D18SED	162 961 050	24 444 157 500	150	53.11	30.39	97.93	62.93
E03SED	208 581 878	15 643 640 850	75	59.75	32.56	98.86	80.39

La composición, distribución y función de las comunidades microbianas se regionaliza en función de una plétora de factores bióticos, abióticos y sus interacciones[81], haciendo de los océanos uno de los ambientes más difíciles de analizar[82]. Entre dichos factores, la temperatura ha demostrado ser el principal factor que determina la composición microbiana hasta la zona fótica[35]; la cual abarca los primeros 200m de profundidad y donde es más común encontrar microorganismos fotoautótrofos[45]. Sin embargo, variables como la presencia de macrofauna o flora marina, fuentes de energía (aire, corrientes marinas, fosas termales), latitud, pH, salinidad, concentración de O_2 disuelto, etcétera, también tienen efecto y sobre todo a mayores profundidades (zona afótica o mesopelágica, 200m a 1000m de profundidad)[83,84]. En estas regiones, los microorganismos quimiótrofos que usan compuestos orgánicos e inorgánicos como fuente de energía empiezan a ser más comunes.

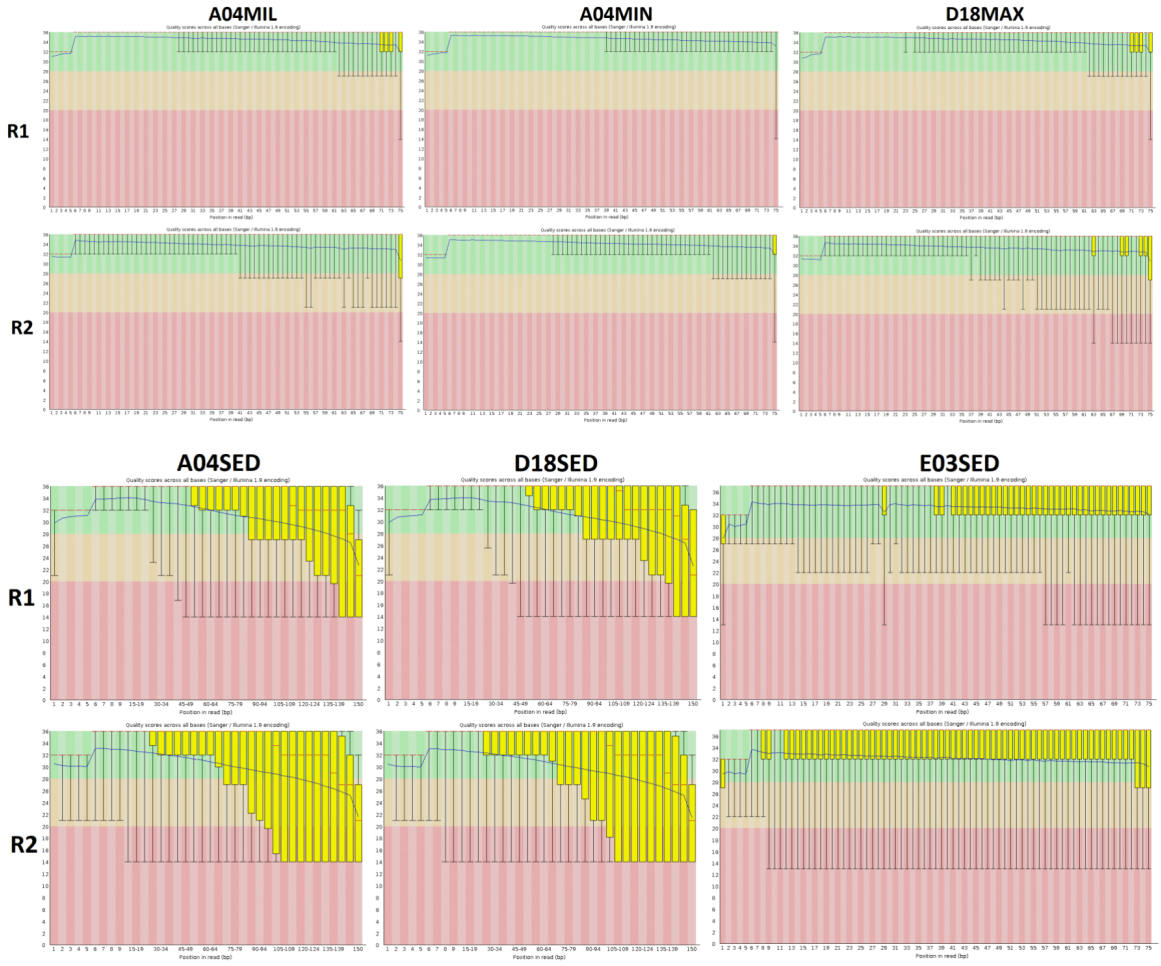


Figura 9. Calidad Phred de los metagenomas de la fracción de agua y sedimentos. El eje de las ordenadas representa la calidad Phred de los reads mientras que en el eje de las abscisas se representa por cada posición la dispersión de los valores de calidad Phred en rangos intercuartiles (25% y 75%) en cajas de color amarillo y los valores mayores y menores en los bigotes (10% y 90%). La línea roja punteada representa la mediana mientras que la línea azul representa la media.

Por ejemplo, las zonas de mínima oxigenación son regiones anóxicas consideradas sumideros globales de nitrógeno y metano en donde predominan microorganismos con metabolismo anaerobio (o anaerobio facultativo)[45,85]. Los océanos profundos son un hábitat único y extremo donde la vida ocurre en cámara lenta pues se favorece el mínimo gasto de energía al ser un ambiente extremadamente oligotrófico y que además se caracteriza por una alta presión, baja temperatura, falta de luz, salinidad y concentración de O_2 variables[86]. En la parte más profunda de los océanos, los sedimentos se constituyen como uno de los hábitats más extremos por su carácter oligotrófico y niveles de presión, sin embargo, los primeros 50cm de los sedimentos profundos contienen ~10E29 microorganismos[45], los cuales usan

bicarbonatos, sulfatos y nitratos como fuente de energía[87]. Para este tipo de especies a las que el movimiento a ambientes más favorables es imposible, la adaptación (evolución) surge como mecanismo de supervivencia, y por ello, microorganismos con poblaciones grandes, plasticidad genómica y tiempos de reproducción rápidos tienen un alto potencial de adaptación, promoviendo ecotipos que ocupan un nicho especifico[45] y surgimiento de interacciones ecológicas complejas como la sintrofía, permitiendo el intercambio de fuentes de carbono y allanando el camino para la adaptación y colonización de nuevos ambientes[88].

Para poder evaluar la composición taxonómica de los seis metagenomas, primero fue necesario analizar la calidad de los reads con los que se realizaron todos los procesos posteriores. Se determinó una calidad Phred promedio >30 para todos los metagenomas y así mismo, los valores de la mediana en general se encuentran en valores de Phred >30, tanto para los metagenomas de agua como de sedimentos, aunque estos últimos con una mayor dispersión de los datos respecto a los primeros (Tab. 2; Fig. 9). Debido a la buena calidad de los reads en general, se determinó que no era necesario realizar un filtro de calidad de las secuencias. La posterior clasificación de los reads usando la base de datos nt como referencia determinó que la mayoría de los reads de los seis metagenomas (>60%) no pudieron ser clasificados. Sin embargo, de los que pudieron ser clasificados, la mayoría corresponde a procariontes (Fig. 10).

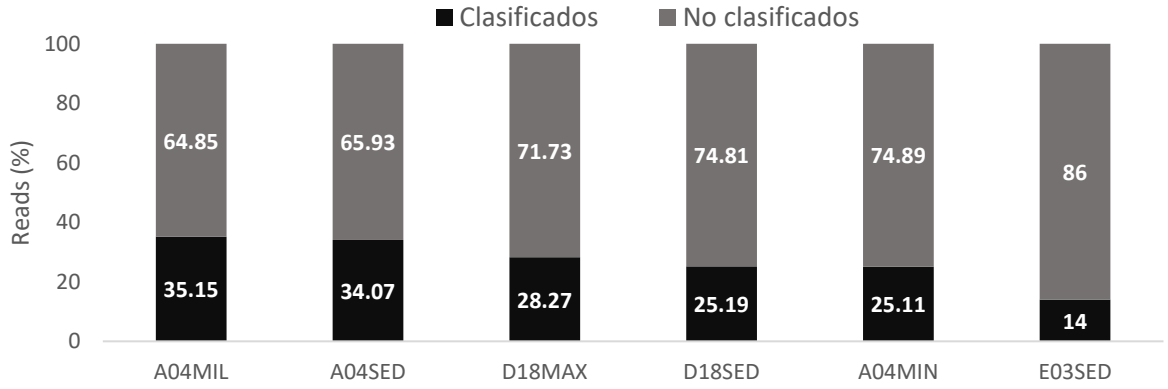


Figura 10. Porcentajes de clasificación taxonómica de los reads contra la base de datos nt. La clasificación completa por categoría taxonómica se encuentra en las Tablas suplementarias 1 y 2 [53]. Dentro de los reads clasificados, la mayoría se identificó como procariontes, aunque también se identificaron 1.4, 1.53, 1.9, 3.12, 5.18 y 6.28 millones de reads de eucariontes en los metagenomas D18MAX, A04MIN, A04MIL, E03SED, A04SED Y D18SED, respectivamente. Además, se identificaron 56.2k, 61k, 69.6k 100k, 124k y 165k de reads de virus en los metagenomas A04MIL, E03SED, D18MAX, A04SED, D18SED Y A04MIN, respectivamente.

Al respecto, las arqueas de la fracción de sedimentos son bastante escasas en relación con las secuencias de bacterias que representan <3% y >95% de los reads, respectivamente. Del pequeño porcentaje de reads de arqueas, los principales phyla representados son Euryarchaeota y Thaumarchaeota y dentro de este último, el orden de los Nitrosopumilales (~1.03% de los reads) es el grupo mejor representado. Por su parte, las bacterias se encuentran representadas principalmente por Proteobacterias, representando en general ≥50% de los reads en los tres metagenomas. Aunque también es notable la presencia de las Actinobacterias, Firmicutes, Bacteroidetes y Cyanobacterias. Dentro de estos grupos *Pseudomonas*, *Streptomyces*, *Bradyrhizobium*, *Azospirillum* y *Burkholderia* son los géneros mejor representados en los tres metagenomas de sedimentos (con una representatividad entre 2.2% a 6.8% de los reads).

En general, en los tres metagenomas de sedimentos se encontraron 2166 géneros diferentes de procariontes, 148 de los cuales corresponden al grupo de las arqueas. En los metagenomas de la fracción de agua se identificaron 1722 géneros de procariontes, de los cuales 120 son arqueas. A nivel de filo dentro del grupo de las bacterias, las Proteobacterias nuevamente son las mejores representadas (>37% de los reads en los tres metagenomas) y de manera muy marcada en el metagenoma A04MIL, donde representan el 76.06% de los reads. Otro filo abundante y casi exclusivo del metagenoma D18MAX son las Cyanobacterias, representando el 34.55% de los reads, mientras que en A04MIN y A04MIL representa tan solo el 2.14 y 0.36%, respectivamente. Otros phyla considerablemente representados en los metagenomas de la fracción de agua son Actinobacteria, Firmicutes y Bacteroidetes y dentro de ellos, existen distintos géneros muy bien representados en diferentes metagenomas: *Vibrio* en A04MIL (41.54% de los reads), *Synechococcus* en D18MAX (33.76%) y *Alteromonas* en A04MIN (29.05%); Aunque también se identificaron otros géneros no tan marcadamente abundantes en los metagenomas, como *Marinobacter*, *Ralstonia*, *Candidatus Pelagibacter*, *Prochlorococcus*, *Pseudomonas*, *Agrobacterium*, *Escherichia*, *Tistrella*, *Alcanivorax*, *Sulfitobacter*, *Streptomyces*, *Candidatus Thioglobus*, *Candidatus Actinomarina* y *Halomonas* (Fig. 11).

