VOTOS DE APROBATORIOS

Secretaria Ejecutiva del Instituto de Investigación en Ciencias Básicas Aplicadas de la Universidad Autónoma del Estado de Morelos.
P r e s e n t e .

Por medio de la presente le informamos que después de revisar la versión escrita de la tesis que realizó la C. **HERNANDEZ JAIMES LIZBETH** con número de matrícula **10002751** cuyo título es:

"Reconstrucción de la microbiodiversidad asociada a sedimentos contaminados por la actividad antropogénica en la cuenca del río Apatlaco"

Consideramos que **SI** reúne los méritos que son necesarios para continuar los trámites para obtener el título de **LICENCIADO EN CIENCIAS ÁREA TERMINAL DE BIOQUÍMICA Y BIOLOGÍA MOLECULAR**

Cuernavaca, Mor a 26 de abril del 2022

Atentamente
Por una universidad culta

Se adiciona página con la e-firma UAEM de los siguientes:

DR. ARMANDO HERNÁNDEZ MENDOZA	**(PRESIDENTE).**
DRA. SONIA DÁVILA RAMOS	**(SECRETARIO).**
DR. AYIXON SÁNCHEZ REYES	**(VOCAL).**
DRA. LUZ DE MARÍA BRETÓN DEVAL	**(SUPLENTE).**
DR. JOSÉ HERNÁNDEZ ELIGIO	**(SUPLENTE).**

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

ARMANDO HERNANDEZ MENDOZA | Fecha:2022-08-18 09:09:46 | Firmante
hpCD3HMwKRKUjcTEwodqzy0jRSDqlrcmjJmPGiS8q7YI3VnuDL7vISnuV6YUwOFSGoEj1PGQiy0cqsIpQmsL6aAO4Z0jcG0REy54fluwvnBDf8ENZQGlmVY4FV/7Wpho1sbtCU
Nc29W8phyGJ+0aJ5lxKEslnmfo53fmvfTlWlHcjXVCF/6CaPtMa0uS0QdPRwbEfRfvV7IUOrlwA4MaUufg+9EJELKk8YIwqQYzs81Hy5b3imfVqYyQJpgQTDpIGSeYWUeTm9ChoT
NvigEPFddo4kzbVCmbremiPa9T7pCDi5jzr5Cxk+SjfRDuAGextWy3IfGtV/nlnxCRm6H0Qw==

SONIA DAVILA RAMOS | Fecha:2022-08-18 10:14:25 | Firmante
MZdhFDredoQnHazEoz4yVd05JvtB33xZEfZeBS5euvWG86NFGdY/8+IBS7blisFrLGFPf/8ciqIDTEVIL2e3S4x/jBHmdQwncn6SF3JXZMfuM5p/oz/Bn/tt7u8YiW0xX3CEPs7rFYKZ
PlliAq843yz3k6bNI+esRY8SRXAvhjjUILVP5mdo6P0J9tNmn98nLMUcd8BMxeKwLZ90icYbJjPKy7kAcBeUGJMRZKDzyoYpV3+5ypMlh5ODGYYdHLE8entFW7lU+grWJ4N/Zkuj/
b9Ezqz3dva3YyjULyhMR7Y+6WYDJKiOHi0NZZ7q1DLFQVDhbviofv3F/VSgKYGqyA==

JOSÉ ALBERTO HERNÁNDEZ ELIGIO | Fecha:2022-08-18 10:38:48 | Firmante
NxPozlyfZavn3odf1RqrnVEii3pMpaEytV6Pv3A2nnMP7sMPYrZTXIykmQn1peZ0Zr0timmdEdC3yTdMSlImI1CLAjd9BFh/DEJ77OS89vKamjCTlN31bvodkwkrCMvH1SYWKdipW9
sDgs3Mi0iRohJMLbGKBDW76rrR3xXIQNagD+AAKNo2OO36AwvclXOs+DOJvmzrZxS2B7ubofmyMWCO01hRPfJsJlptgs0MBMfPOKgbZyYgJ0ZSF2SEu8fxxl6NLfY6aj/3XVTcb
SXKIMlzPxl1xNbBKWZUb7GxfCvCyIGM6aiM8ZzO7NkgKAiubSLUrhFC4Jn8VvaA2zYdJg==

AYIXON SÁNCHEZ REYES | Fecha:2022-08-18 10:55:44 | Firmante
YRhVsvh65Zc1qC4TGbF0S9iMcT5SmVazRmhDqbBjoxxgmqavclG/xnzFuuA8E8ieExGn423rc97xyKx+EcMexvm5tLJV+hrY7aSewgQ1m8tfRz9lRaG9IqFvuMCVinAXniQRmyQE
p/U3jw6286c4uxMCdW8GabuZdr/8/Jd2qGjcN62T4XrCcbeKHX68PyN2yCiSyrogwf3E0HnDcxtjeJGptOhdS4monca2AbchY20Tp0z4mFpGWebOqnhXN1BsM7iXe3O7ummsPSR
A8UhsAMrf003jni48voum+FP50Opza34opPknOb3RZ+KHBVKIUFedmnzZ6xZ3fedbIMKI8w==

LUZ DE MARIA BRETON DEVAL | Fecha:2022-08-25 20:36:02 | Firmante
i31PRWOt6MyHVP48Zw1gbwzj9xxMGHo8GFRzenR+cRYMHnI/awYDVILDX7ZxCPbtY+3GGlcvVWFvWZdK3OpRQqdfJoWjw4RhvV76ldZ8+FFGpBrltVp2IoC3bQSilsV2KhkgwdpS6hHQMOI3xhOwNBZiP2mkeP7RPIVvJhjkcmkUZfu8oOxdqPLX1bu4IGHQVq93IrCuU5WRKYgiSgHeuA4y4aUZf1cnWJFCo/o+sCDnp/DH1CSeK4b7tS6Eago8sGE0cXKzMxg
b0RK01u7kJefXooD9EMq819nqt6jHojhPQ8xspTOVE+Y3clLOIDDXgD/2nJ9+STccBN3KaLIGLg==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

3jA0Rdbws

https://efirma.uaem.mx/noRepudio/VpAZJcwFGNvuLdIgjsCwlgcQt9CjtF44

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Instituto de Investigación en Ciencias Básicas y Aplicadas

Centro de Investigación en Dinámica Celular

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS

Área Terminal en Bioquímica y Biología Molecular

"RECONSTRUCCIÓN DE LA MICROBIODIVERSIDAD ASOCIADA A SEDIMENTOS CONTAMINADOS POR LA ACTIVIDAD ANTROPOGÉNICA EN LA CUENCA DEL RÍO APATLACO"

T E S I S

Que presenta:

LIZBETH HERNANDEZ JAIMES

Para obtener el Grado de

LICENCIADO EN CIENCIAS

Director de tesis:

Dr. Ayixon Sánchez Reyes

Sinodales:

Dr. Armando Hernández Mendoza

Dra. Sonia Dávila Ramos

Dra. Luz de María Bretón Deva

Dr. José Hernández Eligio

CUERNAVACA, MORELOS 2022

AGRADECIMIENTOS

A Yari, Mar y Carito por brindarme su amistad y apoyo incondicional. Al Dr. Ayixon y al Dr. Maikel por confiar en mí para realizar este proyecto y ser guías para lograr el objetivo. A todos mis profesores a lo largo de mi carrera universitaria, ¡sin ustedes no hubiese llegado hasta aquí!

A mi familia, especialmente a mi mamá y mi papá por siempre estar conmigo en todo lo que me propongo. En general, a todas las personas que me apoyaron e hicieron que confiara en mi para lograr concluir este proyecto.

¡Muchas gracias por todo lo que me enseñaron y por su apoyo!

RESUMEN

En este proyecto, utilizamos herramientas bioinformáticas para comparar dos metodologías de asignación taxonómica (ARNs ribosomales y marcadores de copia única) de una muestra sometida a selección nutricional proveniente de sedimentos del río Apatlaco. Ambas aproximaciones constituyen actualmente la principal dicotomía metodológica para reconstruir perfiles taxonómicos en metagenomas ambientales; con importantes consecuencias en la inferencia y descripción de la microbiodiversidad asociada a biomas naturales. Con la ayuda de dos populares herramientas (FOCUS y Kaiju) se pudo realizar la asignación taxonómica y con esta, una comparación para determinar que método replicaría mejor los patrones de abundancias relativas obtenidas desde un modelo nulo de asignación con base en lecturas crudas de secuenciación masiva. Además, se reconstruyeron genomas individuales desde el metagenoma representativo del río, con el objetivo de identificar las firmas especie-específicas de microorganismos presentes en este bioma. Utilizando índices globales de relación genómica, se infirió el género y/o la especie de cada genoma individual, con el fin de contribuir a la descripción del paisaje taxonómico y funcional de los microorganismos del río Apatlaco. Por último, se realizó una exploración funcional sobre el set de genomas individuales resueltos, utilizando herramientas canónicas de anotación funcional como GhostKOALA y KeegMapper; ello permitió inferir probables determinantes genéticos relacionados con diferentes funciones biológicas del metagenoma *in situ*. Con este proyecto, se pudo concluir que los marcadores de copia única describen mejor el perfil taxonómico de la muestra en estudio tomando como referencia el modelo nulo. Además, se identificaron enzimas con potencial para degradar xenobióticos, entre ellas destacan las participantes en la vía de degradación de benzoato, degradación de cloroalcanos y cloroalquenos, degradación de nitrotolueno, degradación de naftaleno, metabolismo de xenobióticos y fármacos mediado por enzimas tipo citocromos p450.

ÍNDICE

Tabla de contenido

LISTA DE SIGLAS, SIMBOLOS Y ABREVIATURAS.

1. ADN: ácido desoxirribonucleico.
2. ANI: Identidad Promedio a nivel de Nucleótidos.
3. ARNr: ácido ribonucleico ribosomal.
4. ARN: ácido ribonucleico.
5. ASM: Sociedad Americana de Microbiología.
6. CF: Coliformes Fecales.
7. COG: Clústeres de grupos de proteínas ortólogos.
8. CONAGUA: Comisión Nacional del Agua.
9. DQO: Demanda Química de Oxígeno.
10. FA: Fracción de Alineamiento.
11. Ground truth: Modelo nulo.
12. Hi-C: biblioteca de ligadura de proximidad.
13. H_0: Hipótesis nula.
14. IMTA: Instituto Mexicano de Tecnología del Agua.
15. INEGI: Instituto Nacional de Estadística y Geografía.
16. KAAS: KEGG Automatic Annotation Server.
17. KEGG: Kyoto Encyclopedia of Genes and Genomes.
18. KO: números K u ortólogos de KEGG.
19. Koala: anotación de enlaces y ortología de KEGG.
20. l/s: litros sobre segundo.
21. NCBI: National Center for Biotechnology Information.
22. pb: pares de bases.
23. SEMARNAT: Secretaría de Medio Ambiente y Recursos Naturales.
24. UBCG: Genes del metabolismo central bacterianos (marcadores de copia única).

LISTA DE TABLAS

LISTA FIGURAS

Capítulo 1

1.1 INTRODUCCIÓN

Ubicada en el estado de Morelos, la microcuenca del río Apatlaco pertenece al río Amacuzac, nace de los escurrimientos de agua de las lagunas de Zempoala y desemboca en el río Yautepec. El río Apatlaco abarca 746 km² de extensión territorial y se estima que 907,473 personas residen en los alrededores de la cuenca. Su agua ha sido utilizada para riego y actividades recreativas, sin embargo, actualmente es evidente la contaminación crónica que sufre esta cuenca, ya que recibe 723.3 l/s de descargas de aguas residuales de origen industrial y municipal, lo que contribuye al deterioro de la calidad del agua (CONAGUA, 2012 & CEAGUA, 2017). Una consecuencia de la contaminación en la cuenca es el desarrollo de patógenos, la emisión de malos olores y el peligro por el probable contacto con sustancias tóxicas, poniendo en riesgo la salud de toda la población que vive en los alrededores. Adicionalmente, se asocia un impacto negativo en la economía del estado, al limitar el uso del agua para fines industriales, de agricultura y para consumo humano.

La contaminación de la cuenca del río Apatlaco ha recibido mayor atención en años recientes (Breton-Deval et al., 2019). Sin embargo, aún se carece de estudios que evalúen la composición microbiana de la cuenca y sus fluctuaciones como respuesta a la actividad antropogénica. Por ello, el presente proyecto tiene como objetivo reconstruir la microbiodiversidad asociada al río Apatlaco, mediante el análisis de un metagenoma proveniente de sedimentos sujeto a selección nutricional. Esto es importante porque, nos permitirá explorar y describir ventanas locales asociadas con eventos de contaminación, identificar probables riesgos infectocontagiosos e inferir sus orígenes, así como proponer alternativas de tratamiento enfocadas en degradación de contaminantes ambientales. Específicamente, proponemos utilizar estimadores de microbiodiversidad filogenética, entre los que se encuentran genes marcadores de copia única y los tradicionales genes ribosomales (ARNs ribosomales); bajo la hipótesis de que los marcadores de copia única podrían ser mejores estimadores para explorar los

perfiles taxonómicos de comunidades ambientales (una hipótesis largamente discutida por la comunidad científica*)* (Segata et al., 2012). Sabemos que los genomas completos o cuasi-completos son la huella filogenética más concluyente para asignar categorías taxonómicas, y permiten la inferencia del potencial funcional directamente relacionado con los taxones del metagenoma estudiado. También proponemos la reconstrucción de genomas a partir del metagenoma, para comprobar si estos pueden o no considerarse estimadores composicionales de diversidad en muestras ambientales. Actualmente, la microbiodiversidad mexicana ha sido un tema de poca atención, por lo que pretendemos crear un catálogo de nuevos microorganismos autóctonos del río Apatlaco, con sus respectivas funciones asociadas. Esto permitiría en el futuro, profundizar en las diversas estrategias evolutivas de los microorganismos para adaptarse a la vida en cuerpos de agua contaminados.

1.2 Planteamiento del problema

Se sabe que el agua del río Apatlaco está fuertemente contaminada a causa del derrame de aguas residuales de los diferentes sectores (fábricas, rastros, basura de la población, entre otros) (CONAGUA, 2012). Este problema propicia el crecimiento de microorganismos patógenos, poniendo en riesgo la salud de las personas que residen en los alrededores de la cuenca. Además, existe la posibilidad de que en el río Apatlaco haya microorganismos con capacidades para la biorremediación, ya que la contaminación puede funcionar como un medio selectivo que permite crecer a los microorganismos aptos para resistir y lidiar con las condiciones de extrema contaminación de algunos sitios. Por lo que, es importante describir taxonómica y funcionalmente la microbiodiversidad que contiene este cuerpo de agua.

Para fines educativos y de importancia ambiental, se deber identificar el método más eficiente y evaluar la precisión de los diferentes métodos empleados para realizar la asignación taxonómica y determinar cuál describe mejor la diversidad microbiológica de nuestro metagenoma; nos propusimos realizar un perfilado taxonómico y funcional de una muestra sometida a selección nutricional,

representativa de sedimentos provenientes del río Apatlaco, utilizando marcadores de copia única y ARNs ribosomales. También, con el objetivo de explorar si el perfil taxonómico de los genomas individuales es similar al del metagenoma completo se ensamblaron genomas individuales de los microorganismos que hay en una fracción representativa de los sedimentos del Apatlaco. Nuestro proyecto es importante para identificar microorganismos con relevancia ambiental, específicamente, aquellos con capacidad de degradar xenobióticos, así como, para conocer la microbiodiversidad existente en el río Apatlaco.

1.3 Hipótesis

Los sedimentos impactados por la actividad antropogénica en la cuenca del río Apatlaco reflejarán paisajes de microbiodiversidad diferenciales, en dependencia de los estimadores filogenéticos empleados (ARNs ribosomales *vs* marcadores de copia única).