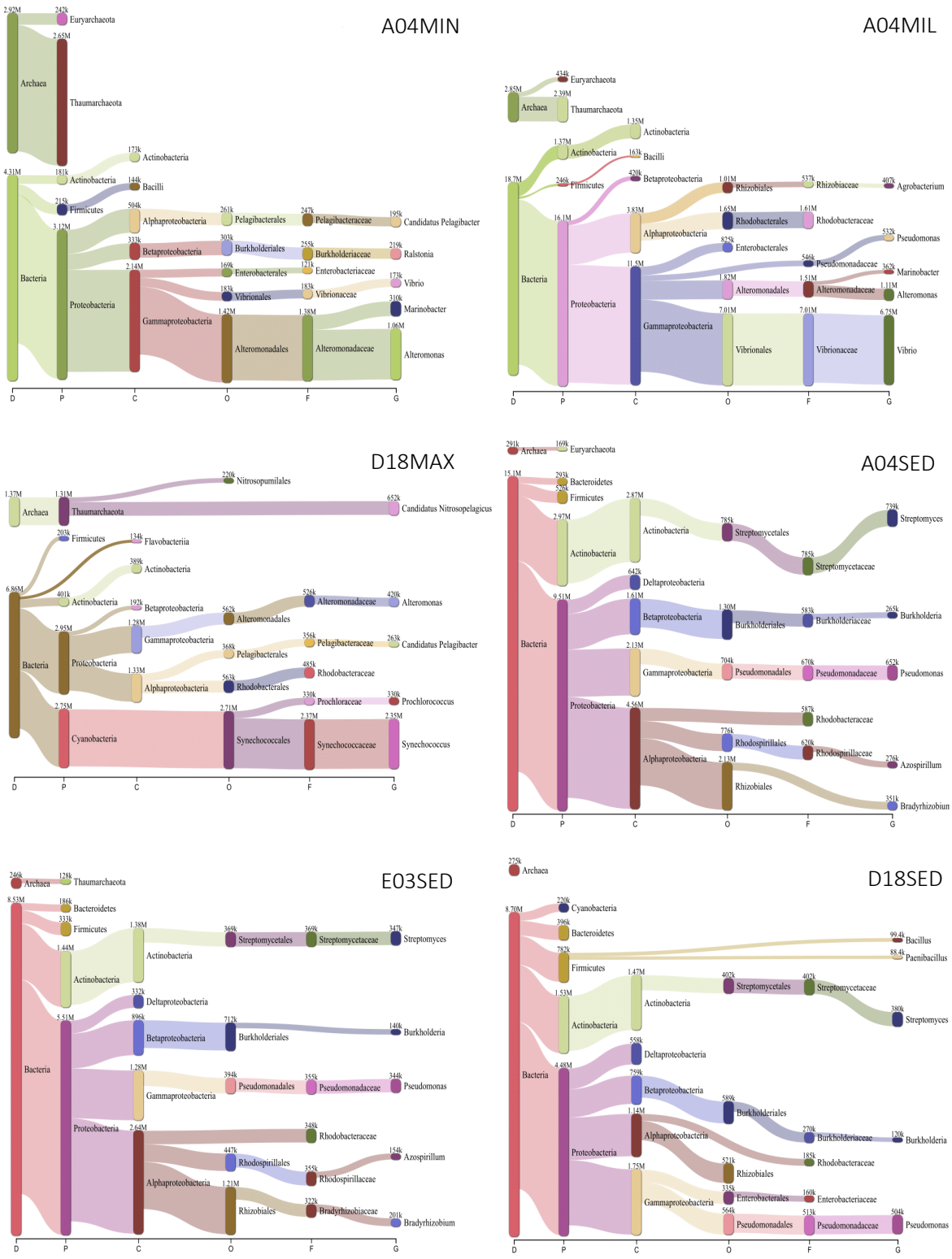


Figura 11. Diagramas aluviales de la composición taxonómica y sus abundancias de los seis metagenomas definidas usando Kraken2 contra la base de datos nt. En el eje de las ordenadas se representan las categorías taxonómicas Dominio (D), Phyla (P), Clase (C), Orden (O), Familia (F), Género (G). En el eje de las ordenadas se representa la con barras la abundancia de los taxa indicando el número de reads identificados, representado a los grupos con mayores abundancias con conexiones más anchas.

A diferencia de los metagenomas de sedimentos, las arqueas se encuentran bien representadas en la columna de agua, principalmente el filo Thaumarchaeota representando el 39%%, 16.54% y 11.38% de los metagenomas A04MIN, D18MAX y A04MIL, respectivamente; aunque la presencia del filo Euryarchaeota también es notable a exceptas del metagenoma D18MAX, donde solo representa el 0.37 de los reads calificados. Dentro de estos dos phyla, los géneros mejor representados en los metagenomas son *Candidatus Nitrosopelagicus*, *Candidatus Nitrosomarinus* y *Nitrosopumilus*, aunque es notable que solo se encuentran bien representados en el metagenoma D18MAX (Fig. 11). Para ilustrar mejor la abundancia de los géneros bacterianos en los metagenomas y el rendimiento de la clasificación de reads, en la Figura 12 se despliegan las abundancias relativas de los taxa, las cuales, permiten visualizar que una parte considerable en los metagenomas de la fracción de agua y la mayor parte de los metagenomas de sedimentos aún no han sido registradas usando la base de datos nt del NCBI como referencia.

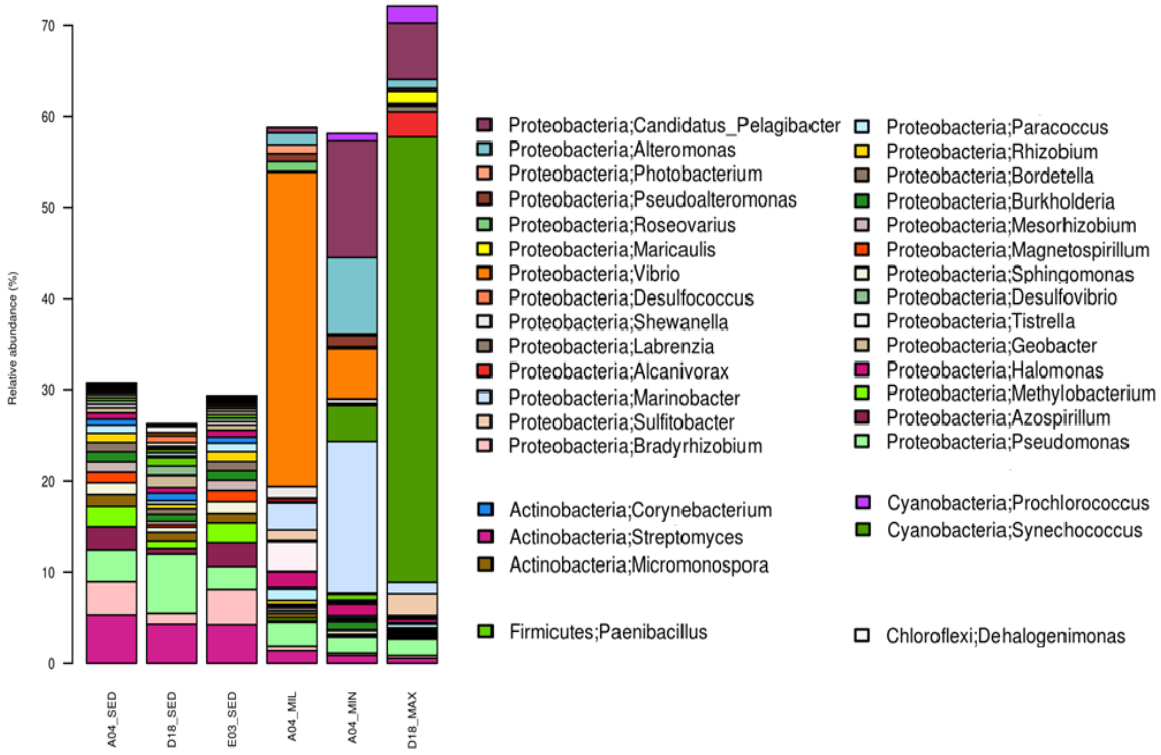


Figura 12. Abundancias relativas de los taxa bacterianos identificados usando Kraken2 con la nt (A. Escobar, Comunicación personal). En el eje de las abscisas se despliegan los seis metagenomas y en las ordenadas los porcentajes de abundancias relativas junto con las leyendas para cada taxón a nivel de filo y género. La lista con las abundancias y taxa completa se encuentra en las Tablas suplementarias 1 y 2 [53].

6.2 MAGs reconstruidos

Los ensambladores metagenómicos tienen distintos rendimientos, siendo los más actuales los que tienden a mostrar un mejor desempeño generalmente, sin embargo, también dependen de la calidad de la secuenciación, la complejidad de las muestras y demás factores [9,10,89–91]. En este sentido se realizó la breve comparación entre Megahit e IDBA-UD, demostrando que este primero logro ensamblar más pares de bases en contigs de mayores tamaños con el ajuste '*meta-senstive*', razón por la cual fue seleccionado para los análisis a *posteriori*, refiriéndose a tales ensambles como contigs primarios (Fig. 13).