1.4 Predicciones

De acuerdo con la hipótesis planteada, una posibilidad es que el paisaje de microbiodiversidad capturado mediante marcadores de copia única sea similar al capturado mediante análisis de los ARNs ribosomales en términos composicionales relativos.

Alternativamente, los paisajes de microbiodiversidad obtenidos mediante análisis de secuencias de los ARNs ribosomales y marcadores de copia única, serán composicionalmente diferenciales entre sí.

Un punto importante que considerar es que todas nuestras evaluaciones fueron realizadas sobre una fracción representativa del metagenoma de sedimentos del río Apatlaco (un ensamble metagenómico). Aunque se cuenta con las lecturas crudas del mismo, el perfil taxonómico de estas lecturas constituirá un modelo nulo, para ponderar qué tanto se puede capturar taxonómicamente desde un ensamble metagenómico.

1.5 Objetivo General

- Describir los paisajes funcionales y de microbiodiversidad de sedimentos del río Apatlaco impactado por actividades antropogénicas.

1.5.1 *Objetivos específicos*

- Perfilar taxonómicamente un metagenoma representativo de sedimentos acuáticos provenientes del río Apatlaco, empleando dos tipos de marcadores: ARNs ribosomales *vs* marcadores de copia única.
- Evaluar si la información contenida en compósitos de genomas del río Apatlaco, replica los perfiles taxonómicos del metagenoma.
- Crear un catálogo taxonómico y funcional de los microorganismos autóctonos del río Apatlaco.

Capítulo 2: FUNDAMENTACIÓN TEÓRICA

2.1 Antecedentes

La cuenca del río Apatlaco en el Estado de Morelos atraviesa por los municipios de Cuernavaca, Jiutepec, Emiliano Zapata, Temixco, Xochitepec, Puente de Ixtla, Tlaltizapán, Zacatepec, Huitzilac y Jojutla (IMTA, 2018). Según la CONAGUA (2012) 13.5% de la superficie del estado, corresponde a la cuenca. Además, una aproximación con datos del INEGI (2020) y de CEAGUA (2017), estima que hay 907,473 personas (48.55% de la población total en Morelos) que viven alrededor de la cuenca, quienes se encuentran en riesgo por la contaminación del río, el cual se calcula que recibe 723.3 l/s de descargas de aguas residuales, y adicionalmente una fracción considerable de las 491,154 toneladas anuales de desechos sólidos que se disponen en el estado (CONAGUA, 2012).

En 2019 y 2020 Breton-Deval et al., al analizar el agua del río Apatlaco identificaron bacterias de los géneros *Acinetobacter, Arcobacter, Prevotella* y *Aeromonas*, destacados por ser patógenos oportunistas (microorganismo que normalmente no es infeccioso, sin embargo, en personas inmunodeprimidas puede llegar a originar y desarrollar una enfermedad (Cisterna, 2007)); también, encontraron microorganismos con potencial en procesos de biorremediación,

evidenciando que algunos microorganismos del Apatlaco tienen capacidad metabólica para tolerar y transformar contaminantes de origen antropogénico. Por otro lado, en Sánchez-Reyes et al., 2020, utilizaron un metagenoma proveniente de los sedimentos de la cuenca, el resultado de la anotación funcional, mostró que habían presentes microorganismos con potencial en procesos de biorremediación, y el perfil taxonómico reveló una composición fundamentalmente anaeróbica o microaerofílica, con los géneros *Methanobacterium* y *Clostridium* como principales miembros en el metagenoma (Sánchez-Reyes et al., 2020a). En otro estudio se determinó la presencia del patógeno emergente *Stenotrophomonas maltophilia* en el río Apatlaco (Ochoa-Sánchez y Vinuesa, 2017). Los estudios sobre el microbioma asociado a la cuenca son limitados, por tanto, hay mucho por explorar sobre la composición y diversidad microbiana presente tanto en el agua superficial como en los sedimentos del río Apatlaco.

Históricamente, la composición taxonómica en el estudio de microbiomas, se estima a partir de la asignación de lecturas de secuenciación masiva al linaje microbiano más probable, por medio del gen ARNr 16S; sin embargo, este método presenta algunas desventajas relacionadas con el diverso número de copias de genes 16S que pueden ir de 1 hasta 16 en bacterias, o de 1 a 4 en arqueas (Sun et al., 2013), así como la heterogeneidad intragenómica entre las diferentes copias dentro de una célula; con diferencias estimadas hasta en un ~6%, con lo que se corre el riesgo de cometer errores de sobreestimación de la diversidad procariótica (Sun et al., 2013). Aunque el marcador ARNr 16S ha sido el estándar para la estimación de perfiles taxonómicos por su bajo costo y facilidad de secuenciación, existen alternativas que emplean marcadores de copia única, estos son genes que están presentes en la mayoría de las especies bacterianas conocidas (en la tabla 7 se enlistan algunos ejemplos de maracdores de copia única) (Na et al., 2018), y por lo general participan en el metabolismo central microbiano. No están sujetos a diversidad intragenómica, rara vez son sujetos de transferencia horizontal y han demostrado mejor robustez para delinear especies y cepas procarióticas en múltiples estudios (Sánchez-Reyes y Folch-Mallol, 2020; Sun et al., 2013). Para una asignación taxonómica eficiente, se debe utilizar la estrategia con menos

sesgo en la estimación de la diversidad, por ello, la importancia de comparar estas dos aproximaciones (16S *vs* marcadores de copia única).

2.2 Marco teórico

2.2.1 Extensión y localización de la cuenca del río Apatlaco

La cuenca del río Apatlaco, nace en la laguna de Zempoala, al norte del Estado de Morelos y desemboca en el río Yautepec (CONAGUA, 2012). En la Figura 1, se observa que la cuenca pasa por los municipios de Cuernavaca, Jiutepec, Emiliano Zapata, Temixco, Xochitepec, Puente de Ixtla, Tlaltizapán, Zacatepec, Huitzilac y Jojutla (IMTA, 2018).

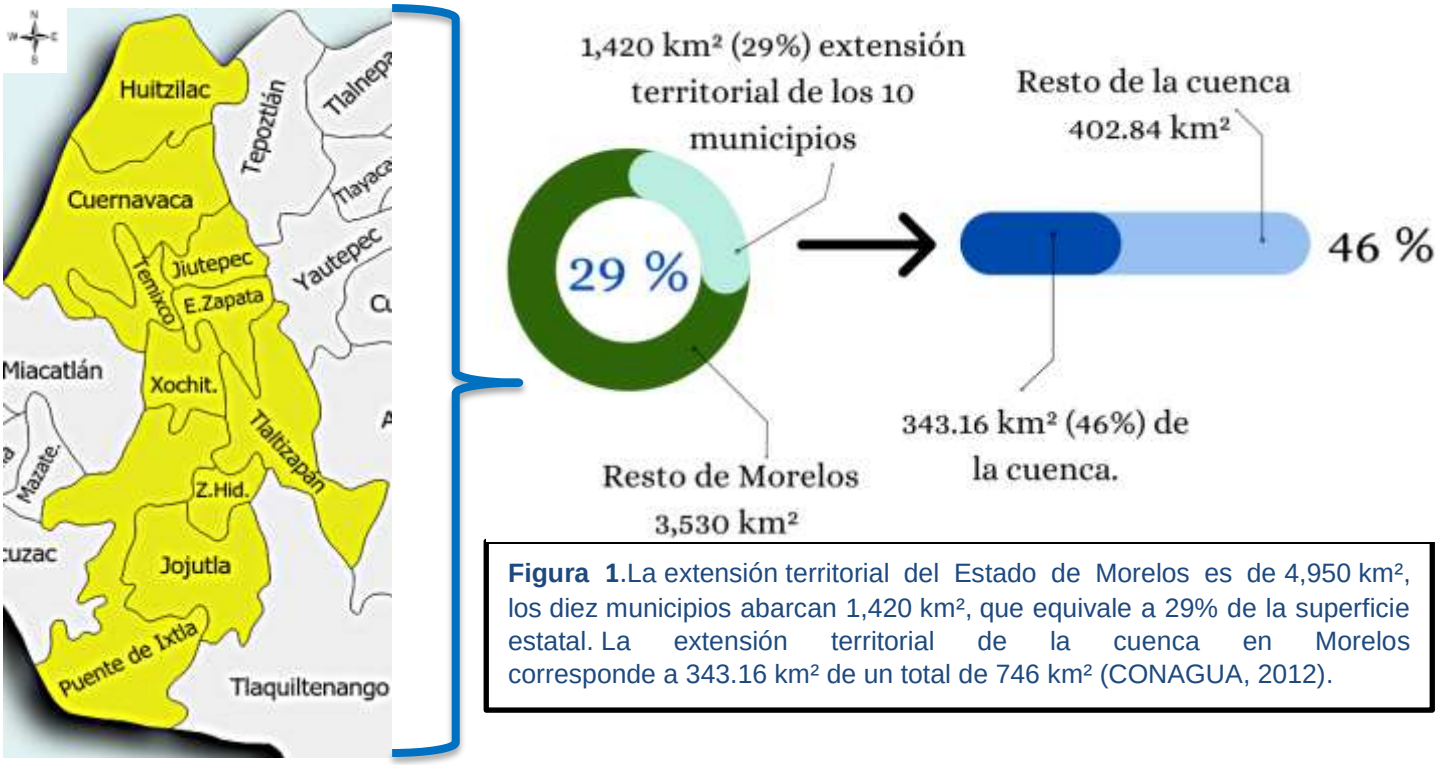


Figura 1.La extensión territorial del Estado de Morelos es de 4,950 km², los diez municipios abarcan 1,420 km², que equivale a 29% de la superficie estatal. La extensión territorial de la cuenca en Morelos corresponde a 343.16 km² de un total de 746 km² (CONAGUA, 2012).

2.2.2 Población en los alrededores de la cuenca

Una aproximación realizada con datos del INEGI (2020) y de SEMARNAT (2017), estima que hay 907,473 personas que viven alrededor de la cuenca. Esta cifra corresponde al 46.02% de la población total en el estado de Morelos, quienes se identifican como habitantes en riesgo por la contaminación del río (Figura 2).

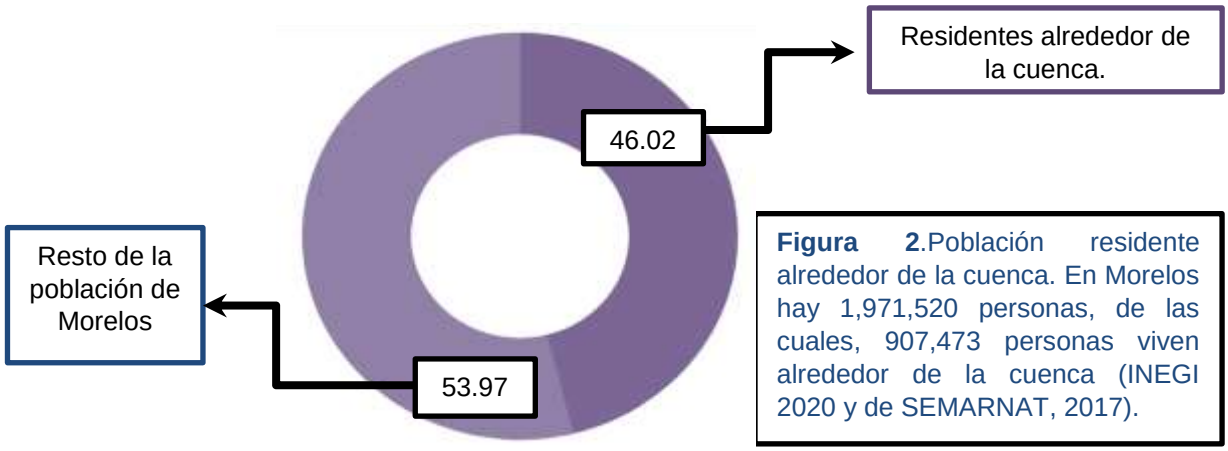


Figura 2.Población residente alrededor de la cuenca. En Morelos hay 1,971,520 personas, de las cuales, 907,473 personas viven alrededor de la cuenca (INEGI 2020 y de SEMARNAT, 2017).

2.2.3 Calidad del agua de la cuenca del río Apatlaco

Un estudio de la Comisión Nacional del Agua, 2012, determinó que el río está contaminado en algunas partes y fuertemente contaminado en otras. El análisis de calidad del agua realizado en 2019 por la Comisión Nacional del Agua (2021) concluyó en que la única zona del río que cumple con la calidad establecida para ser utilizada para el riego, la pesca, la recreación y además cuenta con un ecosistema acuático saludable, es al norte del Estado de Morelos, donde nace la cuenca del Apatlaco. Sin embargo, la mayor parte de la cuenca se encuentra fuertemente contaminada. Los parámetros de calidad utilizados para determinar qué tan buena es la calidad del agua del Apatlaco, según la CONAGUA (2021), se evidencian en las Tablas 1 y 2.

Tabla 1. Parámetros máximos y mínimos de la Demanda Química de Oxígeno (DQO) y Coliformes Fecales (CF), indicadores de contaminación industrial y doméstica respectivamente (CONAGUA, 2021).

Parámetros máximos y mínimos de la DQO		
Estado	DQO	Descripción
Excelente	DQO≤10	Agua no contaminada.
Buena Calidad	10< DQO ≤20.	Aguas superficiales con bajo contenido de materia orgánica biodegradable y no biodegradable.

Aceptable	20< DQO ≤40.	Con indicio de contaminación. Aguas superficiales con capacidad de autodepuración o con descargas de aguas residuales tratadas biológicamente.
Contaminada	40< DQO ≤ 200.	Aguas superficiales con descargas de aguas residuales crudas, principalmente de origen municipal.
Fuertemente Contaminada	200<DQO	Aguas superficiales con fuerte impacto de descargas de aguas residuales crudas municipales y no municipales.
	Parámetros máximos y mínimos de C.F.	
Excelente	CF≤100.	Agua no contaminada o condición normal. No hay evidencia de alteración en los valores de la calidad bacteriológica para el cuerpo de agua superficial.
Buena Calidad	100<CF ≤200.	Aguas superficiales con calidad satisfactoria para la vida acuática y para uso recreativo con contacto primario, así como para otros usos. Indicios de alteración de la calidad bacteriológica.
Aceptable	200< CF ≤1000.	Aguas superficiales con calidad admisible como fuente de abastecimiento de agua potable y para riego agrícola. Muestra bajos niveles de alteración como resultado de la actividad humana.
Contaminada	1000< CF ≤10000.	Aguas superficiales con contaminación bacteriológica. Indica alteración sustancial con respecto a la condición normal.
Fuertemente Contaminada	10000<CF	Aguas superficiales con fuerte contaminación bacteriológica. Alteración severa.

Tabla 2. Resumen de calidad del agua del río Apatlaco. DQO y CF según datos de la (CONAGUA, 2021).

Zona	Lugar	DQO(mg/L)	Estado	CF	Estado
1	Apatlaco abajo Chalchihuapan	<10	Excelente	166	Buena Calidad
2	Apatlaco antes descarga municipal	<10	Excelente	2382	Contaminada

	Tetela				
3	Apatlaco antes Carcamo de San Anton	<10	Excelente	24196	Fuertemente Contaminada
4	Apatlaco arriba Chipitlan	<10	Excelente	24196	Fuertemente Contaminada
5	Apatlaco antes arroyo Chapultepec	<10	Excelente	241960	Fuertemente Contaminada
6	Apatlaco aguas abajo 2 derivadora	<10	Excelente	24196	Fuertemente Contaminada
7	Puente Temixco	<10	Excelente	24196	Fuertemente Contaminada
8	Arriba Ptar El Rayo	<10	Excelente	24196	Fuertemente Contaminada
9	Apatlaco antes arroyo Panocheras	<10	Excelente	24196	Fuertemente Contaminada
10	Apatlaco después de la cascada Real del Puente	<10	Excelente	241960	Fuertemente Contaminada
11	Apatlaco Xochitepec	24.36	Aceptable	241960	Fuertemente Contaminada
12	Apatlaco después arroyo Palo	115.51	Contaminada	241960	Fuertemente Contaminada

2.2.4 Contaminación del agua con énfasis en descargas de la industria textil

En 2012 la CONAGUA, estimó que el 80% de la contaminación del Apatlaco es causada por las descargas municipales vertidas sin tratamiento. Estas descargas son domésticas como industriales. Años atrás en 2008, se declaró que el sector industrial estatal consume 32 millones de m^3 de agua, desechando un total de 25,600,000 m^3 (80%) de aguas residuales a la cuenca (CONAGUA et al. 2008). Del sector industrial, las descargas residuales de la industria textil equivalen a

51,200 m^3, que equivale a un 2% de las descargas de origen industrial (IMTA, 2006).