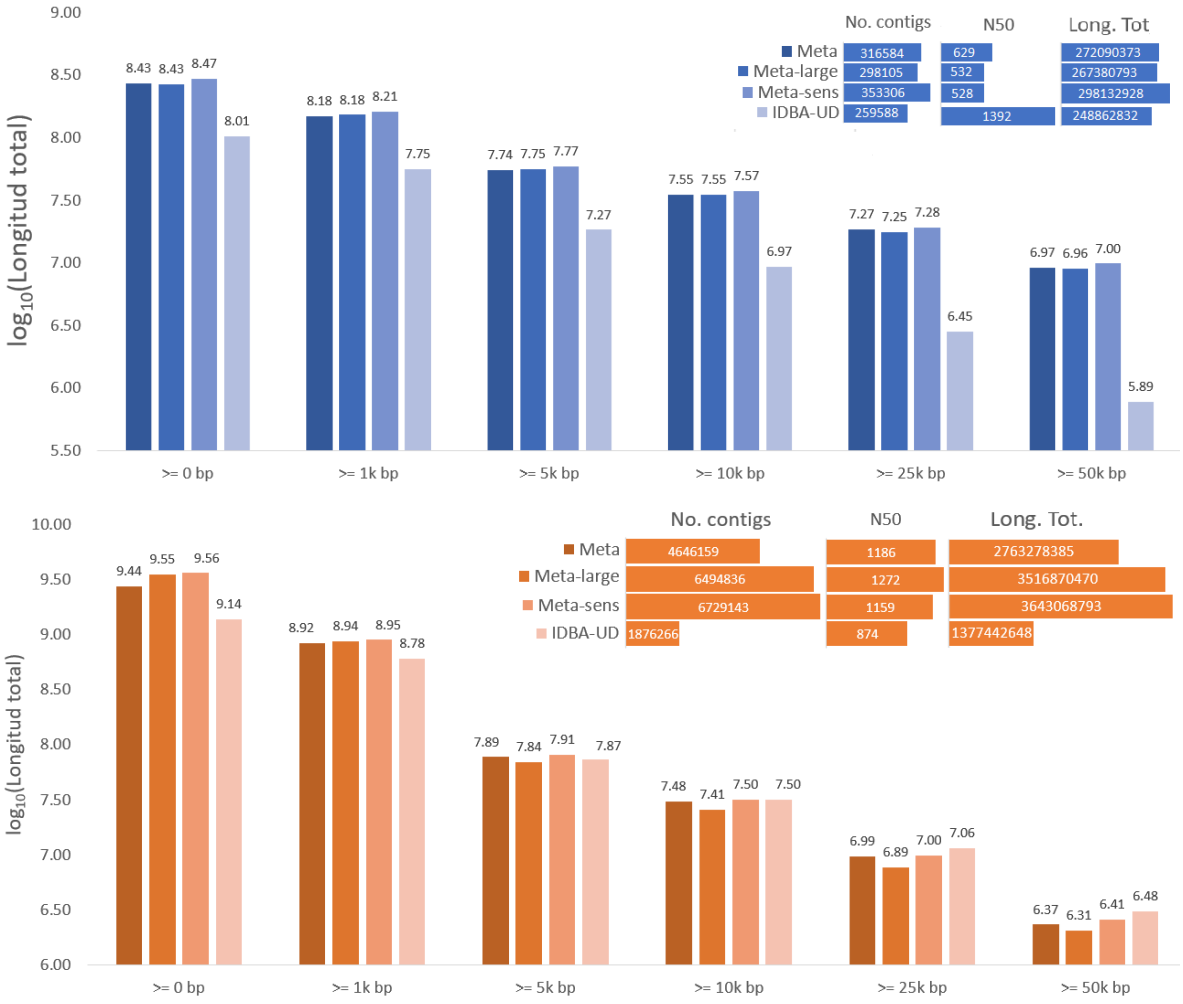


Figura 13. Comparativa del desempeño de ensamble entre IDBA-UD y Megahit en los metagenomas de la fracción de agua (en azul) y sedimentos (en marrón). Se representa en el eje de las ordenadas el $\log_{10}$ del número de pares de bases contenidas en contigs con tamaño ≥0pb, ≥1k pb, ≥5k pb, ≥10k pb y ≥50k pb (eje de las abscisas). En la esquina superior derecha se muestra el número de contigs, N50 y la longitud total de los contigs ≥0pb.

Tabla 3. Número y tamaño de los contigs primarios, secundarios y de los contigs sometidos a binning. Se muestra en paréntesis los porcentajes de mapeos determinados con Bowtie2 contra sus respectivos metagenomas. Notar, especialmente los porcentajes de mapeo de los contigs sometidos a binning.

Ensamble	≥ 0 bp (% de mapeo)	≥1k bp	≥5k bp	≥10k bp	≥ 25k bp	≥50k bp
Fracción de agua						
Contigs primarios	353,306 (54.78)	61,568	4,792	1,627	395	119
Contigs secundarios	303,322 (54.42)	57,367	5,230	1,768	412	120
Binning	21,011 (41.70)					
Fracción de sedimentos						
Contigs primarios	6,729,143 (39.49)	530,240	9,334	1,828	257	40
Contigs secundarios	5,868,011 (39.21)	515,236	12,815	2,536	350	64
Binning	44,093 (10.52)					

El proceso de coensamble para generar los contigs secundarios con Minimus2 demostró ser de utilidad al estructurar contigs de mayor tamaño como puede verse en la Tabla 3 con los contigs de longitud ≥5kpb, tanto para la fracción de agua como la de sedimentos. La longitud y numero de los contigs son factores importantes en el procedimiento de binning debido a que el requerimiento de memoria RAM de los algoritmos suele ser exponencial, particularmente con el algoritmo de propagación por afinidad implementado en Binsanity, ~100k contigs requieren de ~1TB de memoria RAM[63]; además, los algoritmos pueden clasificar erróneamente los contigs de tamaños pequeños[92], razón por la cual es recomendable someter contigs con longitudes ≥2000pb[15,93].

Los ciclos de binning y refinamiento implementados generaron 1314 bins, de los cuales no todos representan MAGs, razón por la cual se realizaron descartes hasta conservar un total de 116 MAGs en los que se realizó la detección del aparato ribosomal (Fig. 7B). De acuerdo con los estándares de calidad MIMAG y los valores de completeness y contaminación determinados con CheckM, ninguno de los MAGs reconstruidos puede ser considerado de alta calidad debido la ausencia de alguna de las subunidades ribosomales o la presencia de <18 tRNAs, sin embargo, cinco MAGs cuentan con un completeness >90% y contaminación <10%, los cuales pueden ser considerados para realizar análisis metabólicos (Fig. 14; Tab. 4).

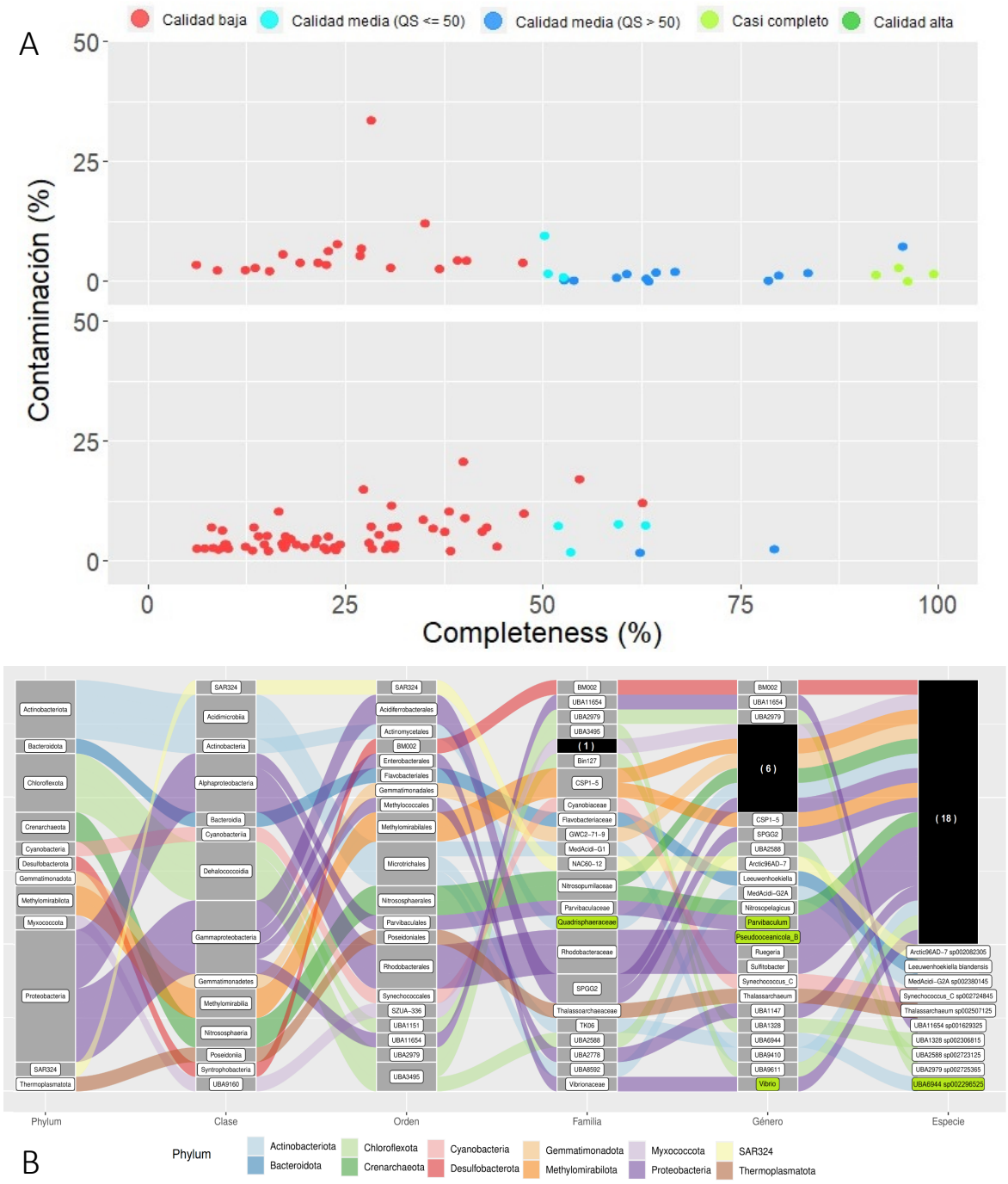


Figura 14. Clasificación taxonómica y de calidad de los MAGS. (A) Calidad de los MAGs de acuerdo con los estándares MIMAG. El panel superior corresponde a los MAGs de la fracción de agua mientras que el panel inferior a los MAGs de la fracción de sedimentos. Las figuras se realizaron con el script *mags-quality_extfig2.R* del repositorio https://github.com/Finn-Lab/MGS-gut[29]. (B) Diagrama aluvial de la clasificación taxonómica basada en la GTDB de los 28 MAGs de calidad media y casi completos. Se señala en negro el número de linajes sin previo registro en la GTDB_R04_RS89 y en verde los MAGs con *completeness* y contaminación >90% y <10%, respectivamente. Ambas figuras se realizaron usando ggplot2[71] dentro del entorno de desarrollo de R-studio[70]. La información genómica detallada de todos los bins y los MAGs se encuentra en las Tablas Suplementarias 3 y 4 asociadas al repositorio en GitHub de la presente tesis[53].

Tabla 4. Características detalladas de los MAGs seleccionados para análisis metabólico y de mapeo.

MAGs	PASS1-11	PASS1-refine-98	PASS1-refine-97	PASS1-refine-69	PASS1-refine-95
Características genómicas generales					
Completeness (%)	99.21	95.93	95.3	94.78	91.91
Contaminación (%)	1.53	0	7.26	2.8	1.34
Strain heterogeneity (%)	0	0	20	0	0
5S (%)	68.91	0	94.96	0	0
23S (%)	99.86	0	99.79	0	0
16S (%)	86.76	0	99.67	0	0
Número de tRNAs	17	18	16	19	20
Quality Score (%)	91.56	95.93	59	80.78	85.21
Tamaño (Mpb)	4.35	3.43	2.36	2.91	4.93
Número de contigs	127	244	186	114	127
GC (%)	48.9	73.8	52.8	62.9	66.7
Número de genes	4051	3495	2444	2870	4802
Clasificación taxonómica según la GTDB					
Filo	Proteobacteria	Actinobacteriota	Actinobacteriota	Proteobacteria	Proteobacteria
Clase	Gammaproteobacteria	Actinobacteria	Acidimicrobiia	Alphaproteobacteria	Alphaproteobacteria
Orden	Enterobacterales	Actinomycetales	Microtrichales	Parvibaculales	Rhodobacterales
Familia	Vibrionaceae	Quadrisphaeraceae	TK06	Parvibaculaceae	Rhodobacteraceae
Genero	Vibrio	__	UBA6944	Parvibaculum	Pseudooceanicola_B
Especie	__	__	UBA6944 sp002296525	__	__

La clasificación taxogenómica realizada con GTDB-tk solo considero los MAGs con ≥50% completeness, particularmente, se evaluaron 20 y ocho MAGs de la fracción de agua y sedimentos, respectivamente (Fig. 14B). Los ocho MAGs de la fracción de sedimentos fueron clasificados por el método de asignación en el árbol de referencia de la GTDB y representan una familia, cinco géneros y ocho especies no registradas en la GTDB. De los ocho MAGs, solo cuatro cuentan con genomas de referencia determinados por valores de *Average Nucleotide Identity* (ANI) y solo un MAG fue clasificado como Arquea. Respecto a los 20 MAGs de la fracción de agua, 10 fueron clasificados por el método de asignación en el árbol de referencia de la GTDB mientras que los 10 MAGs restantes por el método de asignación en el árbol de referencia y con soporte con valores de ANI de sus genomas de referencia. En su conjunto, los 20 MAGs de la fracción de agua representan un género y 10 especies no registradas en la GTDB y solo dos fueron clasificados como arqueas (Tablas Suplementarias 3 y 4[53]). Además, 19 de los MAGs cuentan con genomas de referencia con valores de ANI >80% y solo el MAG

PASS1-refine-98 no cuenta con genoma de referencia inferido por ANI. Considerando el conjunto de 25 genes de copia única para bacterias y arqueas[67] empleado en la construcción del árbol filogenómico, solo 10 MAGs de los 28 clasificados con GTDB-Tk cuentan con un completeness y contaminación ≥80% e igual a 0%, respectivamente. Para evitar el mayor sesgo posible debido a la contaminación, se descartaron todos los MAGs que no cumplieran con tales características. Como se esperaría, los 10 MAGs evaluados con este enfoque se agruparon con sus respectivos genomas de referencia, incluyendo el MAG PASS1-refine-98 que debido a que representa un género sin previo registro en la GTDB no cuenta con referencias inferidas por ANI, razón por la cual se seleccionaron los únicos dos genomas disponibles de familia Quadrisperaceae. Si bien las ramas muestran un buen soporte para la mayoría de los taxa, la politomía entre los phyla Chloroflexota, Actinobacteriota y Proteobacteria sería un reflejo de la falta de diversidad analizada para el filo Chloroflexota.