En el año 2018 se registraron descargas clandestinas de efluentes coloreados a las márgenes del manantial Chapultepec, en Cuernavaca; el cual constituye un brazo del río Apatlaco. La naturaleza de dichas descargas y sus responsables se desconocen en la actualidad. Derivado de lo anterior y de la constante liberación de desechos en el río, el Consejo de Cuencas y Barrancas de Cuernavaca *-una asociación civil instituida para la protección de los cuerpos de aguas estatal-* junto con las personas que viven alrededor de la cuenca dieron a conocer en redes sociales y periódicos, el derrame de aguas residuales de colores; sin embargo, nunca se pudo dar con la procedencia de estas aguas de la industria textil (Editorweb, 2018). También se interpusieron varias demandas municipales para evitar la liberación de efluentes industriales de cualquier naturaleza en la cuenca.

Aparte del problema ambiental causado en el río, se sabe que la contaminación por colorantes textiles puede dar como resultado la selección de microorganismos con capacidad para degradar xenobióticos de estructura relacionadas. Los colorantes textiles más abundantes en el sector industrial poseen naturaleza azoica, antraquinónica o bien suelen ser derivados del trifenilmetano. Estas características químicas los hacen estables a la degradación biológica, química y fotolítica, de ahí que se consideren compuestos recalcitrantes. Una de las apuestas investigativas para biodegradar xenobióticos de estructura compleja, es aprovechar los fenómenos de selección positiva de ocurren en los ambientes contaminados de manera natural. En este sentido el río Apatlaco constituye un reservorio de potenciales microorganismos degradadores de compuestos orgánicos persistentes, como los colorantes textiles.

2.2.6 La microbiodiversidad en la cuenca del río Apatlaco

No existen estudios faunísticos específicos del río Apatlaco, sin embargo, el agua de la cuenca presenta un desequilibrio en el ciclo de nutrientes, lo cual afecta la calidad del agua en color, sabor y en olor; también se registraron mayores florecimientos de algas. Sin embargo, los parámetros fisicoquímicos del Apatlaco

en general no rebasan los límites permisibles para protección de la vida acuática (IMTA, 2007).

Son pocos los análisis realizados, donde se describe la microdiversidad que podemos encontrar en el río Apatlaco. En 2019 Breton-Deval et al., analizó 17 sitios de la cuenca e identificó bacterias de la clase *Gammaproteobacteria, Alphaproteobacteria, Epsilonproteobacteria, Betaproteobacteria,* los géneros *Limnohabitans, Polaromonas, Pedobacter, Thiomonas, Limnohabitans, Polynucleobacter, Myroides, Arcobacter, Pseudomonas, Cellulo phaga, Gillisia, Riemerella, Pedobacter,.* Un análisis reciente de Bretón-Deval et al, (2020) realizó un estudio de 4 sitios seleccionados por sus características fisicoquímicas del Apatlaco, encontró que los sitios más contaminados están enriquecidos en *Acinetobacter, Arcobacter, Prevotella* y *Aeromonas*, destacados por ser patógenos oportunistas potenciales; mientras que el sitio más limpio, es rico en bacterias planctónicas como *Limnohabitans* y *Polaromonas* (Breton-Deval et al., 2020) (Tabla 3). También encontró microorganismos con potencial en la biorremediación, es importante destacar, que entre ellos el género *Pseudomonas* está implicado en la degradación de colorantes textiles y otros xenobióticos (Cortazar-Martínez et al., 2012).

Tabla 3. Microorganismos encontrados en el río Apatlaco (Breton-Deval et al., 2020)**.**

Microorganismo	Patógeno	Potencial en la biorremediación	Descripción
Thiomonas sp.	No	Si	Moreira y Amils, 1997
Polaromonas sp.	No	Si	Irgens et al. 1996
Pedobacter sp.	Si	Si	Steyn et al. 1998
Myroides sp.	Si	Si	Vancanneyt et al. 1996
Pseudomonas sp.	Si	Si	Migula 1894
Acinetobacter sp.	Si	Si	Brisou y Prévot 1954
Aeromonas sp.	Si	Si	Stanier 1943
Arcobacter sp.	Si	No	Vandamme et al. 1991

Prevotella sp.	Si	No	Shah y Collins 1990

2.2.7 Bacterias con potencial para degradar colorantes textiles

Hay numerosos métodos para tratar aguas residuales y degradación de colorantes textiles, específicamente de antraquinona, donde se utilizan métodos como la oxidación de Fenton, oxidación fotocatalítica, oxidación del ozono, oxidación catalítica ultrasónica y catálisis por microondas; sin embargo, es preferible utilizar métodos biológicos, ya que son económicos y no causan contaminación secundaria. Por tanto, se han descrito varias bacterias con potencial para degradar colorantes textiles (Li et al., 2019). En la Tabla 4, se muestran algunos organismos con capacidad para degradar colorantes antraquinónicos.

Tabla 4. Bacterias con potencial para degradar colorantes textiles antraquinónicos (Li et al., 2019)**.**

Bacterias	Tipo de colorante
Pseudomonas sp.	Azul reactivo 2 Verde ácido 27
Staphylococcus hominis subsp. hominis DSM 20328	Azul reactivo 4
Aeromonas hydrophila	Azul brillante reactivo K-GR, Azul ácido 25, Azul ácido 56
Shewanella decolorationis S12 (Xu et al., 2006)	Azul brillante reactivo K-GR
Bacillus cereus	Azul ácido 25, Rojo disperso 11, Azul brillante reactivo K-GR, Azul ácido 56
Bacillus subtilis	Cuba Azul 4, Cuba Azul 4, Cuba Azul 4
Enterobacter sp.	Azul reactivo 19, Azul reactivo 19
Sphingomonas xenophaga	Ácido bromamínico
Serratia liquefaciens	Azul brillante Remazol R
Staphylococcus sp. K2204	Azul brillante Remazol R

Los colorantes textiles derivados de antraquinonas pueden ser degradados por medio de dos mecanismos: adsorción y biodegradación.

a) Adsorción. En este mecanismo, el colorante se adhiere a las bacterias, para posteriormente ser degradado por enzimas.

b) Biodegradación. Este mecanismo conlleva una reacción de reducción mediada por una reductasa, se pierden los cromóforos y posteriormente los anillos

aromáticos se separan, finalmente se descomponen en dióxido de carbono y agua (condiciones anaeróbicas) (Li et al., 2019).

En la Tabla 5, se muestran algunas enzimas para la degradación de colorantes antraquinónicos.

Tabla 5. Enzimas de algunas bacterias involucradas en la degradación de colorantes de antraquinona (Li et al., 2019).

Enzima	Colorante que degrada
Peroxidasa de una Anabaena	Reactive Blue 5, Reactive Blue 4, Reactive Blue 114, Reactive Blue 119 y Acid Blue 45
Peroxidasa de Vibrio cholerae(VcDyP)	Reactive Blue 19
Lacasa de Klebsiella pneumoniae	Colorantes de antraquinona

2.2.8 Marcadores filogenéticos

En un metagenoma (secuenciado por shotgun), se busca describir las comunidades presentes en una muestra significativa del ambiente de interés, para ello, se realiza un perfilado taxonómico de la comunidad y se identifican los organismos potencialmente presentes en la muestra, usando bases de datos taxonómicas preestablecidas, con la ayuda de marcadores moleculares y programas de computación.

2.2.8.1 ARNr 16S

El ARNr 16s, está presente en todos los organismos procariontes, tiene un tamaño promedio de 1500 pb y ha constituido por muchos años el estándar de clasificación taxonómica en los dominios Bacteria y Archaea. La asignación taxonómica empleado marcadores ribosomales es considerada la más usada para realizar perfiles taxonómicos (“Alberts - Molecular Biology Of The Cell,” 2003) ya que consume pocos recursos computacionales (Janda y Abbott, 2007) y se cuenta con grandes bases de datos representativas. Por ejemplo, la base de datos del NCBI para genes 16S (*16S ribosomal RNA (Bacteria and Archaea type strains*)), está compuesta por secuencias curadas de ARN ribosomal 16S que provienen de cepas tipo, cuya taxonomía descrita es muy precisa. Esta base de datos

actualmente posee ~22163 records (consulta 2022/02/03: https://www.ncbi.nlm.nih.gov/nuccore?term=33175%5BBioProject%5D+OR+33317%5BBioProject%5D).

Sin embargo, este método, puede presentar errores debido a las variaciones en el número de copias del ARNr 16S y a la heterogeneidad en la secuencia entre cada una de ellas (~2-6%) (Case et al., 2007; Sánchez-Reyes y Luis Folch-Mallol, 2020). En 2013 Sun et al., estimaron que el número de copias del ARNr 16S van de 1 a 15 en bacterias y de 1 a 4 en arqueas, en la Tabla 6 se observa el número de copias encontradas en algunos filos (Sun et al., 2013). Por ello, se ha considerado otra alternativa para realizar el perfilado taxonómico, que se basa en utilizar genes marcadores filogenéticos de copia única (Segata et al., 2012, Mende et al., 2013, Folch-Mallol y Sánchez-Reyes, 2020).

Tabla 6. Número de copias del ARNr 16S en algunos filos de Bacterias y Arqueas (Sun et al., 2013).

Filo	Promedio de número de copias del gen ARNr 16S
Bacterias	
Firmicutes	6.01 ± 2.82
Fusobacteria	5.40 ± 1.36
Proteobacteria	3,94 ± 2,62
Tenericutes	1.6 ± 0.5
Chlamydiae	1.6 ± 0.8
Acidobacteria	1,3 ± 0,4
Arqueas	
Euryarchaeota	4

2.2.8.2 Marcadores de copia única

Los marcadores bacterianos de copia única son genes homólogos de una sola copia que están presentes en la mayoría de las especies bacterianas conocidas (Na et al., 2018). Además, no están sujetos a diversidad intragenómica y han demostrado mejor robustez para delinear especies y cepas procariotas en múltiples estudios (Mende et al., 2020, 2013).

En la Tabla 7, se muestran algunos marcadores de copia única.

Tabla 7 Algunos genes marcadores filogenéticos universales de copia única empleados en filogenias basadas en metagenómica para la delimitación de especies procariotas (Mende et al., 2013)**.**

Clústeres de grupos de proteínas ortólogos (COG)	Nombre de la proteína	Categoría funcional COG
COG0048	Proteína ribosómica S12	Traducción, estructura ribosomal y biogénesis (J)
COG0049	Proteína ribosómica S7	Traducción, estructura ribosomal y biogénesis (J)
COG0052	Proteína ribosómica S2	Traducción, estructura ribosomal y biogénesis (J)
COG0080	Proteína ribosomal L11	Traducción, estructura ribosomal y biogénesis (J)
COG0081	Proteína ribosomal L1	Traducción, estructura ribosomal y biogénesis (J)
COG0085	ARN polimerasa dirigida por ADN, subunidad beta	Transcripción (K)
COG0087	Proteína ribosomal L3	Traducción, estructura ribosomal y biogénesis (J)
COG0088	Proteína ribosomal L4	Traducción, estructura ribosomal y biogénesis (J)
COG0090	Proteína ribosómica L2	Traducción, estructura ribosomal y

		biogénesis (J)
COG0091	Proteína ribosomal L22	Traducción, estructura ribosomal y biogénesis (J)

2.2.9 Programas para predecir el perfil taxonómico de un metagenoma

Se han desarrollado varios programas computacionales para inferir la composición taxonómica general de un metagenoma (Tabla 8). Para ello, se utilizan métodos como el alineamiento de lecturas contra bases de datos prederminadas, el mapeo de k-meros y su estimación composicional, la extracción de marcadores filogenéticos y su asignación *a posterior*; o bien el alineamiento de genes marcadores traducidos a secuencias de proteínas (Breitwieser et al., 2018).

Tabla 8. Softwares para clasificación taxonómica de metagenomas.

Programa	**Método de clasificación taxonómica**
Focus	Agrupa fragmentos de tamaño *k* (k-mero) y utiliza una base de datos de referencia de 2766 genomas, se basa en la frecuencia de K-meros, emplea mínimos cuadrados no negativos (Genivaldo Gueiros Z. Silva et al., 2014). La base de datos ha sido reciente actualizada a 14,551 genomas provenientes del material tipo: https://github.com/ayixon/RaPDTool
Kaiju	Cada lectura es asignada a un taxón en la taxonomía NCBI y es comparada con una base de datos de referencia (Menzel et al., 2016).
Kraken	Asigna etiquetas taxonómicas a secuencias de ADN, usa alineación de k-mers y un algoritmo de clasificación (Wood y Salzberg, 2014).
MetaPhLan	Perfila la composición de comunidades microbianas utilizando genes marcadores específicos de clado únicos (Segata et al., 2012).

2.2.10 Alcances del perfilado taxonómico

Normalmente el alcance de la asignación taxonómica en un metagenoma es hasta el nivel taxonómico de género, y usualmente no se considera significativa la clasificación a nivel de especie debido a las dificultades intrínsecas del concepto y su demarcación en procariontes. Los niveles taxonómicos superiores por encima del nivel de género (Familia, Clase, Orden, Filo), son cubiertos con bastante éxito

en los clasificadores actuales, ya sea que utilicen amplicones ribosomales o secuencias de genes codificantes. Algo que ha contribuido a la mejora en la clasificación es el enriquecimiento en secuencias de las bases de datos internacionales, sustentado en el auge de la secuenciación masiva y el abaratamiento de los costos de secuenciación. Un estudio realizado en 2019 por Ye et al., donde compararon el rendimiento de programas que asignan taxonomía, arrojó que los clasificadores basados en k-meros (Kraken, Focus, etc), tienen un mejor desempeño para asignación taxonómica, en comparación con los que usan alineamientos de genes marcadores. Por tanto, concluyeron que es mejor realizar una asignación taxonómica basada en ADN que en secuencias de proteínas; en dependencia del tamaño de la base de datos que se utilice como referencia (Ye et al., 2019).

2.2.11 Genomas individuales desde un metagenoma complejo

Una manera de explorar el complemento taxonómico de un metagenoma, es extrayendo o reconstruyendo los genomas individuales que lo componen, a este método se le denomina generalmente binning metagenómico (agrupamiento). En un metagenoma tenemos fragmentos contiguos de ADN (contigs) pertenecientes a diversos microorganismos; y a partir de un agrupamiento de estos contigs se pueden obtener genomas individuales con base en sus propiedades de regularidad composicional, esto es, *frecuencia de tetranucleótidos, porcentajes de Guanina-Citocina* (GC)*, frecuencia de codones*, entre otras. Partiendo de un metagenoma secuenciado por el método shotgun, se puede realizar el binnig a partir de dos estrategias (Sedlar et al., 2017):

a) Agrupando contigs por medio de homología contra un genoma de referencia.
b) Agrupamiento de contigs basándose solamente en características de las secuencias (composición de la secuencia y abundancia) (Sedlar et al., 2017).