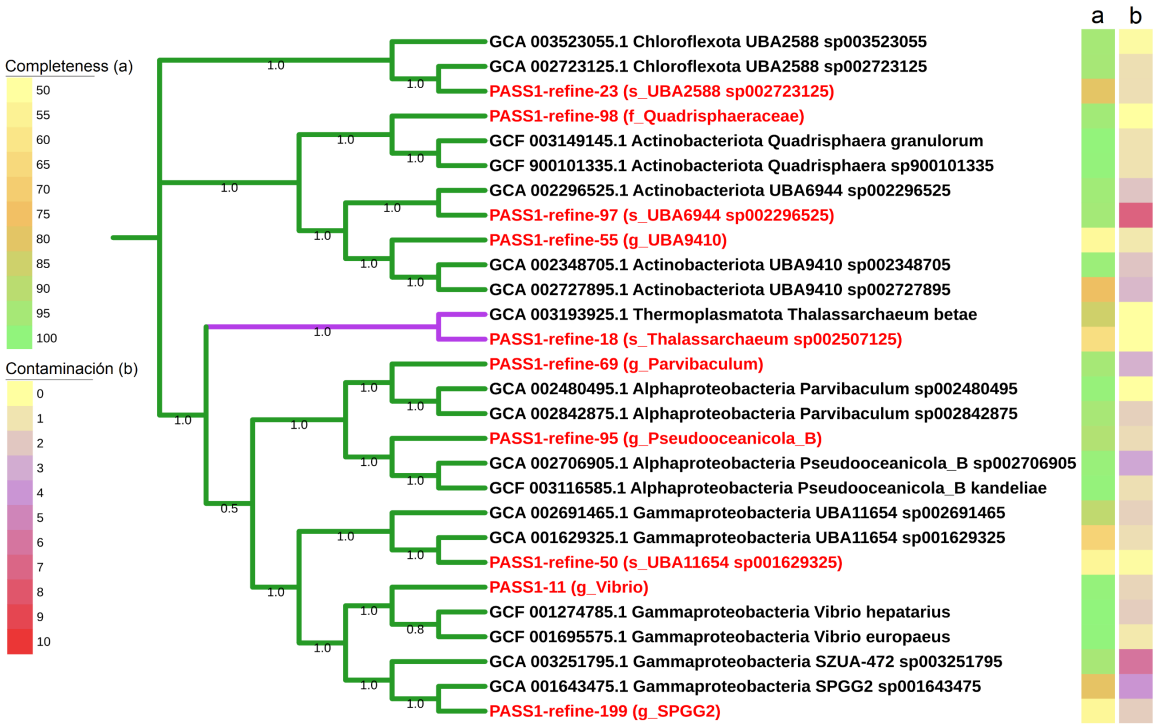
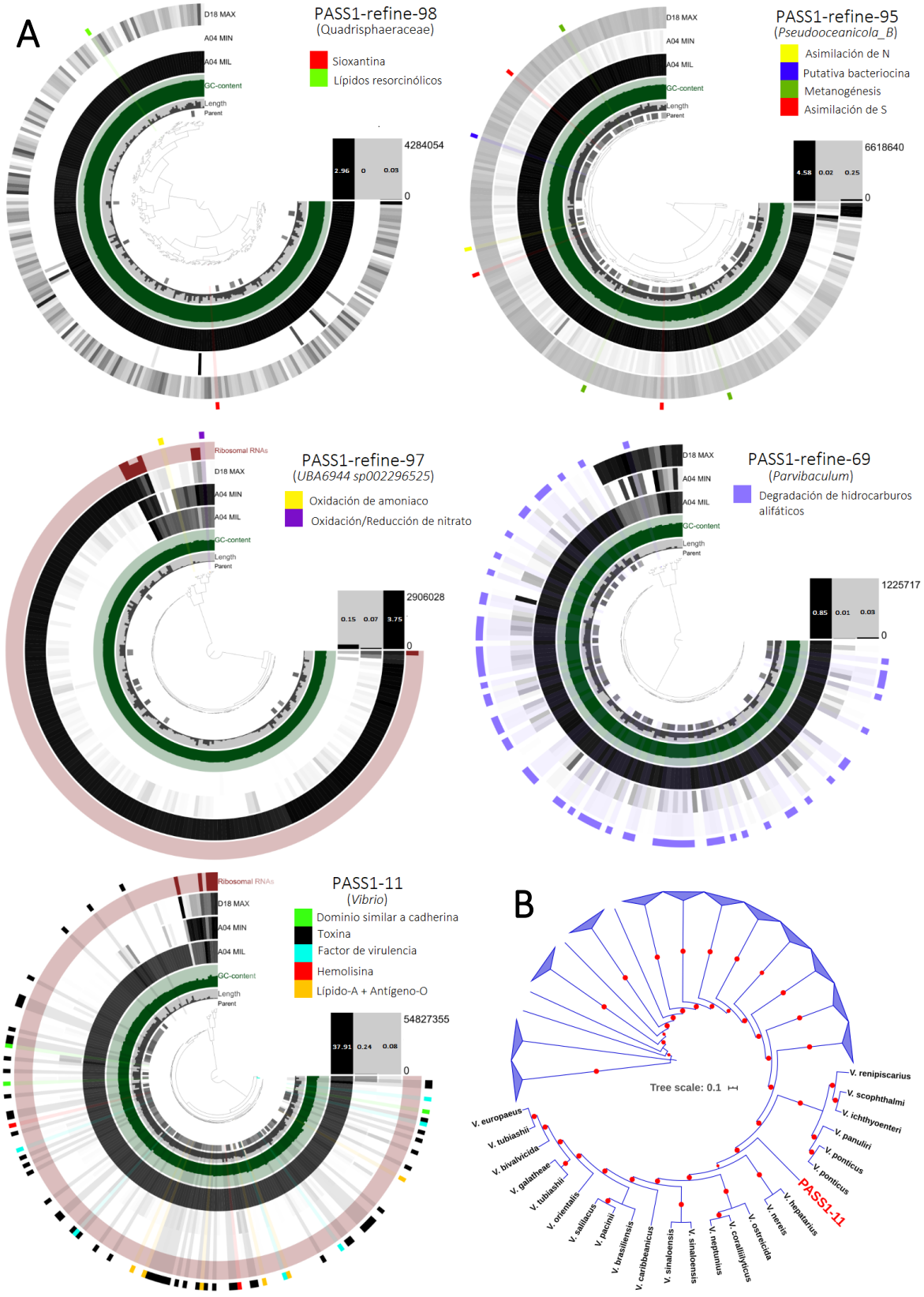


Figura 15. Árbol filogenómico aproximado de máxima verosimilitud con los 10 MAGs seleccionados y sus genomas de referencia. El árbol se basa en un conjunto de 25 genes de copia única específicos de bacterias y arqueas bajo el modelo de sustitución JTT+CAT generado con GtoTree (FastTree). Las ramas en verde y morado representan genomas de bacterias y de arqueas, respectivamente. Se señala con números los valores de soporte SH-aLRT (*Shimodaira–Hasegawa approximate likelihood ratio test)* basado en 1000 remuestreos, número de acceso de NCBI y phyla correspondientes (clase para las Proteobacterias). Se señala el nivel y completeness y contaminación de todos los genomas determinado por CheckM y se indica entre paréntesis la menor categoría taxonómica de los MAGs que pudo ser identificada usando GTDB-Tk (f_: familia; g_genero; s_especie). Árbol disponible en: https://itol.embl.de/tree/1892161564917652158212554 9.

6.3 Inferencias del contexto ecológico de los MAGs

De los cinco MAGs analizados solo el MAG PASS1-refine-97 se reconstruyo prácticamente del metagenoma D18MAX, mientras que restantes provienen principalmente del metagenoma A04MIL (Fig. 16). Se identifico al MAG PASS1-11 como una especie no registrada de *Vibrio* dentro de la GTDB, siendo su referencia más cercana *V. hepatarius* (ANI = 80.14; genoma de referencia: GCF_001274785.1), una especie aislada a partir de camarón patiblanco[94]. *Vibrio* suele ser un género abundante en el océano y tiene cuatro hábitats principales: vida libre o asociado a organismos/superficies de tamaño grande, mediano o pequeño, como peces, rotíferos y partículas de materia orgánica, respectivamente[95,96]; y pese a su plasticidad genómica, las especies suelen agruparse en clados asociados a su estilo de vida[94,96]. Una breve inspección de la vecindad filogenómica del MAG mostró que se agrupo dentro de un clado con especies cuyos aislamientos se obtuvieron a partir macrofauna marina (Fig. 16B), sugiriendo un mismo habito para esta nueva especie.

La presencia de varias vías metabólicas guarda relación con lo anterior al identificarse vías de secreción de proteínas (Tat y Sec-SRP), presencia de flagelo y quimiotaxis, lo cual ayudaría a la identificación de superficies, motilidad y secreción de proteasas y lipasas (Fig. 17). Además de la presencia de sistemas de secreción tipo (SST) l y ll, se identificaron varios genes asociados a la regulación y ensamble de proteínas relacionadas a la patogénesis y mecanismos de depredación bacteriana, que en su conjunto sugieren que se trata de un predador competente y posible patógeno intestinal de fauna marina. Entre dichos elementos se encuentran: lipopolisacáridos, polisacáridos capsulares, proteínas de ensamble de pili tipo IV, toxinas (por ejemplo, toxina-Zeta y putativas toxinas AbiEii), una deacetilasa de peptidoglucano/xilano/quitina, factores de virulencia y el regulador AphA, hemolisinas, proteínas de unión a penicilina, genes de la biosíntesis de lípidos-A (lpxABCDHKLM) y adición del antígeno-O. Curiosamente, también se identificó una luciferasa (Pfam ID: PF00296) así como el gen NorR que confiere resistencia al óxido nitroso (compuesto altamente toxico), ambos, elementos empleados durante la colonización de *V. cholerae* y *Alivibrio fischeri* [97] en sus hospederos



Leyenda en la siguiente pagina →

Figura 16. Características metabólicas clave de los MAGs y su distribución en los metagenomas. (A) Coberturas de los MAGs en los metagenomas visualizadas en Anvi'o. Del centro a la periferia: (1) dendrograma que representa la agrupación jerárquica de los contigs que forman al respectivo MAG en función de la composición de las secuencias (frecuencia de tetranucleótidos y cobertura promedio); (2) *Parent* (Origen) representa el origen de las divisiones de los contigs >20kpb, por defecto Anvi'o realiza este tipo de fragmentación para realizar un análisis de composición de secuencias más detallado, de no indicarse, el espacio en blanco representa un contig <20kpb; (3) *Length* (Longitud) indica el tamaño de cada una de las divisiones; (4) GC con intervalos de 0 - 1; (5) A04MIL, A04MIN y D18MAX representan la cobertura de una región del MAG en los tres metagenomas, una mayor intensidad de color en la división refleja una mayor abundancia de reads mapeados a la misma; (5) *Ribosomal RNA* (RNA ribosomal) señala la ubicación de las secuencias ribosomales en los contigs; (6) En el anillo más externo, se resaltan las regiones genómicas de varias características metabólicas de interés identificadas en los MAGs y señalados en las leyendas con recuadros de colores en sus respectivos paneles; (7) Para cada MAG, se despliega un gráfico de barras debajo de las leyendas indicando el número total de reads mapeados por metagenoma y dentro de cada barra su equivalente porcentual de los respectivos metagenomas. (B) Árbol filogenómico aproximado de máxima verosimilitud de *Vibrio*. Se seleccionaron los genomas de las especies representativas del género registradas en la GTDB para un total de 140 especies. Se realizo el árbol usando GToTree con valores por defecto con base en un conjunto de 172 genes de copia única seleccionados para la clase Gammaproteobacteria bajo el modelo de sustitución JTT+CAT. Se colapsaron los clados restantes donde no se agrupo el MAG PASS1-11 asignado como una nueva especie de *Vibrio* por GTDB-Tk señalado en letras rojas. Se indica con esferas rojas los valores de soporte *SH-aLRT* > 0.9 basados en 1000 remuestreos. De acuerdo con la literatura, entre los tipos de macrofauna asociada a los aislamientos de las especies se encuentran crustáceos, peces, poríferos, bivalvos, rotíferos y cnidarios. Árbol con hospederos identificados disponible en: https://itol.embl.de/tree/189216158494028215783534 65#.

El MAG PASS1-refine-98 fue identificado como un género no registrado dentro de la familia y Quadrisphaeraceae dentro del orden Actinomycetales, en donde se encuentran especies capaces de esporular y sintetizar variedad de productos naturales codificados usualmente en clústers de genes biosintéticos (BGC)[86,98]. Actualmente, la GTDB solo cuenta con dos genomas para dicha familia derivados a partir de metagenomas: *Quadrisphaera granulorum* y *Q. sp900101335*. Dicho género ha logrado ser aislado a partir de reactores biológicos y como endófito de *Oryza sativa*[99,100]; ambas especies se caracterizan por formar tétradas, presentar un alto contenido de GC (~75%) y formar colonias de color naranja. Debido a que se identificó en el MAG la ruta de biosíntesis de retinal casi completa, se analizó a detalle la presencia de BGCs utilizando el servidor de antiSMASH[75] y se identificaron seis putativos BGC, dos de ellos con un 100% de similitud a lípidos resorcinólicos (*Alkylresorcinol*) y sioxantina (*Sioxanthin*), respectivamente. La sioxantina es una carotenoide glicosilado que se caracterizó por primera vez en la Actinobacteria marina *Salinispora*, donde además de determinar la pigmentación naranja de las colonias, cumple funciones biológicas como mecanismo contra el estrés oxidativo, estabilidad de la membrana y al estar glicosilado, sirve como un sitio de adhesión[101]. Por su parte, los lípidos resorcinólicos se han identificado como

elementos esenciales para la formación y resistencia de quistes en la Gammaporobacteria *Azotobacter*[102]. Dichas características, aunadas a la mediana capacidad de formación de biofilms, capacidad de asimilación de azufre, oxidación de sulfuro y nitrato, desnitrificación, presencia de vías anapleróticas y fermentación de ácidos mixtos, sugieren que el MAG PASS1-refine-98 es un microorganismo anaerobio facultativo capaz de asociarse a superficies u organismos y que en condiciones adversas, es capaz de entrar en un estado de dormancia y dispersión mediante esporas hasta volver a encontrar condiciones favorables.

PASS1-refine-95 fue identificado como una especie no registrada de *Pseudooceanicola_B* dentro de la GTDB (ANI = 80.9; genoma de referencia; GCF_003116585.1) dentro de la familia Rhodobacteraceae de Alphaproteobacterias, dichos grupos son reconocidos como uno de los principales encargados de metabolizar el azufre orgánico disuelto a profundidades epipelágica y mesopelágicas[103]. Este tipo de metabolitos con azufre derivan de exudados de organismos, nieve marina (detritos) y del contenido celular de microorganismos vertido en el medio marino por lisis virales y muerte celular, los cuales se precipitan hasta el lecho oceánico[87,104]. La prospección metabólica identifico un metabolismo bastante versátil de este organismo respecto a los demás MAGs, incluyendo vías metabólicas involucradas en la asimilación de azufre (sulfito, sulfuro, sulfato, tiosulfato, dimetilsulfoniopropionato/DMSP), metanogénesis, asimilación de nitrógeno (DNRA, reducción de óxido nitroso y amonificación de nitrito/nitrato), fijación de CO_2 y además presenta movimiento flagelar y quimiotaxis. Otro aspecto interesante del MAG es la presencia de SST I, IV y VI implicados en la adquisición de nutrimentos, virulencia, toxicidad, patogénesis, conjugación bacteriana y como mecanismos de competencia intraespecífica[105]. En conjunto, dichas características y otras más identificadas sugieren que el MAG cuenta con un sistema predatorio bien articulado y diversos mecanismos de adquisición y asimilación de nutrimentos que lo hacen competente en su hábitat, posiblemente habitándolo como forma de vida libre como quimioorganoheterótrofo. Si bien antiSMASH logro identificar la presencia de un BGC que codifica a una bacteriocina, esta no guarda similitud algún otro BGC, por lo que es necesario una investigación más detallada de la putativa bacteriocina identificada en el MAG.

PASS1-refine-69 fue identificado como una especie no registrada de *Parvibaculum* dentro de la GTDB (ANI = 82.68; genoma de referencia: GCA_002480495.1) correspondiente a la clase Alphaproteobacteria. *Parvibaculum* es un género quimioorganoheterótrofo aerobio, capaz de ser cultivado y que suele encontrarse en ambientes acuáticos con forma motil de vida libre en sitios donde ocurre la degradación de hidrocarburos, degradando surfactantes e incluso utilizando hidrocarburos alifáticos (alcanos y alquinos, de 8 a 16 moléculas de C) como única fuente de energía[106,107]. Con base en ello, se buscaron e identificaron genes relacionados a la utilización de alcanos y alquinos encontrándose nueve secuencias predichas como citocromo P450 (alcano-monooxigenasa), 30 alcohol deshidrogenasas, siete aldehído deshidrogenasas, 15 acil-CoA sintetasas, 33 acil-CoA deshidrogenasas, 20 enoil-CoA hidratasas, 11 acil-CoA acetil-transferasas, cinco tioesterasas y 16 putativas acil-CoA tioéster hidrolasas de cadena larga; sugiriendo capacidades de asimilación de compuestos alifáticos similares a *P. lavamentivoran*s[106]. Además, se identificó la capacidad de quimiotaxis, movimiento flagelar, SSTIV parcial, vías anapleróticas parciales, derivación de glioxilato y otras vías implicadas en la β-oxidación de hidrocarburos y ácidos grasos[108,109].

El MAG PASS1-refine-97 fue identificado como la especie *UBA6944 sp002296525*, familia TK06 en el orden de los Microtrichales (Genoma de referencia: GCA_002296525.1; ANI = 97.74%), todas ellas son nuevas categorías taxonómicas definidas a partir de datos metagenómicos por la GTDB y que se encuentran dentro de la clase Acidimicrobiia, la cual ha sido recientemente definida mediante análisis de secuencias ribosomales 16S[110] y además ha sido identificada como uno de los principales grupos encargados de metabolizar el azufre orgánico disuelto a profundidades mesopelágicas[103]. El MAG se distingue del resto debido a que es el único que se encuentra bien representado en el metagenoma D18MAX, presenta un genoma pequeño (2444 genes en 2.36 Mpb) y capacidades metabólicas reducidas respecto a los demás MAGs; sus capacidades fermentativas son reducidas, así como parte del metabolismo central de azucares y ácidos carboxílicos (glucolisis, Entner-Doudoroff, glioxilato). Sin embargo, destaca por su aparente especialización en el metabolismo de N con

genes de varias implicadas en la oxidación de amoniaco, oxidación de nitrato y reducción de nitrato (*dissim nitrate reduction*)[111]. En los ecosistemas acuáticos, este tipo de vías suelen ocurrir en zonas donde la disponibilidad de O_2 y nitrato son dinámicas (Fig. 6) y además se suelen asociarse a bacterias que son simbiontes de eucariontes como moluscos, diatomeas, corales, poríferos y pastos marinos[112–114].

Figura 17. Mapa de calor de la integridad de las rutas metabólicas presentes en los cinco MAGs seleccionados. Se despliega la escala de integridad de las rutas metabólicas, así como las categorías taxonómicas de los MAGs. Para conocer detalladamente la lista de genes analizados de las rutas metabólicas consultar el archivo KOALA_definitions.txt dentro del repositorio en GitHub de KEGG-Decoder (Anexo apartado 9.2).

6.4 Limitantes

Hasta este punto ha quedado claro como la metagenómica puede ser implementada para conocer las comunidades microbianas desde diferentes perspectivas, sin embargo, hay que considerar que aunque resulta una herramienta excepcionalmente útil, también lleva implícita una serie de sesgos logísticos, experimentales y computacionales[91,115]. Por lo tanto, los resultados aquí expuestos deben de tomarse como meras hipótesis sujetas a validación mediante estrategias culturómicas[82,116] y no como demostraciones de la fisiología de los microorganismos, la cual al ser una propiedad emergente, resulta difícil de conceptualizar incluso con aislamientos. Un buen ejemplo de esto y que además resalta la dificultad de trabajar de microrganismos no cultivables es la reciente caracterización de la arquea Asgardiana '*Candidatus' Prometheoarchaeum syntrophicum*[117] *,* la cual ha requerido más de 10 años en lograr ser cultivada en laboratorio así como otros ejemplos con representantes de los phyla del grupo CPR Saccharibacteria y Absconditabacteria que al ser relativamente muy escasos en sus ambientes se habían resistido a ser cultivados[25].

Entre los procedimientos bioinformáticos aquí empleados para reconstruir los MAGs destacan tres puntos clave capaces de determinar los resultados e interpretaciones: (1) la colecta y procesamiento de las muestras; (2) el ensamble metagenómico y (3) el procedimiento de binning. Por ejemplo, al haber usado los filtros Sterivex/Millipore de 0.22µm de diámetro durante el procesamiento de las muestras, se está seleccionando solo aquellos microorganismos con dimensiones adecuadas para el filtro, aunque los procariontes usualmente se encuentran entre los 0.22–1.6 µm[35,46]. Otro aspecto para considerar en este punto es que los metagenomas representan parte de la comunidad microbiana presente en el ecosistema al tiempo específico en el que fue muestreado, y que esta estructura poblacional puede variar en el tiempo a escalas cortas (horas del día) o largas (estaciones del año) en función de la dinámica de la comunicad microbiana con su ambiente.