El binnig es efectivo para explorar las lecturas que no pudieron ser agrupadas taxonómicamente mediante los marcadores filogenéticos, como suele suceder con

organismos que no son cultivables (Breitwieser et al., 2018). Sin embargo, es computacionalmente costoso y sujeto a cierto sesgo *c*omposicional, puesto que los organismos más abundantes tienen mayor probabilidad de ser resueltos en compósitos genómicos más o menos completos. Los microorganismos menos abundantes (< 1% abundancia relativa) rara vez pueden ser recuperados en fracciones individuales significativas con tamaños mayores a 100 Kpb. Los agrupamientos menores a este límite se descartan del proceso de binning puesto que toman como referencia el tamaño mínimo de un genoma bacteriano (*Carsonella ruddii* 159,662 Kpb) (Ball, 2006). Entre las herramientas más populares para hacer binning están los softwares Metabat2, Maxbin, Concoct entre otros (Kang et al., 2019; Alneberg et al., 2014; Wu et al., 2014).

Al obtener los genomas individuales desde un metagenoma, el siguiente paso es aplicar una estrategia para determinar a qué especie y/o género pertenece, o bien establecer su contexto taxonómico más fino posible. En este sentido se han desarrollado varias herramientas de muy alta resolución como son GTDB-Tk y Anvio (Chaumeil et al., 2020; Eren et al., 2021).

2.2.12 Delimitación de género y especie de genomas individuales

Desde aproximadamente el año 1963, se han propuesto más de 20 conceptos de especie; el concepto más popular de especie es el denominado biológico, dicta que una especie es “una población cuyos individuos se reproducen entre sí y producen descendencia fértil” (Leopardi y Duno, 2010; Shanker et al., 2017). Este concepto posee como limitación que no es aplicable al grupo de los procariontes dado que en estos no hay reproducción sexual reconocible. Por otro lado, Rosselló-Mora y Amann (2001) describe para los procariontes, que se habla de la misma especie cuando existe coherencia genómica y un alto grado de similitud en características independientes; se presenta una propiedad fenotípica en común y pertenecen a un grupo monofilético (concepto filofenético). Este último concepto ha sido útil durante muchos años para delinear especies microbianas y recientemente los criterios genómicos han ganado mayor relevancia debido a la disponibilidad de secuencias. Estos criterios genómicos se han resumido en los

denominados índices de relación genómica global (OGRI por sus siglas en inglés) (Chun y Rainey, 2014), los cuales estiman desde las secuencias genómicas parámetros estables entre taxones de la misma especie. Podemos citar entre estos a la identidad promedio a nivel de nucleótidos y aminoácidos (ANI y AAI en inglés), la distancia genómica o la frecuencia de alelos estimada con base en la frecuencia de k-meros, ente otros (Tabla 9). Para este proyecto, estudiaremos diferentes parámetros genómicos por medio de herramientas bioinformáticas, para delimitar género y/o especie, tomando como referencia la observación que plantea que una especie procariótica generalmente está representada por un grupo de organismos que presentan un ANI ≥95% y una distancia genómica de Mash (D) ≤0.05. A esto habría que sumarle otros criterios complementarios como la hipótesis filogenética de monofilia y la de los caracteres diagnósticos fenotípicos. La delimitación cuantitativa de género se entiende como un grupo de organismos que presentan 75% ≤ANI< 95% y cuya fracción de alineamiento (FA) sea ≥33.33% (Barco et al., 2020); así como el parámetro complementario de AAI >55%. Sin embargo, estos parámetros no son universales y se deben evaluar caso por caso, dado que existen corrimientos en los límites establecidos que solo dependen del grupo taxonómico estudiado. Además, en especies nuevas o en grupos nuevos a niveles taxonómicos mayores, los estándares genómicos actuales suelen tener variaciones puntuales. Otros parámetros que se pueden utilizar para lograr determinar la especie y/o el género de un genoma se pueden observar en la Tabla 9.

Tabla 9. Parámetros genómicos para delimitar género y/o especie.

Parámetro genómico	Descripción	Rango para delimitar género	Rango para delimitar especies
Distancia genómica de MASH	Calcula la distancia aproximada entre dos secuencias (Ondov et al., 2016).	-	≤0.05 (equivale a ANI≥95%)
FA	La fracción de alineamiento es el promedio del total de alineamiento de genoma A contra el genoma B y viceversa.	≥33.33%	-

%ARN 16S	Es la comparación de las secuencias de ARNr 16S (identidad del ARNr 16S), no se considera oficialmente como un parámetro genómico (Qin et al., 2014).	≥94%	≥97%-98.7% (no se considera adecuado para delimitar especie)
ANI	Identidad promedio de nucleótidos (Konstantinidis et al., 2005).	≥73%	>95%
AAI	Identidad promedio de aminoácidos (Rodriguez-R & Konstantinidis K, 2014).	≥55%	≥85%
DDH	Hibridación DNA-DNA (Wayne et al., 1987).	-	>70% (97%-98.65% de ARNr 16S)
MiSI	Identificador de especies microbianas, que emplea fracciones de alineación (AF) y ANI para la demarcación de especies (Varghese et al., 2015).	FA≥33.33% y ANI≥73%	FA≥33.33% y ANI≥95%
GOC	Sintenia en genes (sentido de traducción de los genes) (Kanyó y Molnár, 2016).	-	<66%
POCP	Es el porcentaje de proteínas conservadas entre un par de genomas (Qin et al., 2014).	≥50%	-

2.2.13 Anotación funcional

La anotación funcional es definida como el proceso de colectar información sobre la identidad biológica de un gen, su función molecular, su función biológica, localización, regulación y parámetros de expresión, entre otros detalles ontológicos. En términos prácticos, la anotación funcional se refiere al proceso de asignar función a una secuencia de ADN (codificante o reguladora). Esto se lleva a cabo por medio de la identificación de genes ortólogos en bases de datos curadas. El proceso generalmente involucra alineamiento de las secuencias a consultar contra la base de datos de referencia, y la determinación de significación de estos alineamientos. La anotación funcional se explora desde dos frentes, uno es el trabajo de confirmación experimental en los laboratorios y otro es el de la inferencia computacional. En el inicio de este párrafo nos hemos referido a este último frente. Constituye un problema biológico que primero se encarga de inferir homología, y luego se encarga de inferir función con base en la homología (Pagni y Jongeneel, 2001).

Existen servidores y herramientas dedicadas a la anotación funcional de secuencias a partir de genomas y metagenomas o bien de sus archivos de predicción de proteínas. Entre estos mencionamos a la base de datos KEGG (Kyoto Encyclopedia of Genes and Genomes) la cual contiene un set de proteínas y vías metabólicas curadas experimentalmente, además de varias herramientas de anotación como los servidores KAAS (KEGG Automatic Annotation Server) y los servicios de anotación automática implementados en las herramientas *Koala* (Blast, Ghost y Kofam). Todas estas herramientas son amigables con el usuario, de libre acceso y están ampliamente documentadas, lo que facilita su utilización. La anotación funcional en el sistema KEGG funciona agrupando los ortólogos en grupos a los que se le asignan etiquetas denominadas KO (números K u ortólogos de KEGG). Estos ortólogos están asociados a una función metabólica-molecular de las secuencias. Los KO representan ortólogos funcionales que participan en diferentes vías metabólicas y que son clasificados manualmente con base a resultados experimentales (Kanehisa y Sato, 2020).

Capítulo 3.

3. METODOLOGÍA

3.1 Obtención de los metagenomas objeto de estudio

Para el desarrollo de este proyecto se partió de dos metagenomas obtenidos del río Apatlaco, uno que proviene de una muestra de sedimentos (Sánchez-Reyes et al., 2020), y el segundo metagenoma proveniente de agua superficial (Breton-Deval et al., 2020). A continuación, se describen los detalles de toma de muestras y su procesamiento, sin embargo, es importante aclarar que estos procedimientos constituyen el antecedente de este trabajo, y que no fueron desarrollados directamente en esta tesis.

3.1.2 Metagenoma de agua superficial

De acuerdo con lo descrito por Breton-Deval et al., 2020, a lo largo del río se seleccionaron 4 sitios con base a un estudio integrativo de calidad del agua realizado en 2019 (Breton-Deval et al., 2019). Dicho estudio exploró un total de 17

puntos de muestreo que fueron clusterizados para seleccionar aquellos que fueran más representativos según sus parámetros físico-químicos. Los puntos y su ubicación se detallan en la Tabla 10.

Tabla 10. Puntos de muestreo a lo largo del río Apatlaco y su ubicación (Breton-Deval et al., 2019).

Punto de muestreo	Nombre	Latitud	Longitud	Estado de contaminación
P1	Apatlaco abajo Chalchihuapan	−99.26872	18.97372	Buena calidad
P7	Puente Temixco	−99.2187	18.83	Fuertemente Contaminado
P10	Apatlaco después de la cascada Real Del Puente	−99.23337	18.78971	Fuertemente Contaminado
P17	Río Apatlaco Tlatenchi	−99.18278	18.60914	Fuertemente Contaminado

De estos cuatro puntos se tomaron muestras de agua superficial para extracción de ADN utilizando el kit DNeasy PowerWater (QIAGEN, Hilden, Alemania). Para cada muestra, se preparó una biblioteca de Illumina a partir de ADN total utilizando el kit TruSeq v2 (Illumina, Inc., San Diego, CA, EE. UU.). Posteriormente, la secuenciación se realizó en la plataforma NextSeq500 (Illumina, Inc., San Diego, CA, EE. UU.), generando 108,785,988 lecturas shutgun con una longitud de 75 pb. Con estas lecturas se creó un ensamble consenso representativo del río (sitios P1-P17) utilizando el ensamblador MEGAHIT versión 1.2.9 (Li et al., 2015) con los parámetros implementados por defecto en el mismo. Este ensamble posee un tamaño de 550,428,717 pb y 746,031 contigs, fue utilizado como referencia para la resolución de genomas individuales a partir de otro set de datos de secuenciación masiva que será descrito más adelante.

3.1.3 Metagenoma de sedimentos

Según lo descrito por Sánchez-Reyes et al., 2020, se recolectaron cuatro muestras de sedimentos del río Apatlaco de los cuatro puntos señalados anteriormente (~0.5 kg por cada punto); las muestras se mezclaron para crear un solo compósito representativo de los sedimentos del Apatlaco, para

posteriormente ser enriquecido y cultivado por 30 días con 200mg mL^{-1} de un colorante comercial de naturaleza antraquinónica (Deep-Blue 35). Finalizado este tiempo, se extrajeron 10g de sedimentos más 10mL de agua remanente para ser sometidos a un tratamiento con un crosslinker (formaldehído al 1%), con el objetivo de crear uniones covalentes en el ADN intracelularmente. Con esta muestra, se creó una biblioteca tipo Hi-C (compatible con Illumina), empleando el kit Microbioma ProxiMeta Hi-C de la compañía Phase Genomics (Seattle, WA). Las lecturas tipo Hi-C obtenidas (72,007,031 lecturas de 150 pb) son de naturaleza quimérica, ya que, poseen uniones formadas por regiones de DNA cercano espacialmente pero no contiguo. La principal ventaja de este tipo de lecturas es que permiten crear un mapa de distancias intracromosomal, útil para la resolución de genomas individuales a partir de alguna referencia metagenómica. Alternativamente, las lecturas fueron divididas (Split) en dos partes para eliminar las regiones de quimerismo, y posteriormente fueron ensambladas con MEGAHIT v1.1.3 obteniéndose un ensamble de ~ 158,650,210 pb (Fernández-López et al., resultados no publicados).

3.2 Esquema metodológico general desarrollado en esta tesis

Este proyecto se compone de 3 objetivos que se ilustran esquemáticamente en el siguiente trazado metodológico (Figura 3).

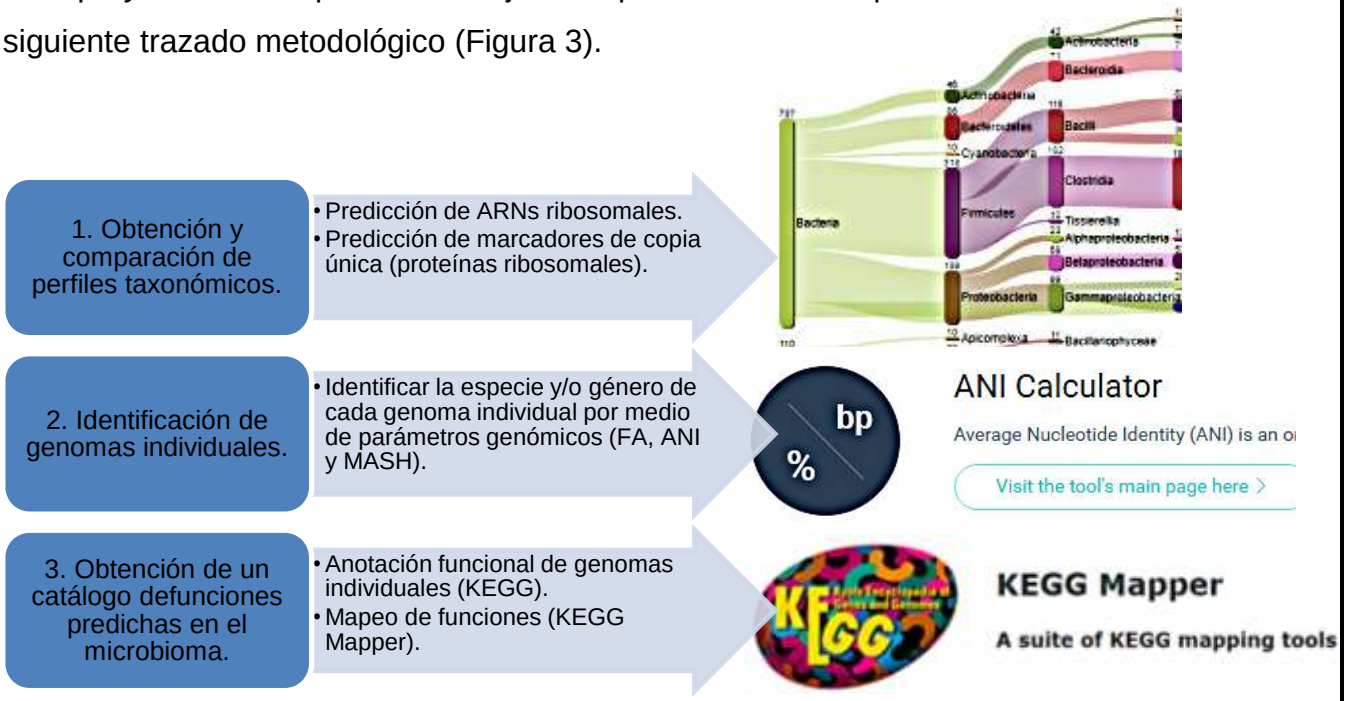


Figura 3. Resumen de las actividades planeadas para alcanzar los objetivos del proyecto y llevar a cabo la prueba de hipótesis.

3.3 Obtención y comparación de perfiles taxonómicos en el metagenoma proveniente de sedimentos

A partir del ensamble obtenido de los sedimentos, se extrajeron las secuencias de ARNs ribosomales y los marcadores de copia única para posteriormente predecir el paisaje taxonómico con cada método.

3.3.1 Predicción de ARNs ribosomales

Para ejecutar la mayoría de las aplicaciones computacionales necesarias y llevar a cabo los análisis se instaló una terminal bash de Ubuntu en Windows (Windows Subsystem for Linux -WSL- en inglés), en su versión 18.04. Esta terminal nos permitió ejecutar código bioinformático de manera similar a un sistema UNIX nativo.