Además, una de las principales cualidades de Megahit, el coensamble de los reads, no fue aprovechada debido a la estrategia de optimización de memoria implementada,

sustituyendo este paso con el coensamble de los contigs secundarios realizado con Minimus2. El coensamble a nivel de reads ha demostrado ser una estrategia que aprovecha los patrones de cobertura y profundidad de secuenciación a través de varias muestras para reconstruir genomas de microorganismos con baja abundancia en las muestras [21,29,30,46,47,92,118,119]. En este sentido, la cantidad de muestras analizadas resultan relativamente escasas para aprovechar tales enfoques y un mayor número sería necesario[92]; pese a ello, los trabajos realizados por parte del CIGoM resultan pioneros en el área y han sorteado variedad de retos ofreciendo información valiosa con la cual trabajar a futuro conforme se vaya invirtiendo más en investigación y desarrollo[49–51] (Anexo apartado 9.2).

Respecto al binning, Binsanity implementa el algoritmo de propagación por afinidad y utiliza la cobertura de los contigs como parámetro principal para realizar el clustering y posteriormente, refinamientos basados en la composición de las secuencias[63]. Dicha estrategia, al menos en teoría, elude el sesgo que otras herramientas de binning pueden generar al utilizar únicamente parámetros de composición, como incluir especies estrechamente relacionadas en un mismo bin. Sin embargo, comparaciones realizadas entre este tipo de herramientas[12,120] han demostrado que algoritmos como MetaBAT y MaxBin tienen un mejor desempeño y aún más, herramientas como DASTool, Binning_refiner o dRep que integran los genomas generados por diferentes herramientas han sido altamente recomendados en la reconstrucción de MAGs. De hecho, resultados generados usando el pipeline de SqueezeMeta[121] que hacen uso de dicha estrategia aunada a el coensamble de las muestras con Megahit, demostraron reconstruir un número mayor de MAGs de buena calidad llegando a reconstruir mayores porcentajes de los RNA ribosomales (Resultados no mostrados). Con este enfoque fue posible reconstruir MAGs con ≥90% de completeness de géneros como *Synechococcus*, *Pseudomonas*, *Leeuwenhoekiella*, *Nitrosopelagicus* entre otros, los cuales habían sido identificados como potenciales candidatos a reconstruirse con una buena calidad mediante el análisis de los reads.

La identificación de la identidad de los microorganismos es un aspecto rutinario y fundamental en los estudios metagenómicos que llega a cobrar aún más importancia cuando se trata de microorganismos patógenos en casos clínicos. En esta tesis, la herramienta seleccionada para este fin fue GTDB-Tk, un algoritmo automatizado que integra la taxogenómica de procariontes propuesta por la GTDB[47,78] que ha sido adoptada por una considerable parte de la comunidad de científicos dentro del área metagenómica desde su planteamiento, sin embargo, como propuesta taxonómica que refleje la biología de los microrganismos, genera una inflación de hasta 128 phylas de bacterias y arqueas para resolver las relaciones polifiléticas, un rasgo que aunado al desarrollo matemático de la normalización de los rangos taxonómicos hace lucir su naturaleza un tanto más operacional que biológica[39]. Pese a ello, requiere de menos decisiones subjetivas respecto al tipo de información genómica que analizar así como del tiempo de trabajo que ello implica e incluso suele coincidir con la mayoría de los resultados generados por análisis *ad hoc* utilizando genes de copia únicos así como del gen 16S de taxa particulares[122]. Este fue el principal motivo por el cual fue seleccionada, y aunque este tipo de software automatizado resulta bastante útil cuando se cuenta con conocimientos y experiencia en el área limitada al ajustarse a un amplio rango de variedad de datos, podrían no ser adecuadas para resolver ciertos problemas que necesitan de un mayor detalle analítico.

6.5 Perspectivas

Recientemente se pudo identificar que la profundidad, el contenido de materia orgánica, así como la concentración de hidrocarburos aromáticos, son los principales factores que determinan la composición de las comunidades microbianas en el Golfo de México[50]. Sin embargo, resulto un hecho curioso que solo uno de los MAGs evaluados, PASS1-refine-69 clasificado como *Parvivaculum*, se haya relacionado con la capacidad de degradación de hidrocarburos alifáticos. Por dicho motivo sería interesante realizar el escrutinio de los contig secundarios en busca de alguno de los 19 genes claves en la degradación de alcanos e hidrocarburos aromáticos policíclicos identificados por Liang *et al*.[123]. Incluso, los reads podrían ser analizados directamente a nivel de proteína con algoritmos como PLASS[124] para

identificar la diversidad metabólica de los metagenomas. Además, análisis de mapeo de los reads con Anvi'o podrían realizarse usando genomas de microorganismos reconocidos por su especialización en la degradación de hidrocarburos aromáticos como *Alcanivorax*, *Cycloclasticus* o *Pseudomonas*[125,126]. Así mismo, valdría la pena revisar los MAGs en busca de secuencias codificantes o *motifs* que se relacionando con hábitos simbiontes con organismos eucariontes como anquirinas, dominios, WD40, tetratricopeptidos, repeticiones ricas en leucina y pirrolo-quinolina quinona[127,128].

Debido a la abundancia de microorganismos poco representados en los metagenomas (Fig. 12; Tab. sup. 1-2), otra estrategia interesante a implementar es el algoritmo LSA[129], el cual al menos en teoría es capaz de ensamblar genomas bacterianos con abundancias relativas de hasta 0.00001%. Así mismo y en línea con las tendencias actuales sería interesante realizar un estudio masivo de metagenomas provenientes del Golfo de México, un proyecto que de hecho en las últimas etapas de la presente tesis fue publicado por Karthikeyan *et al*.[130] quienes crearon la base de datos *Genome Repository of Oiled Systems* que incluye 2021 MAGs y SAGs que derivan principalmente de metagenomas analizados a partir del derrame petrolero *Deepwater horizon* y otros ambientes con presencia de hidrocarburos con la cual valdría la pena realizar comparaciones así como con otras bases de datos especializadas en hidrocarburos como AnHyDeg[131]. Este proyecto además ilustra la poca investigación dentro de los limites mexicanos del Golfo de México, que hasta la creación del CIGoM no habían sido evaluados a detalle.

La interpretación ecológica de los microrganismos a partir de MAGs se ve menos sesgada en cuanto a mayor sea su calidad, siendo el ideal recuperar genomas circularizados (cerrados)[13,132]. Si bien es posible reconstruir MAGs con tales características, resulta relativamente poco probable al existir en la actualidad solo 59 MAGs de este tipo pese a la cantidad masiva de metagenomas[15]. La calidad de un MAG no solo depende de los valores de completeness y contaminación que presente, en este sentido, los análisis realizados en esta tesis con los MAGs <90% completeness deberán de evaluadas detalladamente e incluso,

vale la pena que se realice el refinamiento manual de los cinco MAGs seleccionados para el análisis metabólico para valorar si las inferencias aquí vertidas se encuentran sesgadas. Esto se puede realizar mediante comparaciones con microorganismos similares para evaluar la presencia de contigs quiméricos, así como la sintenia de los taxa y la taxonomía de los genes que los componen. Esta misma idea puede aplicarse a los MAGs reconstruidos usando SqueezeMeta respecto a los reconstruidos con Binsanity. Finalmente, un punto que tiene que ser considerado de la presente tesis es la unicidad de su carácter meramente descriptivo al no incluirse comparaciones estadísticas entre las muestras. Usualmente los estudios metagenómicos emplean métricas de disimilitud de Bray–Curtis acompañados de análisis de componentes principales y sus variantes así como índices de diversidad como Simpson o Shannon–Weaver y α, β o γ diversidad para evaluar la composición microbiana entre muestras[1,81,133]. Así mismo, los análisis metabólicos pueden llevarse a cabo con herramientas como EnrichM, METACyc, o redes de correlación que permiten conocer vías metabólicas enriquecidas e incluso las posibles redes de interacción que existen entre las comunidades. Sin embargo, hay que tomar dichas inferencias con cautela debido a que la tan sola presencia de los genes puede no ser reflejo de los fenómenos reales entre las comunidades[116].

7. Conclusiones

La metagenómica es una herramienta excepcional para conocer la biodiversidad de procariontes en términos de sus taxa y sus respectivas capacidades metabólicas sin necesidad de cultivarlos laboratorio. Se demostró que es posible reconstruir MAGs a partir de los metagenomas analizados y mediante el análisis filogenómico se lograron identificar microorganismos que en su conjunto representan una familia, seis géneros y 18 especies sin previo registro en la GTDB.

Además, la prospección metabólica de cinco MAGs ayudo a inferir su posible papel ecológico, encontrando metabolismos especializados tal es el caso de *Parvivaculum* en la degradación de hidrocarburos alifáticos o *Vibrio* como un posible patógeno de macrofauna marina. Dichos MAGs a su vez también representan un valioso recurso a las bases de datos genómicas, permitiendo integrarlos en estudios posteriores con análisis más detallados que diluciden más respecto a la biología de estos microrganismos.

La clasificación de los reads realizada permitió describir brevemente la composición de las comunidades microbianas presentes en la columna de agua y sedimentos del Golfo de México, los cuales resultan muy similares a los patrones globales identificados a partir de otros proyectos marinos. Sin embargo, un gran porcentaje de los reads no pudieron ser clasificados usando bases de datos meticulosamente curadas como lo es la nt del NCBI. Particularmente la fracción de sedimentos cuenta con una composición más diversa y representada por microorganismos con baja abundancia respecto a las comunidades de las muestras provenientes de la fracción la columna de agua. Debido a ello y con los métodos implementados en la presente tesis, no fue posible reconstruir MAGs casi completos a parir de las muestras de sedimentos, por lo que un mayor muestreo sería necesario para aprovechar las estrategias actuales de cómputo, así como considerar otros enfoques diseñados para la reconstrucción de MAGs.