1) Se instaló desde la terminal el paquete Barrnap (Seemann, 2013), que constituye un predictor de ARNr empleando modelos ocultos de Markov. Se utilizó el comando:

```
$ conda install -c bioconda -c conda-forge barrnap
```

2) Partiendo del ensamble crudo representativo del metagenoma de sedimentos del río Apatlaco, se estimaron las secuencias presentes de ARNr eucarióticos y procarióticos siguiendo las siguientes instrucciones:

Bacterias

```
$barrnap --kingdom bac -o bacterias.txt <nombre del ensamble>
```

Arqueas

```
$barrnap --kingdom arc -o arqueas.txt <nombre del ensamble>
```

3) Posteriormente se concatenaron los dos archivos resultantes para obtener un solo archivo de genes ribosomales procariontes:

$cat bacterias.txt arqueas.txt > ARNs_concatenado.txt

3.3.2 Predicción de marcadores de copia única

Para la predicción de marcadores de copia única se utilizó la herramienta UBCG: https://www.ezbiocloud.net/tools/ubcg (Na et al., 2018), que predice un set máximo de 92 marcadores compartidos en bacterias y archaeas. Para cubrir la población de marcadores presentes en el metagenoma, este se dividió en fragmentos de aproximadamente 3 Mpb con la función: *$split --bytes=4000000*. Estos fragmentos fueron considerados como "*genomas artificiales*" representativos ya que poseen un tamaño promedio cercano al tamaño de un genoma procariote. Para cada "*genoma artificial"* se predijo su contenido de marcadores de copia única con el comando de UBCG:

$ java -jar UBCG.jar extract -bcg_dir bcg -i fasta/$i -label $i –t 2

La variable *$i* fue añadida a un ciclo de bash y representa cada uno de los "*genomas artificiales"*. Su ejecución fue de la siguiente forma: *$ ./ cicloUBCG.sh*

Los marcadores de copia única obtenidos para cada fragmento de genoma artificial fueron posteriormente concatenados en un solo archivo y representan la totalidad predicha de marcadores de copia única en el metagenoma, la cual ascendió a 6236.

3.3.3 Perfilado taxonómico

Para llevar a cabo el perfilado taxonómico en el metagenoma de sedimentos usamos tres estrategias. Primero realizamos un perfil sobre el ensamble metagenómico crudo; este perfil lo consideramos nuestro modelo nulo ya que contiene una mayor representación de secuencias provenientes de las lecturas. Posteriormente utilizamos los archivos con los marcadores ribosomales (956 secuencias) y los marcadores de copia única (6236 secuencias) para crear perfiles taxonómicos marcador-específicos y poder comparar su desempeño en la exploración del paisaje de microbiodiverdidad. Se usaron dos herramientas

perfiladoras, Focus (https://github.com/metageni/FOCUS) que posee una base de datos curada de 2,766 genomas microbianos completos (DOI: 10.7717/peerj.425/supp-1) para comparación (Genivaldo Gueiros Z Silva et al., 2014); y la implementación en línea de Kaiju (https://kaiju.binf.ku.dk/server) empleando la base de datos de nucleótidos no redundantes del NCBI (NCBI BLAST nr +euk) (Menzel et al., 2016). Las salidas fueron procesadas en Excel y la salida taxonómica de Kaiju fue transformada a un formato compatible con la taxonomía del paquete Kraken y Pavian para análisis y visualización de los resultados de la clasificación taxonómica (Breitwieser y Salzberg, 2020; Wood y Salzberg, 2014). Un diagrama general del proceso de perfilado taxonómico se presenta en la Figura 4.

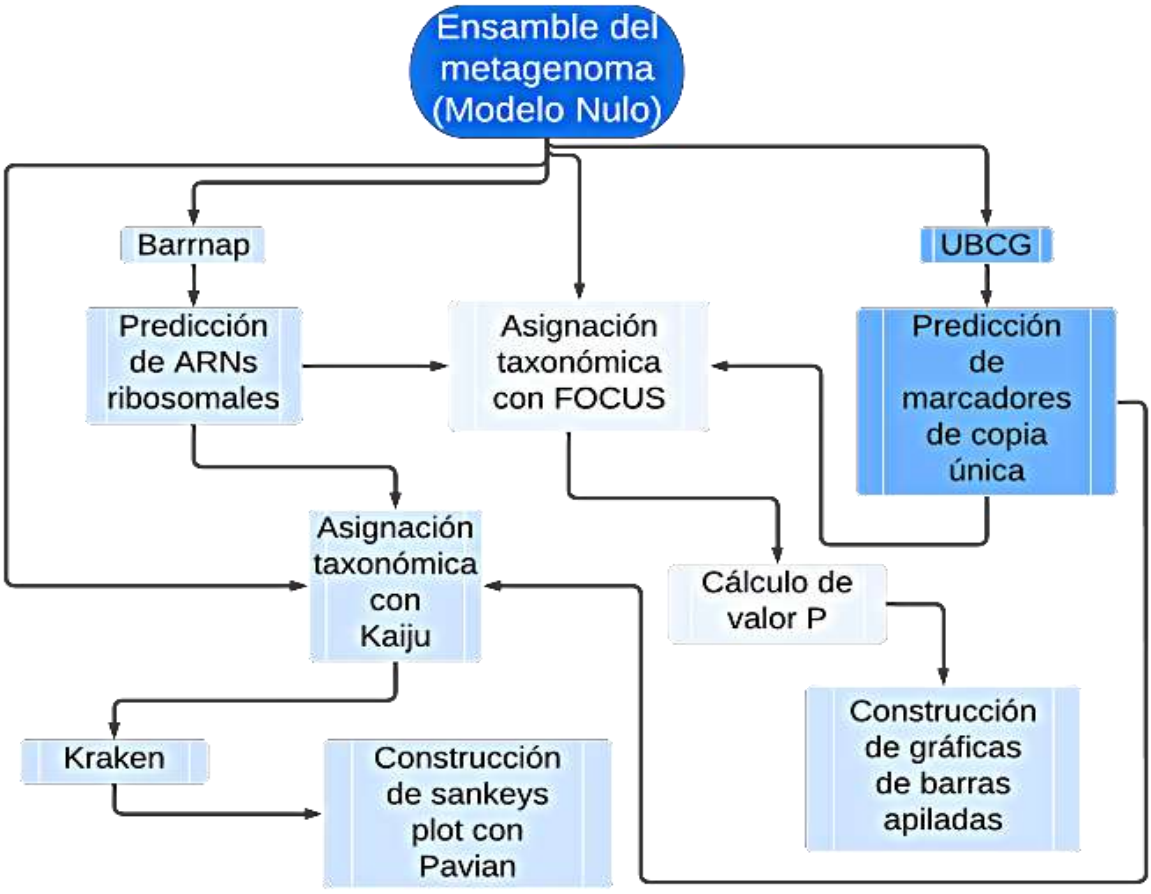


Figura 4. Diagrama de flujo representando el proceso de clasificación taxonómica. Partiendo del ensamble del metagenoma crudo, se extrajeron los marcadores de copia única y los ARNs ribosomales, posteriormente se predijo taxonomía utilizando Kaiju y Focus, para finalmente realizar los cálculos estadísticos y determinar que método (marcadores de copia única vs ARNs ribosomales) describe mejor el perfil taxonómico de nuestro

3.3.4 Comparación de los perfiles taxonómicos.

Se compararon los perfiles de abundancia relativa resultantes del perfilado con el software FOCUS para cada ejercicio de asignación taxonómica (ARNs ribosomales *vs* marcadores copia única). Se utilizó como modelo nulo o *Ground truth* los perfiles de abundancia relativa obtenidos para el ensamble metagenómico crudo. La comparación estadística fue realizada mediante análisis binomial, con el estadístico Chi^2 teniendo en cuenta un test de dos colas, con la ayuda de la calculadora en línea: http://www.vassarstats.net/binomialX.html? El contraste de hipótesis se realizó teniendo en cuenta un nivel de significancia de 0.05.

3.3.5 Identificación de genomas individuales (objetivo 2)

Previamente, en el grupo de laboratorio se había deconvolucionado un set de 34 genomas individuales a partir de los metagenomas de agua superficial y de sedimentos anteriormente descritos. En esta tesis llevamos a cabo la identificación de estos genomas por medio de la comparación de los índices de relación genómica ANI y distancia genómica de Mash (Lee et al., 2016a; Ondov et al., 2016b). Se utilizó una base de datos personalizada que contiene 31,910 records genómicos procedentes de la base de datos GTDB (Genome Taxonomy Database). Esta base de datos está disponible en: https://figshare.com/ndownloader/files/30863182. Además, se exploró si los genomas individuales obtenidos del metagenoma están representados en los perfiles obtenidos desde los marcadores filogenéticos (Figura 5).

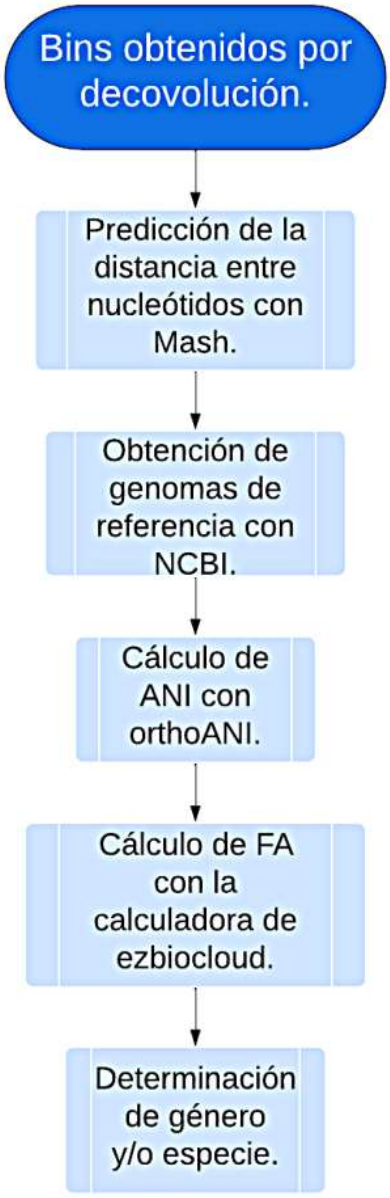


Figura 5. Diagrama de flujo del objetivo 2. Partiendo de 34 bins obtenidos por decovolución, se comparó cada uno contra la base de datos GTDB. Calculamos su distancia de Mash, ANI y fracción de alineamiento (FA), para estimar el género y en su caso la especie del genoma en cuestión.

3.3.5.1 Calculo de Identidad Promedio de Nucleótidos (ANI)

La Identidad Promedio de Nucleótidos se estimó mediante una implementación en Python del algoritmo orthoANI https://pypi.org/project/orthoani/ (Lee et al., 2016), en el cual solo los ortólogos de cada genoma comparado tributan al cálculo del índice: *$ orthoani -q sequence1.fasta -r sequence2.fasta*.

3.3.5.2 Cálculo de la Fracción de Alineamiento (FA)

La fracción de alineamiento se obtuvo mediante la calculadora en línea implementada en el sitio ezbiocloud: https://www.ezbiocloud.net/tools/ani (Yoon et al., 2017).

3.3.5.3 Identificación de especies y/o géneros

Se determinó la especie y el género tomando en cuenta la distancia genómica de Mash, el ANI y la FA, con los parámetros mostrados en la Tabla 11. La clasificación es con respecto a los genomas de referencia comparados en la base de datos. El criterio de ponderación implica que ANI >95% es indicativo de relaciones especie-específicas entre los taxones comparados; ANI<95% es inconclusivo. La distancia de Mash < 0.05 es indicativo de relaciones a nivel de especie, pero se debe corroborar con el ANI, ya que, la distancia de Mash por sí sola no es suficientemente resolutiva. En ausencia de criterios para asignación a nivel de especie, se puede estimar el género haciendo uso de la FA (Barco et al., 2020).

Tabla 11. Parámetros utilizados para delimitar género y/o especie (Barco et al., 2020).

ANI	FA	Distancia Mash	Mismo género	Misma especie
≥73.10%	≥33.33%	≤0.2	Sí	No
≥95%	≥33.33%	≤0.05	Sí	Sí
<73.10%	<33.32%	>0.2	No	No

3.3.6 Anotación funcional (objetivo 3)

Cada bin o genoma reconstruído desde el metagenoma del punto anterior, se introdujo en el servidor MetaGeneMark (http://exon.gatech.edu/meta_gmhmmp.cgi) (Zhu et al., 2010) para predecir las secuencias de genes codificantes y aminoácidos en cada genoma individual. Posteriormente se usó el servidor en línea GhostKOALA (https://www.kegg.jp/ghostkoala/) para realizar la anotación funcional por medio de asignación de un número K (KO) correspondiente a una función molecular

especifica. La salida es un archivo de dos columnas que identifica cada gen con su KO correspondiente. Este archivo fue posteriormente usado para mapear las funciones metabólicas en el servidor KEGGMapper (https://www.genome.jp/kegg/tool/map_pathway.html).

3.3.6.1 Catálogo funcional de microorganismos autóctonos del río Apatlaco

Se realizó una tabla resumen con la frecuencia de genes con funcionalidad en cada genoma, junto con su porcentaje de anotación. Cabe señalar que en este proyecto, no nos limitamos a identificar enzimas relacionadas a la degradación de colorantes textiles, ya que las bases de datos utilizadas no son especializadas en este tema, por ello, se identificaron enzimas con potencial para la degradación de xenobióticos en general.

4. RESULTADOS Y DISCUSIÓN

4.1 Predicción de ARNs ribosomales

A partir de un ensamble metagenómico obtenido de sedimentos del río Apatlaco, se llevó a cabo la predicción de secuencias ribosomales empleando la herramienta Barrnap (Seemann, 2018). Para ello se consideraron los dominios Bacteria y Archaea y Eucariotas. Como se muestra en la Tabla 12, se obtuvo mayor cantidad de secuencias inferidas del dominio bacteria (491) en comparación con arqueas (465) y eucariotas (173).

Tabla 12. Cantidad de secuencias (16S, 23S, 5S, 18S, 28S y 5.8S) obtenidas con Barrnap ordenadas por dominios.

Dominio	Tipo de secuencia	Cantidad inferida (Barrnap)	Total
Bacterias	16S	167	491
	23S	114	
	5S	210	
Arqueas	16S	153	465
	23S	102	
	5S	210	
	5.8s	31	
Eucariotas	18S	71	173
	28S	71	

	5S	210
	5.8S	31

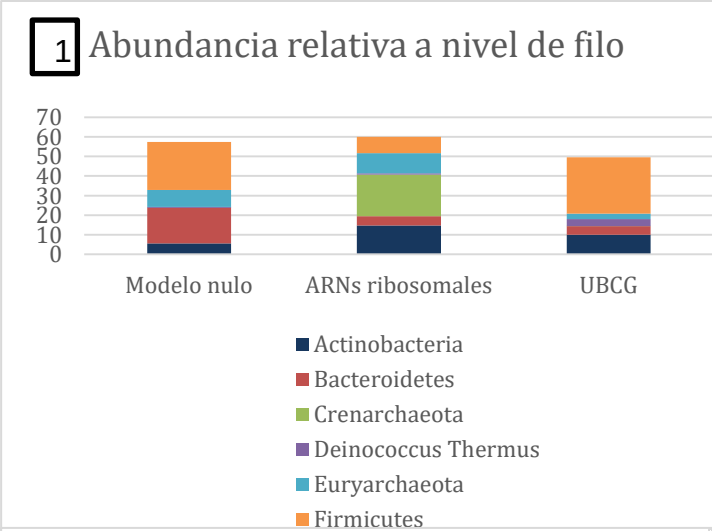
4.2 Predicción de marcadores de copia única.

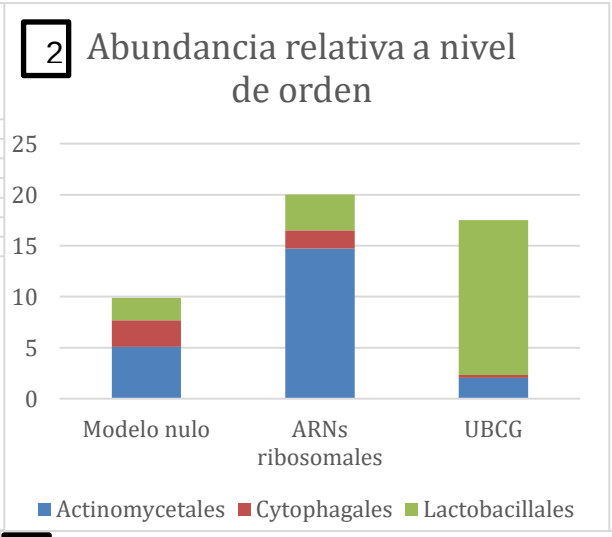
Para la predicción de marcadores de copia única, se fraccionó el metagenoma en 146 fragmentos, cuyo tamaño aproximado fue de 3 Mpb. Con cada fragmento (denominados en este trabajo como "*genoma artificial"*) se llevó a cabo la predicción empleando el software UBCG (Na et al., 2018). Se predijeron en total 6236 marcadores de copia única (UBCGs) que corresponden a funciones moleculares altamente conservadas en procariontes. El promedio de UBCGs por cada "*genoma artificial"* fue de 43, el mínimo de UBCGs predichos fue de 21 y el máximo 66.

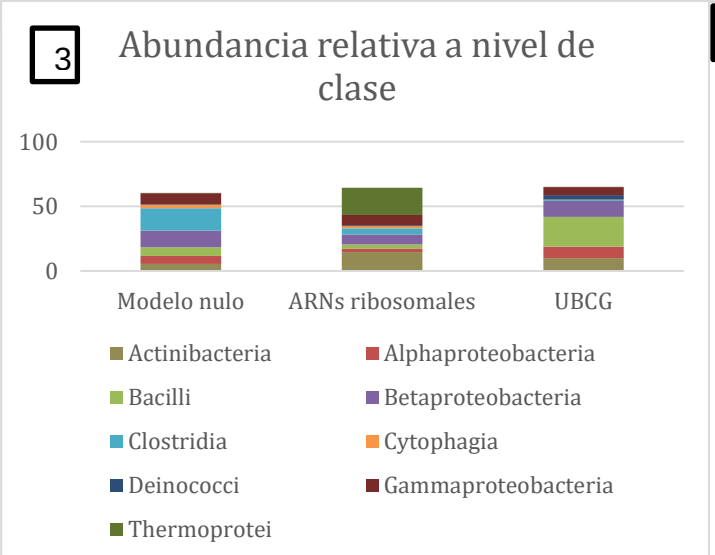
4.3 Perfiles taxonómicos de los sedimentos en tres escenarios

Decidimos perfilar la composición microbiana de los sedimentos empleando el software FOCUS (Silva et al., 2014) bajo tres escenarios de predicción, un modelo nulo que contiene todas las secuencias incluidas en el ensamble metagenómico crudo (contigs), un modelo utilizando solo las secuencias de ARN ribosomal predichas para bacterias y archaeas (956 secuencias). FOCUS no posee referencias eucarióticas en su base de datos por lo que este grupo no se consideró en la predicción. Por último un modelo empleando los 6236 marcadores de copia única (UBCGs) extraídos del metagenoma. Los resultados para diferentes niveles taxonómicos se muestran en la figura 6. Tanto los ARN ribosomales como los UBCGs tienen la capacidad de detectar más filos que el modelo nulo, con diferencias en los patrones de abundancia relativa. Las Actinobacterias poseen patrones similares entre los marcadores (ARN-UBCG), pero están menos representadas en el modelo nulo (~5%). Sin embargo, este último refleja el filo Bacteroidetes con mayor abundancia relativa mientras que el grupo de los Firmicutes, aunque es detectado en los tres modelos, refleja menor abundancia relativa cuando se emplean los ARN ribosomales. La abundancia relativa de las Archaeas se muestra menor en los UBCGs, lo cual puede deberse

a que las archaeas comparten un menor número de estos marcadores respecto a las bacterias (El promedio de UBCG es de 25-30 aproximadamente en Archaeas). Estos mismos patrones diferenciales en la abundancia relativa pueden encontrase hacia niveles más finos de clasificación (orden, clase, familia). Sin embargo, la capacidad de detección varía poco para los grupos más abundantes. Por ejemplo, los órdenes más abundantes (Actinomycetales, Cytophagales y Lactobacillales) son detectados en los tres modelos, pero con significativas diferencias en los perfiles de abundancia relativa. En cualquier caso, el perfil de la comunidad en los sedimentos contiene taxones diversos del grupo de las archaeas y las bacterias, principalmente del grupo de los Metanobacterias y el género Clostridium.







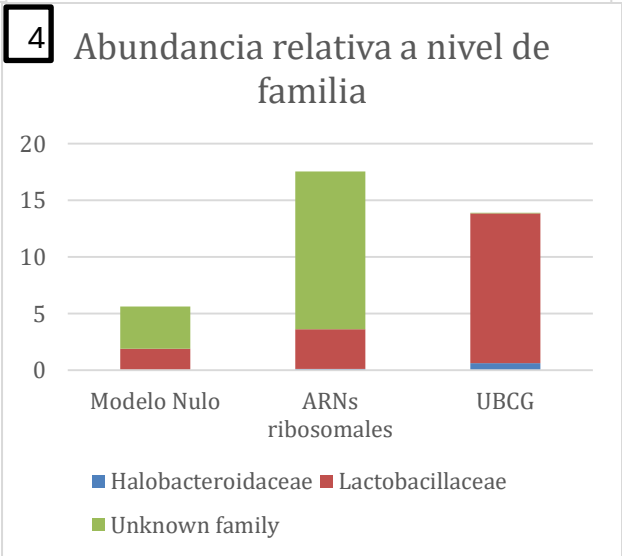


Figura 6. Abundancia relativa de los diferentes niveles taxonómicos, con cada método empleado. 1) A nivel de filo, usando los ARNs ribosomales se obtuvo mayor cantidad del filo Crenarchaeota, mientras que con UBCG y el modelo nulo, se obtuvo mayor cantidad del filo Firmicutes. 2) Se predijo mayor cantidad del orden Actinomycetales utilizando los ARNs ribosomales y el modelo nulo, usando los UBCGs resultó mayor el orden de los Lactobacillales. 3) A nivel de clase, utilizando los ARNs ribosomolas se obtuvo mayor abundancia de Thermoprotei, usando UBCG, hay mayor abundancia de Bacili, mientras que el modelo nulo presenta mayor abundancia de Clostridia. 4) La abundancia de la familia Lactobacillaceae es mayor con UBCGs, mientras que con ARNs ribosomales y el modelo nulo la abundancia de Lactobacillaceae es casi nula.

4.4 Comparación estadística de los perfiles obtenidos mediante ARNs ribosomales *vs* marcadores de copia única, utilizando un modelo nulo (ensamble metagenómico crudo)

Con el objetivo de contrastar la hipótesis relacionada con que los marcadores filogenéticos utilizados definen los patrones de microbiodiversidad diferenciales (en otras palabras, el mapa de microbiodiversidad depende de los marcadores empleados para perfilar la taxonomía); decidimos evaluar estadísticamente las diferencias en cuanto a abundancia relativa entre los perfiles arrojados por los ARN ribosomales y los marcadores de copia única (UBCGs). Para este análisis asumimos que la abundancia relativa asociada al ensamble metagenómico crudo representa un modelo de hipótesis nula, cuyas frecuencias serían las esperadas bajo nuestras condiciones. Con base, en esta premisa, el contraste de hipótesis se resume a evaluar si los patrones de abundancia relativa obtenidos con los ARNs ribosomales y los UBCGs difieren significativamente de aquellos obtenidos con el modelo nulo (H_0). Para el nivel de dominio el patrón de abundancia relativa no difiere significativamente entre el modelo nulo y los UBCGs, mientras que los ARNs ribosomales sí difieren del modelo nulo (Tabla 13), por tanto, para este nivel podemos concluir que los marcadores de copia única (UBCGs), describen perfiles de abundancia relativa en el metagenoma mucho más acordes con el modelo nulo; a diferencia de los perfiles arrojados por las secuencias de ARN.

Tabla 13. Cálculo de la distribución binomial (valor p) a nivel de dominio.

	Distribución binomial (valor P)	Decisión

Arqueas		
UBCG	0.056133	p>0.05 se acepta H_0
ARNs	<0.000001	p<0.05 se rechaza H_0
Bacterias		
UBCG	0.056133	p>0.05 se acepta H_0
ARNs	<0.000001	p<0.05 se rechaza H_0

Los filos más abundantes como Actinobacteria, Euryarchaeota y Firmicutes estimados con con UBCG, se comportan de forma similar al modelo nulo respecto al patrón de abundancia relativa (p>0.05); con la excepción de Bacteroidetes (Tabla 14). Este último caso también es diferencial utilizando los marcadores de RNA ribosomal (p < 0.001002). Esto puede estar relacionado con que el grupo de los Bacteroidetes es taxonómicamente críptico, y su representación en las bases de datos no está tan enriquecida como sucede con otros grupos. Otro detalle importante es que existe un *bias* composicional hacia las secuencias de organismos cultivables, de manera que los fragmentos de secuencias de organismos novedosos a nivel de especie no son fácilmente asociados con taxonomías precisas en los niveles superiores. Una tendencia similar se observa con las clases y órdenes más representativos del perfil de FOCUS; donde es más frecuente encontrar a los UBCGs arrojando valores de abundancia relativa similares al modelo nulo (excepto para las clases Bacilli y Clostridia). Las alfa y gamma proteobacterias son, sin embargo, clases que se resuelven correctamente con los marcadores ribosomales, lo cual no resulta sorprendente pues el grupo de las proteobacterias se considera el más representado en las bases de datos internacionales.

Tabla 14. Cálculo de la distribución binomial (valor p) para diferentes niveles taxonómicos. Se muestran los niveles con clasificación usando los dos tipos de marcadores (UBCG y ARNr).

Filo	**Distribución binomial UBCG**	**Decisión**	**Distribución binomial ARNr**	**Decisión**
Actinobacteria	0.078408	p>0.05 se	0.000078	p<0.05 se

		acepta H0		rechaza H0
Bacteroidetes	0.182087	p>0.05 se acepta H0	0.001002	p<0.05 se rechaza H0
Euryarchaeota	0.06725	p>0.05 se acepta H0	0.764177	p>0.05 se acepta H0
Firmicutes	0.352371	p>0.05 se acepta H0	0.000199	p<0.05 se rechaza H0
Clase				
Actinobacteria	0.078408	p>0.05 se acepta H0	0.000078	p<0.05 se rechaza H0
Alphaproteobacteria	0.30301	p>0.05 se acepta H0	0.136224	p>0.05 se acepta H0
Bacilli	<0.000001	p<0.05 se rechaza H0	0.347218	p>0.05 se acepta H0
Clostridia	0.000026	p<0.05 se rechaza H0	0.001578	p<0.05 se rechaza H0
Gammaproteobacteria	0.46539	p>0.05 se acepta H0	0.992021	p>0.05 se acepta H0
Orden				
Actinomycetales	0.238	p>0.05 se acepta H0	0.000019	p<0.05 se rechaza H0

Sin embargo, a pesar de las ventajas técnicas de usar el perfilador FOCUS (velocidad de ejecución y precisión), hubo diferencias en los niveles taxonómicos que pudieron ser clasificados. De manera que una comparación pareada a todos los niveles (desde Filo hasta género digamos), no fue posible. Observamos que existen categorías que solo eran perfiladas empleando un tipo de marcador (Tabla 15). Destacan con los UBCG el grupo de las Flavobacterias cuya abundancia relativa difiere del modelo nulo. El resto de los niveles representados tienen patrones de abundancia relativa similar, incluso a niveles más finos como el de familia.

Tabla 15. Cálculos de la distribución binomial (valor p) de cada nivel taxonómico utilizando una calculadora en línea http://www.vassarstats.net/binomialX.html?. Considerando exclusivamente a los UBCGs (ya que con los ARNs ribosomales no tuvimos asignación).

Filo	**Distribución binomial UBCG**	**Decisión**
Proteobacteria	0.825871	p>0.05 se acepta H0
Clase		
Betaproteobacteria	0.952156	p>0.05 se acepta H0
Flavobacteriia	0.031555	p<0.05 se rechaza H0
Orden		
Burkholderiales	0.992021	p>0.05 se acepta H0
Flavobacteriales	0.031555	p<0.05 se rechaza H0
Familia		
Comamonadaceae	0.357573	p>0.05 se acepta H0

Estos resultados comparativos, nos permiten concluir que el modelo que usa los UBCG, resulta en perfiles de abundancia relativa similares a los perfiles derivados con el modelo nulo del metagenoma completo. El modelo en el que se usaron los marcadores ribosomales presentó dos problemas fundamentales, un mayor desacuerdo en los perfiles de abundancia relativa respecto al modelo nulo para las categorías comparadas, y falta de resolución taxonómica en varios niveles que no permitieron comparar su desempeño de forma fina. En términos aproximados, los marcadores ribosomales concuerdan con el modelo nulo en un 40% de los casos, contra un 60% de los UBCGs. A pesar de que los marcadores ribosomales pueden estar más representados en las bases de datos, estos tienden a sobreestimar la abundancia relativa en estudios poblacionales, debido a que no siempre se presentan en una sola copia por genoma microbiano. Otro elemento que se debe tener en cuenta es que este análisis comparativo solo se realizó empleando un perfilador FOCUS. Este se eligió porque es amigable con el usuario, tiene precisión aceptable para estudios con microbiomas y requiere muy pocos recursos computacionales. Básicamente FOCUS se ejecuta en pocos minutos en una computadora personal con tan solo 4 GB de memoria RAM, superando a cualquier

otra herramienta existente en la literatura (Kraken, MetaPhlan, Kaiju (Menzel et al., 2016b; Segata et al., 2012b; Genivaldo Gueiros Z Silva et al., 2014; Wood and Salzberg, 2014b). Sería necesario, realizar estudios comparativos con otras herramientas para ponderar estos resultados.

4.5 Asignación taxonómica con Kaiju

Como ejercicio adicional, los tres modelos anteriores fueron también perfilados con la herramienta Kaiju (Menzel et al., 2016b). En este caso nos limitamos a registrar el total de objetos genómicos que pudieron ser clasificados, como un medidor de la utilidad de los modelos para clasificación taxonómica (Tabla 16).

Tabla 16. Resultados generales de la asignación taxonómica utilizando Kaiju.

	Número de secuencias clasificadas	Clasificación taxonómica (%)	Secuencias no clasificadas (%)	Bacterias (%)	Virus (%)	Hongos (%)	Protozoarios (%)
Modelo nulo	664,024	68.87%	31.13%	62.25%	0.2497%	0.9507%	0.01898%
ARNs ribosomales	1330	65.86%	34.14%	55.41%	0	3.233%	1.128%
UBCGs	5470	95.9	4.095	92.78	0	0.1463	0.01828

En la Tabla 16, se observa que la mayor cantidad de secuencias clasificadas corresponden al modelo de los UBCG, con ~96% de las secuencias asignadas a algún taxón; en evidente contraste con los ARNr e incluso el modelo nulo para los que solo se clasificó cerca de un 70% de las secuencias. Sorprendentemente, el grupo más representado fue el de las bacterias. Otro elemento que se resalta es la asignación de secuencias eucarióticas, algo que es biológicamente posible dada la naturaleza de la muestra y que contrasta con FOCUS que solo es capaz de detectar organismos procariontes con su base de datos por defecto.

Los perfiles taxonómicos de los tres modelos son mostrados a continuación en forma de Sankey Plots (Figuras 7, 8 y 9).