8. Referencias

1. Escobar-Zepeda, A., Vera-Ponce De León, A. & Sanchez-Flores, A. The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Front. Genet.* 6, 1–15 (2015).
2. Goodwin, K. D. *et al.* DNA Sequencing as a Tool to Monitor Marine Ecological Status. Front. Mar. Sci. 4, 1–14 (2017).
3. Woese, C. R. & & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci USA. 74, 5088–5090 (1977).
4. Handelsman, J. et al. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol.* 5, (1998).
5. Bergin, C. *et al.* Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* 6, 1–14 (2018).
6. Dharamshi, J. E. *et al.* Marine Sediments Illuminate Chlamydiae Diversity and Evolution. *Curr. Biol.* 30, 1032-1048.e7 (2020).
7. Baker, M. De novo genome assembly: What every biologist should know. *Nat. Methods* 9, 333–337 (2012).
8. Quince, C. et al. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 1211 (2017).
9. Ayling, M., Clark, M. D. & Leggett, R. M. New approaches for metagenome assembly with short reads. *Brief. Bioinform.* 00, 1–11 (2019).
10. van der Walt, A. J. *et al.* Assembling metagenomes, one community at a time. *BMC Genomics* 18, 1–13 (2017).
11. Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4, 1–11 (2016).
12. Hofmann, F. M. P. *et al.* AMBER: Assessment of Metagenome BinnERs. *Gigascience* 7, 1–8 (2018).
13. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731 (2017).
14. Parks, D. H. et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–55 (2015).
15. Lin-Xing C. et al. Accurate and Complete Genomes from Metagenomes. *Genome Re.* (2020).
16. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43 (2004).
17. Rodriguez-R, L. & Konstantinidis, K. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Prepr.* (2016).
18. Rodriguez-R, L. M. *et al.* The Microbial Genomes Atlas (MiGA) webserver: Taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. *Nucleic Acids Res.* 46, W282–W288 (2018).
19. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* 1, 1–6 (2016).
20. Laserson, J., Jojic, V. & Koller, D. Genovo: De novo assembly for metagenomes. *J. Comput. Biol.* 18, 429–443 (2011).
21. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538 (2013).
22. Shendure, J. *et al.* DNA sequencing at 40: Past, present and future. *Nature* 550, (2017).
23. Baudry, L. et al. MetaTOR: A Computational Pipeline to Recover High-Quality Metagenomic Bins From Mammalian Gut Proximity-Ligation (meta3C) Libraries. *Front. Genet.* 10, 1–12 (2019).
24. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* 49, 643–650 (2017).
25. Cross, K. L. *et al.* Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat. Biotechnol.* (2019)
26. Huang, J. M. & Wang, Y. Genomic differences within the phylum Marinimicrobia: From waters to sediments in the Mariana Trench. *Mar. Genomics* 100699 (2019).
27. Wilkins, L. et al. Metagenome-assembled genomes provide new insight into the microbial diversity of two thermal pools in Kamchatka, Russia. *Sci. Rep.* 9, 1–15 (2019).
28. Escobar-Zepeda, A., Sanchez-Flores, A. & Quirasco Baruch, M. Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. *Food Microbiol.* 57, 116–127 (2016).
29. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* 568, 499–504 (2019).
30. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 176, 649-662.e20 (2019).
31. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol.* (2020)
32. Nelkner, J. *et al.* Effect of Long-Term Farming Practices on Agricultural Soil Microbiome Members Represented by Metagenomically Assembled Genomes (MAGs) and Their Predicted Plant-Beneficial Genes. *Genes (Basel).* 10, 424 (2019).
33. Parsons, R. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science (80-.).* 304, 66–74 (2004).
34. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 5, 0398–0431 (2007).
35. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science (80-.).* 348, 1–10 (2015).
36. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463 (2017).
37. Castelle, C. J. & Banfield, J. F. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of

Life. *Cell* 172, 1181–1197 (2018).
38. Zhu, Q. *et al.* Phylogenomics of 10 , 575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun*, (2019).
39. Cavalier-smith, T. & Chao, E. E. Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes , archaebacteria) *Protoplasma*. (2020).
40. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523, 208–211 (2015).
41. Dombrowski,N. et alA. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* (2019).
42. Hug, L. A. Sizing Up the Uncultured Microbial Majority. *mSystems* 3, 1–4 (2018).
43. Lloyd, K. G. et al. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* 3, 1–12 (2018).
44. Salazar, G. & Sunagawa, S. Marine microbial diversity. *Curr. Biol.* 27, R489–R494 (2017).
45. Cavicchioli, R. *et al.* Scientists' warning to humanity: microorganisms and climate change. *Nat. Rev. Microbiol.* (2019)
46. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2 , 631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* 1–8 (2017).
47. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2, 1533–1542 (2017).
48. Rodriguez, J.L & Pardo-Lopez L. Bacterias del Golfo de México: su distribución y potencial aplicación biotecnológica. Biotecnologia en movimiento (2018).
49. Escobedo-Hinojosa, W. & Pardo-López, L. Analysis of bacterial metagenomes from the Southwestern Gulf of Mexico for pathogens detection. *Pathog. Dis.* 75, 1–9 (2017).
50. Godoy-Lozano, E. E. *et al.* Bacterial diversity and the geochemical landscape in the southwestern Gulf of Mexico. *Front. Microbiol.* 9, 1–15 (2018).
51. Raggi, A. L. et al. Metagenomic Profiling and Microbial Metabolic Potential of Perdido Fold Belt (NW) and Campeche Knolls (SE) in the Gulf of Mexico. *Front. Microbiol.* (2020)
52. Pierella Karlusich, P., Ibarbalz, F. M. & Bowler, C. Phytoplankton in the Tara Ocean. *Ann Rev Mar Sci* 233–265 (2020).
53. González-Arias, M.A. Reconstrucción de genomas a partir de metagenomas del Golfo de México. GitHub Repository. Available at: https://github.com/miangoar/Reconstruccion-de-genomas-a-partir-de-metagenomas-del-Golfo-de-Mexico
54. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
55. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257 (2019).
56. Breitwieser, F. P. & Salzberg, S. L. Pavian: Interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics* 36, 1303–1304 (2020).
57. Peng, Y. et al. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428 (2012).
58. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102, 3–11 (2016).
59. Gurevich, A. et al. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075 (2013).
60. Sommer, D. et al. Minimus : a fast , lightweight genome assembler. *BMC Bioinfo.* 11, 1–11 (2007).
61. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–9 (2012).
62. Li, H. *et al.* The Sequence Alignment / Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).
63. Graham, E. D., Heidelberg, J. F. & Tully, B. J. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* 5, e3035 (2017).
64. Kalvari, I. *et al.* Rfam 13.0 : shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* 46, 335–342 (2018).
65. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935 (2013).
66. Chan, P. P. et al. tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes. *bioRxiv* 614032 (2019)
67. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* 1–3 (2019)
68. Chaumeil, P. et al. GTDB-Tk : a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 1–3 (2019).
69. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259 (2019).
70. Team, R. RStudio: Integrated Development for R. vol. RStudio, I (2015).
71. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).
72. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3, e1319 (2015).
73. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* 428, 726–731 (2016).
74. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J.* 12, 1861–1866 (2018).
75. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* 47, W81–

W87 (2019).
76. Breitwieser, F. P. & Salzberg, S. L. Pavian : Interactive analysis of metagenomics data for microbiomics and pathogen identification. *Bioinformatics* (2020)
77. Nawrocki, E. P. Annotating Functional RNAs in Genomes Using Infernal. Methods Mol Biol 1097 (2014).
78. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996–1004 (2018).
79. Lee, M. D. Applications and Considerations of GToTree: A User-Friendly Workflow for Phylogenomics. *Evol. Bioinforma.* 15, (2019).
80. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522 (2012).
81. Liu, J., Meng, Z., Liu, X. & Hua, X. Microbial assembly , interaction , functioning , activity and diversification : a review derived from community compositional data. *Mar. Life Sci. Technol.* (2019).
82. Gutleben, J. et al. The multi-omics promise in context : from sequence to microbial isolate. *Crit. Rev. Microbiol.* 0, 212–229 (2018).
83. Azam, F. & Malfatti, F. Microbial structuring of marine ecosystems. *Nat. Rev. Microbiol.* 5, 782–791 (2007).
84. Karl, D. M. Microbial oceanography: Paradigms, processes and promise. *Nat. Rev. Microbiol.* 5, 759–769 (2007).
85. Bertagnolli, A. D. & Stewart, F. J. Microbial niches in marine oxygen minimum zones. *Nat. Rev. Microbiol.* 16, 723–729 (2018).
86. Subramani, R. & Aalbersberg, W. Marine actinomycetes: An ongoing source of novel bioactive metabolites. *Microbiol. Res.* 167, 571–580 (2012).
87. D'Hondt, S. et al. Subseafloor life and its biogeochemical impacts. *Nat. Commun.* 10, 1–13 (2019).
88. Liby, E. et al. Syntrophy emerges spontaneously in complex metabolic systems. *PLOS Com. Bio*l. 1–17 (2019)
89. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat. Methods* 14, 1063–1071 (2017).
90. Vollmers, J., Wiegand, S. & Kaster, A. K. *Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! PLoS ONE* vol. 12 (2017).
91. Bharti, R. & Grimm, D. G. Current challenges and best-practice protocols for microbiome analysis. *Brief Bioinform* (2019).
92. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* 1, (2014).
93. Sedlar, K., Kupkova, K. & Provaznik, I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.* 15, 48–55 (2017).
94. Thompson, F. L. *et al.* Vibrio fortis sp. nov. and Vibrio hepatarius sp. nov., isolated from aquatic animals and the marine environment. *Int. J. Syst. Evol. Microbiol.* 53, 1495–1501 (2003).
95. Asplund, M. E. *Ecological aspects of marine Vibrio bacteria Exploring relationships to other organisms and a changing environment*. PhD Thesis. University of Gothenburg, Switzerland. (2013).
96. Le Roux, F. & Blokesch, M. Eco-evolutionary Dynamics Linked to Horizontal Gene Transfer in Vibrios. *Annu. Rev. Microbiol.* 72, 89–110 (2018).
97. Negus, D. *et al.* Predator Versus Pathogen : How Does Predatory Bdellovibrio bacteriovorus Interface with the Challenges of Killing Gram-Negative Pathogens in a Host Setting ? *Annu Rev Microbiol* (2017)
98. Jensen, P. R. Natural Products and the Gene Cluster Revolution. *Trends Microbiol.* 24, 968–977 (2016).
99. Maszenan, A. et al.. Quadrisphaera granulorum gen. nov., sp. nov., a Gram-positive polyphosphate-accumulating coccus in tetrads or aggregates isolated from aerobic granules. *Int. J. Syst. Evol. Microbiol.* 55, 1771–1777 (2005).
100. Muangham, S. et al. Quadrisphaera oryzae sp. nov., an endophytic actinomycete isolated from leaves of rice plant (Oryza sativa L.). *J. Antibiot. (Tokyo).* 72, 93–98 (2019).
101. Richter, T. K. S., Hughes, C. C. & Moore, B. S. Sioxanthin, a novel glycosylated carotenoid, reveals an unusual subclustered biosynthetic pathway. *Environ. Microbiol.* 17, 2158–2171 (2015).
102. Horinouchi, S. Combinatorial biosynthesis of non-bacterial and unnatural flavonoids, stilbenoids and curcuminoids by microorganisms. *J. Antibiot. (Tokyo).* 61, 709–728 (2008).
103. Landa, M. *et al.* Sulfur metabolites that facilitate oceanic phytoplankton–bacteria carbon flux. *ISME J.* 2536–2550 (2019)
104. Thume, K. *et al.* The metabolite dimethylsulfoxonium propionate extends the marine organosulfur cycle. *Nature* 563, 412–415 (2018).
105. Costa, T. R. D. *et al.* Secretion systems in Gram-negative insights. *Nat. Publ. Gr.* 13, 343–359 (2015).
106. Schleheck, D. *et al.* Complete genome sequence of Parvibaculum lavamentivorans type strain (DS-1 1). *Stand. Genomic Sci.* 5, 298–310 (2011).
107. Rosario-Passapera, R. *et al.* Parvibaculum hydrocarboniclasticum sp. nov., a mesophilic, alkane-oxidizing alphaproteobacterium isolated from a deep-sea hydrothermal vent on the East Pacific Rise. *Int. J. Syst. Evol. Microbiol.* 62, 2921–2926 (2012).
108. Nabeshima, S., Tanaka, A. & Fukui, S. Effect of Carbon Sources on the Level of Glyoxylate Cycle Enzymes in n-Alkane-utilizable Yeasts. *Agric. Biol. Chem.* 41, 275–279 (1977).
109. Pen, M. De. Choking on Acetyl-CoA, the Glyoxylate Shunt, and Acetyl-CoA-Driven Metabolism. *Handb. Hydrocarb. Lipid Microbiol.* (2010)
110. Hu, D., Cha, G. & Gao, B. A phylogenomic and molecular markers based analysis of the class Acidimicrobiia. *Front. Microbiol.* 9, 1–12 (2018).

111. Kuypers, M. M. M., Marchant, H. K. & Kartal, B. The microbial nitrogen-cycling network. *Nat. Rev. Microbiol.* 16, 263–276 (2018).
112. Fiore, C. L., Jarett, J. K., Olson, N. D. & Lesser, M. P. Nitrogen fixation and nitrogen transformations in marine symbioses. *Trends Microbiol.* 18, 455–463 (2010).
113. Kamp, A., Høgslund, S., Risgaard-Petersen, N. & Stief, P. Nitrate storage and dissimilatory nitrate reduction by eukaryotic microbes. *Front. Microbiol.* 6, 1–15 (2015).
114. Sogin, E. *et al.* Seagrass excretes sugars to their rhizosphere making them the sweet spots in the sea.*bioRxiv* (2019).
115. Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research. *BMC Biol.* 17–20 (2019).
116. Carr, A., Diener, C., Baliga, N. S. & Gibbons, S. M. Use and abuse of correlation analyses in microbial ecology. *ISME J.* (2019)
117. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote – eukaryote interface. *Nature* (2020)
118. Nielsen, H. B. *et al.* Articles Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 32, (2014).
119. Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species , strains , and phage during infant gut colonization. *Genome Res*. 111–120 (2013).
120. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat. Methods* 14, 1063–1071 (2017).
121. Tamames, J. & Puente-Sánchez, F. SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Front. Microbiol.* 10, 1–10 (2019).
122. Coil, D. A. *et al.* Genomes from bacteria associated with the canine oral cavity: A test case for automated genome-based taxonomic assignment. *PLoS One* 14, e0214354 (2019).
123. Hug, L. A., Flynn, T. M. & Sun, B. Long-Term Oil Contamination Alters the Molecular Ecological Networks of Soil Microbial Functional Genes. *Front Microbiol* 7, 1–13 (2016).
124. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* 16, 603–606 (2019).
125. Xu, X. *et al.* Petroleum Hydrocarbon-Degrading Bacteria for the Remediation of Oil Pollution Under Aerobic Conditions: A Perspective Analysis. *Front. Microbiol.* 9, 1–11 (2018).
126. Head, I. M., Jones, D. M. & Röling, W. F. M. Marine microorganisms make a meal of oil. *Nat. Rev. Microbiol.* 4, 173–182 (2006).
127. Robbins, S. J. *et al.* A genomic view of the reef-building coral Porites lutea and its microbial symbionts. *Nat. Microbiol.* (2019)
128. Karimi, E. *et al.* Genomic blueprints of sponge-prokaryote symbiosis are shared by low abundant and cultivatable Alphaproteobacteria. *Sci. Rep.* 9, 1–15 (2019).
129. Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* 33, 1053–1060 (2015).
130. Karthikeyan, S. *et al.* Genome repository of oil systems: An interactive and searchable database that expands the catalogued diversity of crude oil-associated microbes. *Environ. Microbiol.* 22, 2094–2106 (2020).
131. Callaghan, A. V. & Wawrik, B. AnHyDeg: a curated database of anaerobic hydrocarbon degradation genes. GitHub repository (2016) Available at: doi:https://doi.org/10.5281/zenodo.61278.
132. Shaiber, A. & Eren, A. M. Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories. *MBio* 10, 1–3 (2019).
133. Buttigieg, P. L. & Ramette, A. A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. FEMS Microbiol Ecol 90, 543–550 (2014).

9. Anexo

9.1. Líneas de comando (ver ref. 53)

Evaluación de calidad

```
# Basic_stats.pl (cortesía de Alejandra Escobar: ebi.ac.uk/about/people/alejandra-escobar-zepeda)
ls /reads/*R1*.fastq > R1_reads && ls /reads/*R2*.fastq > R2_reads

paste r1_reads.txt r2_reads.txt > reads_list.txt

stats.pl reads_list.txt # Genera el archivo "basic_stats_out.txt" con las estadísticas

# FastQC
fastqc /reads/*.fastq -o /reads/FastQC_results
```

Clasificación taxonómica con Kraken2

```
# Kraken2 + nt_db. Ejemplo solo para el metagenoma A04MIL
kraken2 --db /kraken2/nt_DB/ --threads 20 --report A04MIL.report --use-names \
--paired /reads/A04_MIL_1_R1.fastq /reads/A04_MIL_1_R2.fastq 2> A04MIL.out

# Pavian
# Cargar los archivos con extensión '.report' a Pavian https://fbreitwieser.shinyapps.io/pavian/
```

Ensamble metagenómico y mapeo de secuencias

```
# Megahit. Ejemplo solo para el metagenoma A04MIL
megahit -t 50 -1 /reads/A04_MIL_1_R1.fastq -2 /reads/A04_MIL_1_R2.fastq -o A04MIL_megahit_meta

megahit --presets meta-large -t 50 -1 /reads/A04_MIL_1_R1.fastq -2 /reads/A04_MIL_1_R2.fastq \
 -o A04MIL_megahit_large

megahit --presets meta-sensitive -t 50 -1 /reads/A04_MIL_1_R1.fastq -2 /reads/A04_MIL_1_R2.fastq \
-o A04MIL_megahit_sensitive

# IDBA-UD
fq2fa --merge /reads/A04_MIL_1_R1.fastq /reads/A04_MIL_1_R2.fastq A04MIL_reads_idba_input.fa

idba_ud -r A04MIL_reads_idba_input.fa -o A04MI_idba.fa

# Concatenación de acuerdo con la fracción y ensamblador. Ejemplo solo con los contigs de la
# fracción de agua generados con IDBA-UD.
cat A04MIL_idba.fa A04MIN_idba.fa D18MAX_idba.fa > water_primary_contigs_idba.fa

# Evaluación de los ensambles con Quast
quast.py -o water_primary_contigs_QUAST_STATS -m 0 -t 20 water_primary_contigs_idba.fa \
water_primary_contigs_megahit_meta.fa water_primary_contigs_megahit_large.fa \
water_primary_contigs_megahit_sensitive.fa

# Coensamble con Minimus2
toAmos -s water_primary_contigs_megahit_sensitive.fa \
-o water_primary_contigs_megahit_sensitive_input_2_minimus2.afg

minimus2 water_primary_contigs_megahit_sensitive_input_2_minimus2.afg \
-D OVERLAP=100 MINID=95

# Concatenación de los contigs coensamblados y singletones
cat *input_2_minimus2.fasta *input_2_minimus2.singletons.seq > water_secondary_contigs.fa
```

Binning metagenómico

```
# Bowtie2 y SamTools
bowtie2-build *secondary_contigs.fa water_secondary_contigs_index

bowtie2 -x /mapping/water_secondary_contigs_index --very-sensitive --end-to-end --no-unal \
-q --threads 20 -1 \
/reads/A04_MIL_1_R1.fastq,/reads/A04_MIN_2_R1.fastq,/reads/D18_MAX_1_R1.fastq -2 \
```

```
/reads/A04_MIL_1_R2.fastq,/reads/A04_MIN_2_R2.fastq,/reads/D18_MAX_1_R2.fastq -S water_RAW.sam

samtools view -Sb water_RAW.sam -o water_RAW.bam

samtools sort -o water_SORTED.bam water_RAW.bam

# Descarta los contigs <2kpb
cat water_secondary_contigs.fa | megahit_toolkit filterbylen 2000 > water_secondary_contigs_2k.fa

# Binsanity. Consultar código del script get_bins.sh en el repositorio de la Ref. 50 escrito de
# acuerdo al tutorial de binning de Tully et al. (ver Anexo apartado 9.2)
qsub -V get_bins.sh

# sRNAs y tRNA (cortesía de Alexandre Almeida: ebi.ac.uk/about/people/alexandre-almeida)
for file in PASS*fna; do
   RNA_detect_archaea.sh  $file
done

for file in PASS*fna; do
   RNA_detect_bacteria.sh  $file
done
```

Clasificación taxogenómica de los MAGs

```
# GTDB-Tk
gtdbtk classify_wf --genome_dir selected_MAGs --out_dir GTDB_MAGs_Taxonomy --cpus 25

# GToTree
GToTree -f my_28_MAG.txt -H Bacteria_and_Archaea.hmm -N -j 20 -n 4 -o ONLY_MAGs

GToTree -f my_10_selected_MAGs.txt -a MAGs_ref_accessions.txt -H Bacteria_and_Archaea.hmm \
-j 20 -n 4 -t -L Domain\,Phylum\,Class\,Species -o FINAL_10MAGs_TREE

# iTol
# Cargar el archivo con extencion .tre dentro del directorio FINAL_10MAGs_TREE
```

Análisis metabólico y de cobertura

```
# Anvi'o
anvi-run-workflow -w metagenomics -c config.json

for file in 03_CONTIGS/*.db; do
   anvi-run-pfams -c $file --pfam-data-dir /Database/Anvio/pfam -T 20
done

anvi-get-sequences-for-gene-calls -c 03_CONTIGS/PASS1_11-contigs.db --get-aa-sequences \
-o PASS1_11_genes.fa

# GhostKOALA
# Cargar los genes usando la base de datos genus_prokaryotes y descargar las anotaciones
KEGG-to-anvio --KeggDB /GhostKoalaParser/samples/KO_Orthology_ko00001.txt -i KEGG_PASS1-11.txt \
-o KEGG_PASS1-11_AnviImportable.txt

anvi-import-functions -c PASS1_11-contigs.db -i KEGG_PASS1-11_AnviImportable.txt

# KEGG-Decoder
KEGG-decoder -i KEGG_PASS1-11.txt -o FUNCTION_OUT_PASS1-11.list -v static
mv function_heatmap.svg decoder_PASS1-11.svg

hmmsearch --tblout PASS1_11_expanderv0.3.tbl -T 75 /KEGGDecoder/HMM_Models/expander_dbv0.6.hmm \
PASS1_11-contigs.db_genes.fa

Decode_and_Expand.py FUNCTION_OUT_PASS1-11.list HMM_OUT_PASS1_11.list
mv decode-expand_heatmap.svg heatmap_PASS1_11.svg

# Inspección genómica
anvi-interactive -p 06_MERGED/PASS1_refine_97/PROFILE.db -c 03_CONTIGS/PASS1_refine_97-contigs.db
```

9.2. Sitios de interés

Webinar "*Reconstrucción y análisis de genomas a partir de metagenomas*" impartido por Miguel Ángel González Arias para la consultora bioinformática *Winter Genomics*.
https://www.youtube.com/watch?v=ckIbT93Qhjc

Artículos científicos publicados por parte de miembros del CIGoM
https://cigom.org/noticias/category/articulos-cientificos/

Repositorio a GitHub de la presente tesis donde se detallan los comandos del apartado 9.1 del anexo y se disponen individualmente cada una de las figuras para su consulta (ref. 53).
https://github.com/miangoar/Reconstruccion-de-genomas-a-partir-de-metagenomas-del-Golfo-de-Mexico

Protocolos de ensamble y binning de Tully *et al*.[46]
Ensamble:
https://www.protocols.io/view/assembly-procedure-applied-to-tara-oceans-data-ex-hfqb3mw
Binning:
https://www.protocols.io/view/binning-procedure-applied-to-tara-oceans-dataset-u-iwgcfbw

Protocolos detallados de análisis de Anvi'o[72]
http://merenlab.org/2018/01/17/importing-ghostkoala-annotations/#export-anvio-gene-calls
http://merenlab.org/2018/07/09/anvio-snakemake-workflows/#metagenomics-workflow
http://merenlab.org/2016/06/22/anvio-tutorial-v2/
http://merenlab.org/vocabulary/

KOALA_definitions.txt (KEEG-Decoder)[74]
https://github.com/bjtully/BioData/blob/master/KEGGDecoder/KOALA_definitions.txt