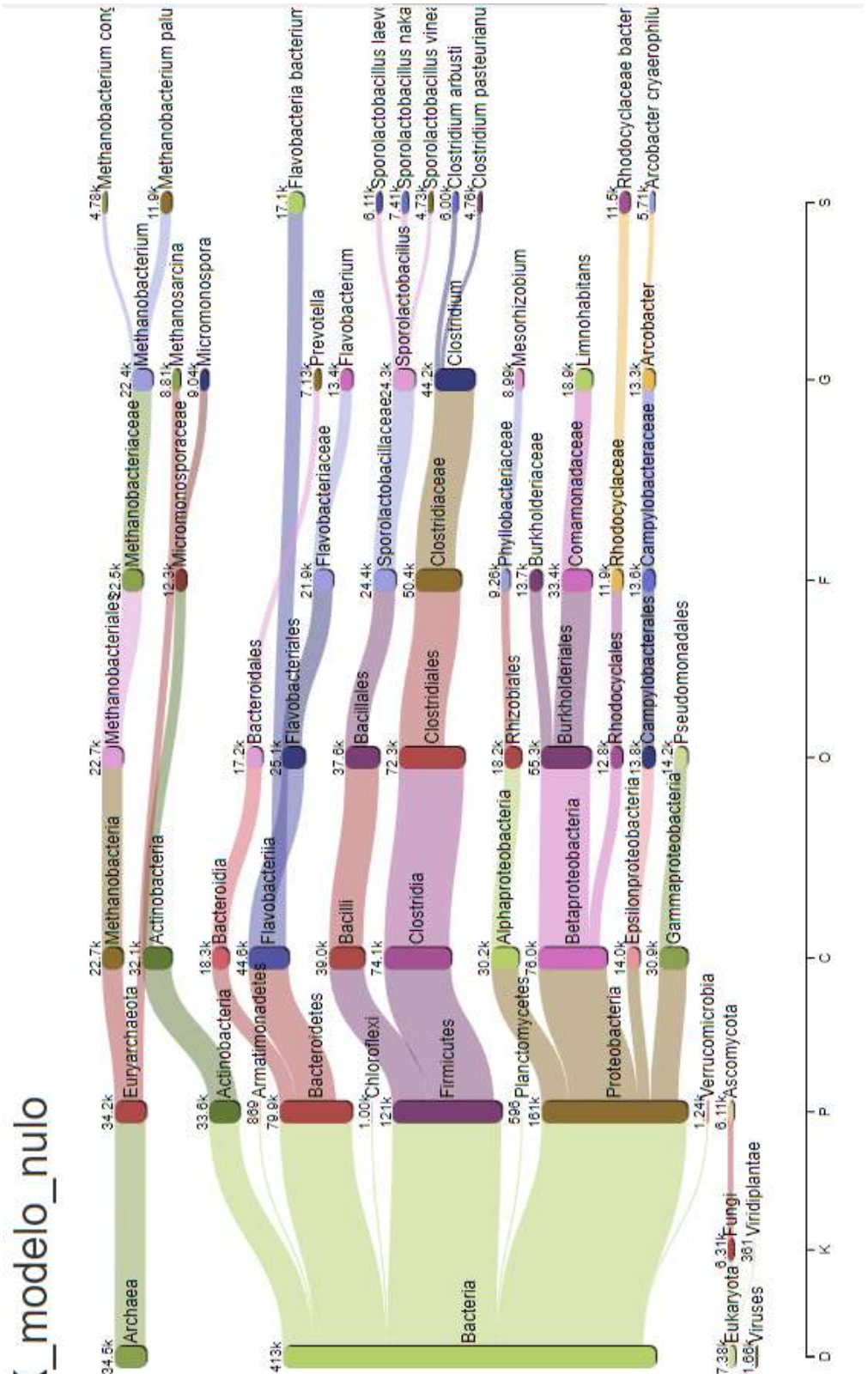


Figura 7. Sankey Plot de la asignación taxonómica utilizando el modelo nulo.

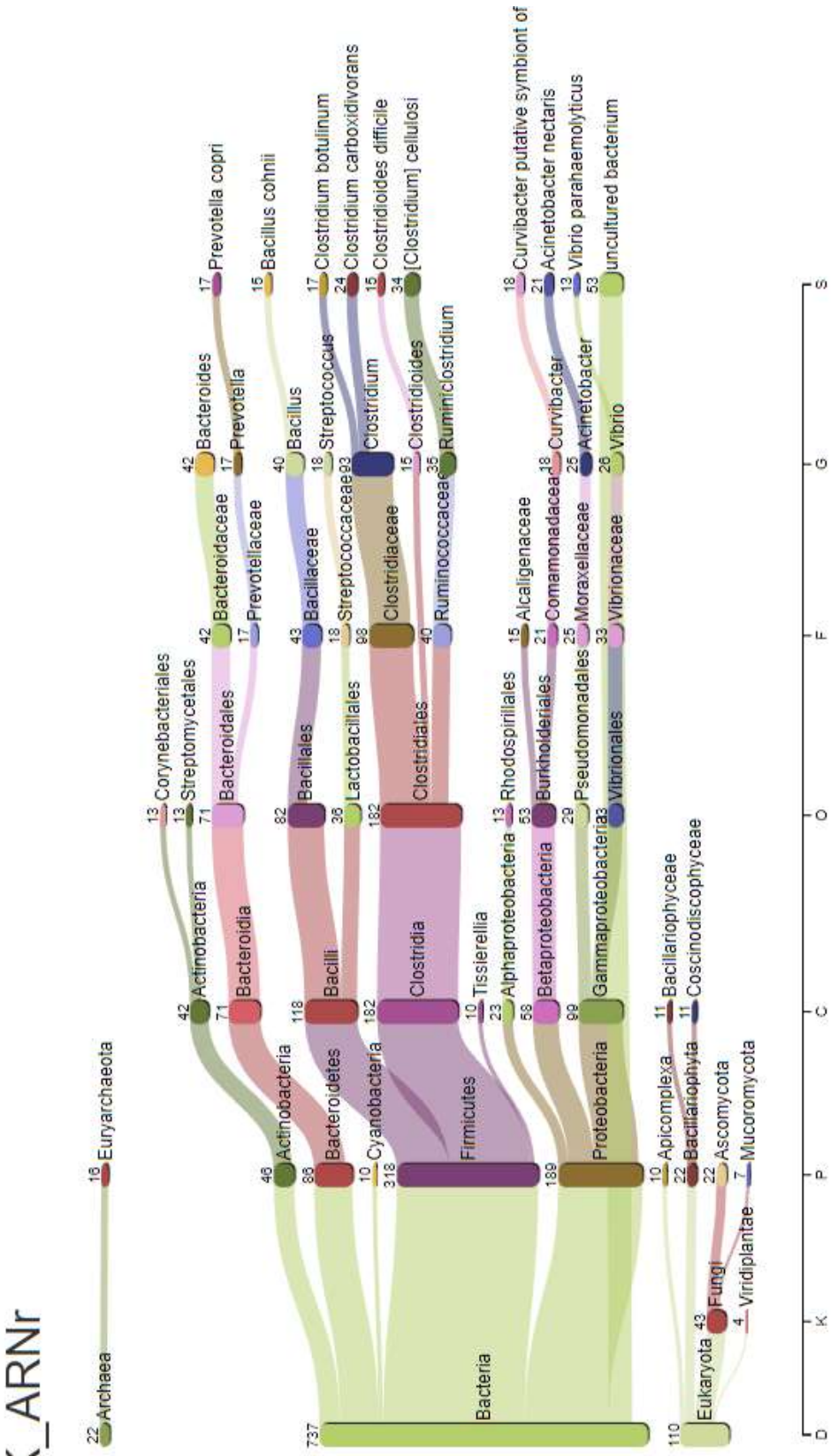


Figura 8. Sankey Plot de la asignación taxonómica utilizando ARNs ribosomales.

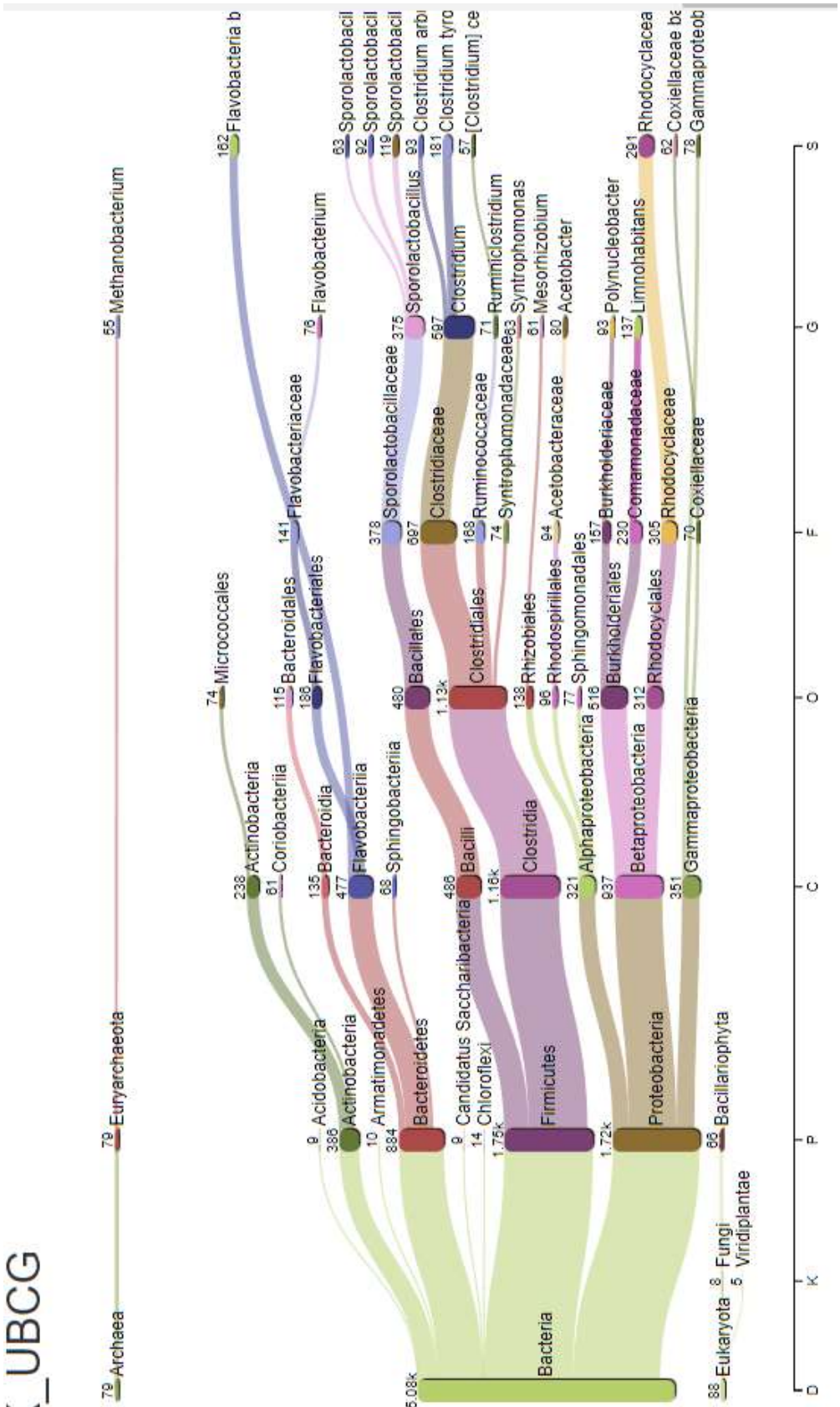


Figura 9. Sankey Plot de la asignación taxonómica utilizando marcadores de copia única.

4.6 Géneros y/o especies de genomas individuales extraídos del metagenoma de sedimentos

Como parte de este proyecto, también se llevó a cabo la identificación taxonómica de un grupo de genomas individuales resueltos por deconvolución metagenómica. Esto fue realizado por la empresa PhaseGenomics (https://phasegenomics.com/) mediante una plataforma privada denominada Proximeta (Press et al., 2017).

Resultado de la deconvolución se obtuvieron 34 compósitos de genomas, los cuales fueron analizados mediante las herramientas Mash y ANI para darle contexto taxonómico (Tabla 17). Al menos 16 genomas pudieron ser asociados a su género o especie empleando los estimadores de relación genómica, los restantes 18 solo pudieron ser asociados a niveles superiores, en la mayoría de los casos a nivel de filo. Estos últimos, sin embargo, es posible que representen nuevos contextos especie-específicos, y por tanto no cuentan con una referencia comparable en la base de datos. Es importante notar que la base de datos del GTDB contiene más de 45,000 genomas de los cuales 18,412 pertenecen al material tipo (poseen taxonomía definida hasta especie) (Federhen, 2015). Entre los genomas resueltos destacan las especies: *Afipia apatlaconensis* (Oren y Garrity, 2022; Sánchez-Reyes et al., 2020), *Clostridium arbusti*, *Clostridium tyrobutyricum* y *Cupriavidus metalidurans*. Varias de estas especies han sido relacionadas con degradación de xenobióticos ambientales y con procesos infectocontagiosos como es el caso del género *Afipia* (Thomas et al., 2006).

Estas estimaciones también nos permitieron inferir la taxonomía de otros bins pertenecientes a los géneros *Sporolactobacillus*, *Acetobacter*, *Methanobacterium*, *Mesorhizobium*, y *Methanosarcina.*

Tabla 17. Delimitación de género y/o especie de cada bin. Se señala con un * los bin que no corresponden al genoma de referencia señalado, el resto se asocian a un género o especie determinada.

Bin	Mash distancia	ANI (%)	FA	Referencia	Asignación taxonómica	Comentario
1*	0.24	60.64	0.42	GCF_002093665.1	*Rouxiella badensis* (enterobacterias)	No pertenece al género ni a la especie sugerido.
2*	0.26	62.72	0.95	GCA_000210715.1	*Fretibacterium fastidiosum* (bacteria)	No pertenece al género ni a la especie sugerido.
3	0.06	94.48	56.64	GCF_000314675.2	*Afipia apatlaconensis* (proteobacteriasa)	Pertenece al género *Afipia*.
4	0.15	82.18	42.09	GCF_000970305.1	*Methanosarcina barkeri 3* (euryarchaeotes)	Pertenece al género *Methanosarcina.*
5	0.02	98.86	60.49	GCF_000246895.1	*Clostridium arbusti* (firmicutes)	Corresponde a la especie *Clostridium arbusti.*
6	0.13	85.08	48.01	GCF_000359585.1	*Clostridium tyrobutyricum* (firmicutes)	Pertenece al género *Clostridium.*
7	0.04	97.51	68,45	GCF_000196015.1	*Cupriavidus metalidurans* (b-proteobacterias)	Corresponde a la especie *Cupriavidus metalidurans.*
8	0.06	92.63	68,77	GCF_006539345.1	*Acetobacter peroxydans* (a-proteobacteria)	Pertenece al género *Acetobacter.*
9*	0.24	63.19	2,33	GCF_900103455.1	*Streptomyces wuyuanensis* (alto GC Gram +)	No pertenece al género ni a la especie sugerido.
10	0.07	90.71	44.08	GCF_001705425.1	*Mesorhizobium hungaricum* (a-proteobacteria)	Pertenece al género *Mesorhizobium.*
11*	0.26	62.00	2.20	GCF_900111815.1	*Bacillus mediterraneensis* (firmicutes)	No pertenece al género ni a la especie sugerido.
12*	0.24	71.35	24.69	GCA_002306975.1	*Bacteria Ruminococcaceae*	No pertenece al género ni a la

					(firmicutes)	especie sugerido.
13	0.03	98.17	48.41	GCF_000359585.1	*Clostridium tyrobutyricum* (firmicutes)	Corresponde a la especie *Clostridium tyrobutyricum.*
14*	0.21	75.64	33.21	GCA_003133285.1	*Bacteria Coriobacteriia* (actinobacteria)	No pertenece al género ni a la especie sugerido.
15*	0.23	70.80	29.58	GCA_002328745.1	*Bacteria actinobacteria* (actinobacteria)	No pertenece al género ni a la especie sugerido.
16	0.20	79.08	46.77	GCF_000377985.1	*Sporolactobacillus vineae* (firmicutes)	Pertenece al género *Sporolactobacillus .*
17*	0.26	62.28	3.28	GCF_900128955.1	*Tissierella praeacuta* (firmicutes)	No pertenece al género ni a la especie sugerido.
18	0.08	91.00	62.3	GCF_900095295.1	*Methanobacterium congolense* (euryarchaeotes)	Pertenece al género *Methanobacterium .*
19*	0.26	62.55	1.44	GCA_002423365.1	*Bacteria desulfuromonadaceae* (d-proteobacteria)	No pertenece al género ni a la especie sugerido.
20*	0.26	64.88	8.13	GCA_900540095.1	*Collinsella sp.* (actinobacterias)	No pertenece al género ni a la especie sugerido.
21*	0.21	73.49	27.75	GCA_900319635.1	Bacteria *Clostridiales* no cultivada (firmicutes)	No pertenece al género ni a la especie sugerido.
22	0.20	83.20	49.64	GCF_900113325.1	*Sporolactobacillus nakayamae* (firmicutes)	Pertenece al género *Sporolactobacillus .*
23	0.18	82.43	39.77	GCF_900095295.1	*Methanobacterium congolense* (euryarchaeotes)	Pertenece al género *Methanobacterium .*
24	0.18	82.79	40.42	GCF_900095295.1	*Methanobacterium congolense* (euryarchaeotes)	Pertenece al género *Methanobacterium .*

25*	0.24	72.32	13.52	GCF_900184925.1	*Bacteria Ruminococcaceae* (firmicutes)	No pertenece al género ni a la especie sugerido.
26*	0.20	93.20	12.8	GCF_000246895.1	*Clostridium arbusti* (firmicutes)	No pertenece al género ni a la especie sugerido.
27	0.17	87.52	23.02	GCF_000314675.2	*Afipia broomeae* (proteobacterias)	Pertenece al género *Afipia*
28	0.19	82.24	53.85	GCF_900095295.1	*Methanobacterium congolense* (euryarchaeotes)	Pertenece al género *Methanobacterium*
29*	0.29	61.26	53.85	GCA_000432195.1	*Bacteria Firmicutes* (firmicutes)	No pertenece al género ni a la especie sugerido.
30*	0.26	74.45	37.18	GCA_002397345.1	Bacteria *Clostridiales* (firmicutes)	No pertenece al género ni a la especie sugerido.
31*	0.29	57.33	0.05	GCA_000238975.1	*Serratia symbiotica str. 'Cinara cedri'* (enterobacterias)	No pertenece al género ni a la especie sugerido.
32*	0.21	83.75	20.46	GCF_000359585.1	*Clostridium tyrobutyricum* (firmicutes)	No pertenece al género ni a la especie sugerido.
33*	0.29	64.86	0.75	GCA_000961805.1	*Bacteria Peptococcaceae* (firmicutes)	No pertenece al género ni a la especie sugerido.
34*	0.29	0	0	GCA_000434555.1	*Clostridium sp* (firmicutes)	No pertenece al género ni a la especie sugerido.

4.7 Anotación funcional de los genomas y posibles vías metabólicas relacionadas con degradación de xenobióticos

Para establecer roles funcionales en los compósitos de genoma estudiados, se anotaron doce de los genomas más completos en la deconvolución, mediante el servidor de KEGG GhostKoala (Kanehisa et al., 2016). Los porcentajes de anotación en los genomas más completos oscilan entre 33.86 % y 61.59 %; estas cifras están dentro de los rangos esperados de anotación en compósitos de

genomas nuevos y de ambientes poco explorados como los sedimentos del río Apatlaco (Tabla 18). De igual manera, la cantidad total de genes asignados es consistente con los conteos de genes en genomas procarióticos discretos (entre 1000 y 5000 genes). Una fracción de los genomas (cerca de un 40 %) no pudo ser asociada con ninguna función molecular conocida. Esto es debido a dos factores principales, la novedad de las secuencias evaluadas y las limitaciones en los algoritmos de exploración para establecer relaciones de homología.

Tabla 18. Resumen de resultados de anotación funcional para los genomas más completos del metagenoma.

Genoma	Cantidad de Genes	Genes con Función	**Porcentaje de Anotación (%)**
bin 1	2539	1376	54.07
bin 2	4091	1826	44.63
bin 3	6770	3127	46.25
bin 4	2268	1134	50
bin 6	2205	1334	60.49
bin 7	2543	1117	43.92
bin 8	2404	1142	47.5
bin 9	1419	874	61.59
bin 10	2571	1381	53.71
bin 11	3486	1378	39.52
bin 12	3476	1177	33.86
bin 13	1071	603	56.3

Debido a que el metabolismo de xenobióticos es un nicho ecológicamente importante en el Apatlaco, decidimos explorar el potencial funcional enriquecido respecto a esta respuesta fenotípica. Para ello, seleccionamos las siguientes vías metabólicas: degradación de benzoato, degradación de cloroalcano y cloroalqueno, degradación de nitrotolueno, degradación de naftaleno, degradación de esteroides, metabolismmo de xenobióticos por citocromo p450, metabolísmo de fármacos citocromo p450 y metabolísmo de fármacos: otras enzimas (Tabla 19).

Tabla 19. Vías metabólicas más representativas en los genomas estudiados

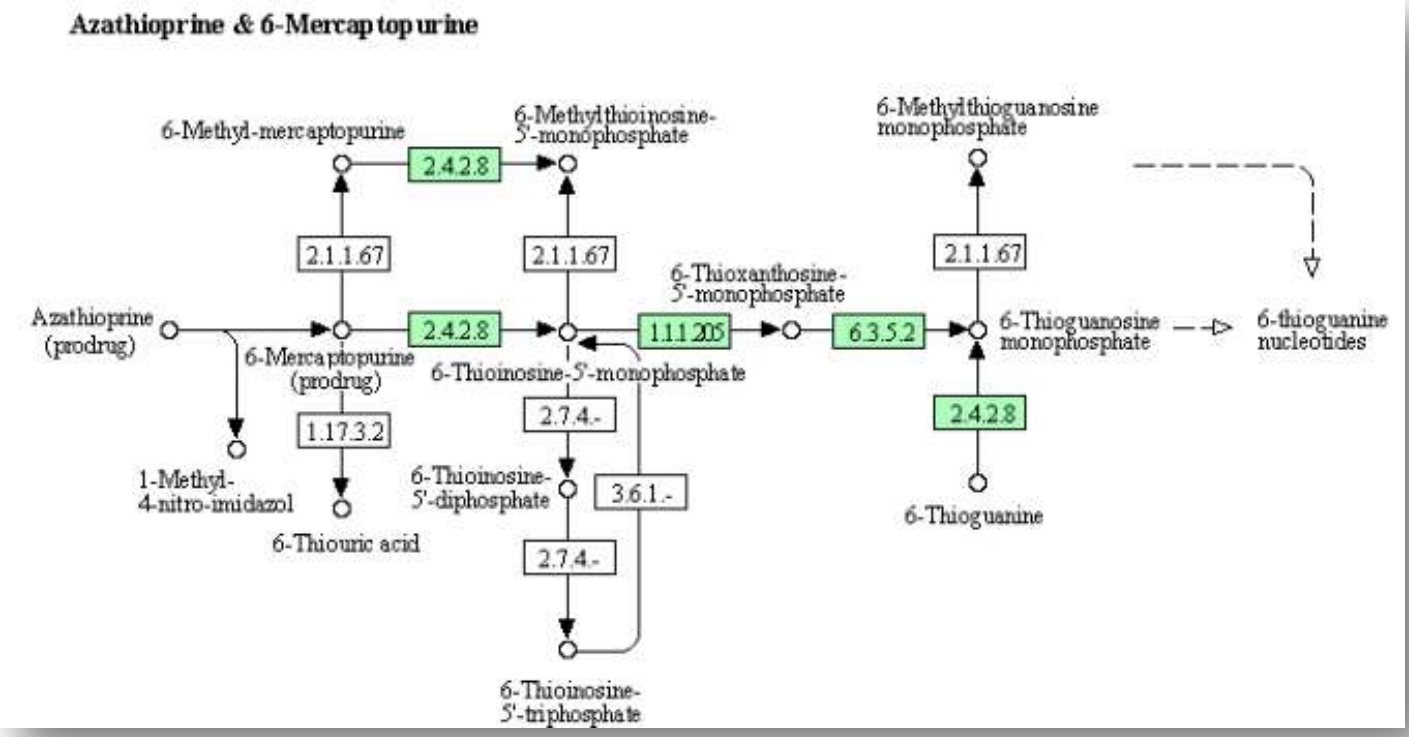
Vía Metabólica	Función Molecular	Definición	Número de Genes
Degradación de Benzoato	K00626	Acetil-CoA C-acetiltransferasa	18
	K01607	Carboximunolactona descarboxilasa	12
	K02588	Proteína de hierro nitrogenasa NifH	19
Degradación de Cloroalcano y Cloroalqueno	K02591	Cadena beta de la proteína nitrogenasa molibdeno-hierro	13
	K02586	Cadena alfa de la proteína nitrogenasa molibdeno-hierro	12
	K13954	Alcohol deshidrogenasa	12
	K04072	Acetaldehído deshidrogenasa / Alcohol deshidrogenasa	11
Degradación de Nitrotolueno	K06281	Subunidad grande de hidrogenasa	12
Degradación de Naftaleno	K05898	Oxoesteroide deshidrogenasa	12
Degradación de Esteroides	K05898	Oxoesteroide deshidrogenasa	31
Metabolísmo de Xenobióticos Citocromo p450	K00799	Glutatión S-transferasa	17
Metabolísmo de Fármacos Citocromo p450	K00799	Glutatión S-transferasa	19
Metabolismo de Fármacos: Otras Enzimas	K00799	Glutatión S-transferasa	19
	K01951	GMP sintasa (hidrolizado de glutamina)	18
	K01195	Beta-glucuronidasa	14
	K00088	IMP deshidrogenasa	13

La mayoría de estos genes son compartidos entre diferentes taxones, lo que sugiere probables relaciones sintróficas entre los miembros de la comunidad. Estas relaciones sintróficas hipotéticas son mecanismos vitales en la colonización de nuevos nichos microbianos sujetos a presiones de selección cambiantes como es el caso de los cuerpos de agua corriente.

La azatioprina es un medicamento xenobiótico usado para prevenir el rechazo al trasplante de órganos en personas que han recibido trasplante de riñón. Así como en otras enfermedades autoinmunes como la artritis reumatoide grave. La

azatioprina pertenece a una clase de medicamentos llamados inmunosupresores. Funciona al reducir la actividad del sistema inmunológico del cuerpo para que no ataque el órgano trasplantado ni a las articulaciones. Sin embargo, presenta efectos adversos en ocasiones graves, como pueden ser aumentar su riesgo de desarrollar ciertos tipos de cáncer y disminución en el número de glóbulos sanguíneos producidos en la médula ósea. La biodegradación de azatioprina por microorganismos constituye un modelo de estudio importante para entender los mecanismos de interacción celular de este medicamento. De manera interesante, en los genomas estudiados se encuentra una vía casi completa para la degradación de este xenobiótico mediante el metabolismo de nucleótidos (Figura 10).

Figura 10. Vía para la degradación microbianana de Azatioprina en el metagenoma estudiado. Las enzimas en verde están presentes en el metagenoma.

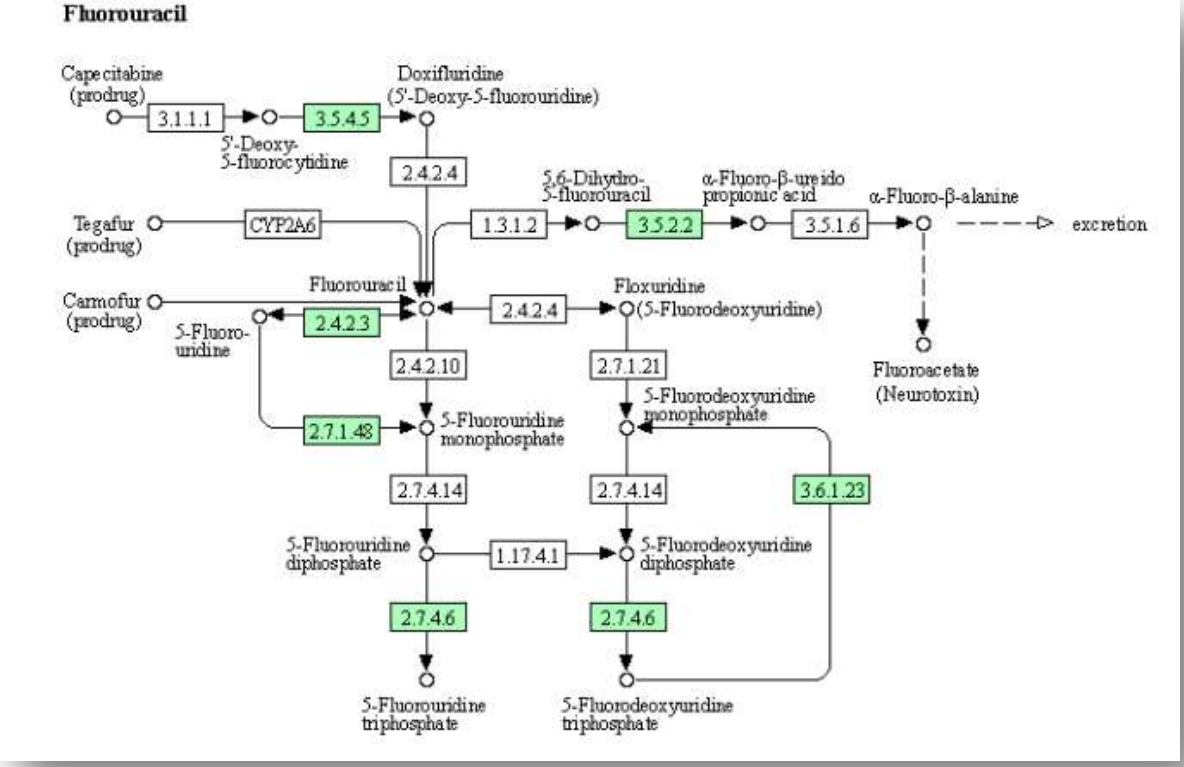


Uno de los principales determinantes enzimáticos encontrados es la hypoxantina fosforibosiltrasferasa [EC:2.4.2.8] la cual cataliza la conversión de hypoxantina a tirosina monofosfato durante la generación de nucleótidos de purina. Es posible

que la activación de la droga a mercaptopurina se lleve a cabo por enzimas oxidativas inespecíficas; se sabe que en humanos estas enzimas pertenecen a la fracción microsomal hepática y en menor cantidad al suero sanguíneo, en bacterias enzimas tipo P450 monoxigenasas se han reportado como posibles candidatas para la activación.

Otro compuesto xenobiótico cuya vía de degradación está representada en los genomas explorados en este estudio es el fluoracilo, un fármaco quimioterapéutico anticanceroso. Ha sido muy utilizado para tratar cáncer de colon y recto, cáncer de mama y cánceres del aparato digestivo entre otros. El fluoracilo puede ser convertido a sus derivados de fluoridina trifosfato mediante la acción concertada de varias enzimas tipo difosfato cinasas, ampliamente representadas en el metagenoma (Figura 11).

Figura 11. Vía de degradación del fluoracilo. Las enzimas en verde están presentes en el metagenoma.



Lo anteriormente expuesto constituye evidencias del potencial genómico metabólico para degradar estos compuestos xenobióticos por los miembros del metagenoma. Dada las características de la zona donde se obtuvieron las muestras, es probable que cierto tipo de selección nutricional o presión de selección esté relacionada con la emergencia de estas capacidades metabólicas, ya que el río Apatlaco constituye un receptáculo de diversas descargas domésticas e industriales.

5. CONCLUSIONES

Podemos concluir que los sedimentos en la cuenca del río Apatlaco sí reflejaron paisajes de microbiodiversidad diferenciales, cuando se usaron dos estimadores filogenéticos distintos (ARNs ribosomales vs UBCGs). Siendo estos últimos más acordes con un modelo de hipótesis nula que representa a un ensamble metagenómico crudo.

Los sedimentos del río Apatlaco constituyen un nicho diverso genómica y microbiológicamente, con abundancia de taxones novedosos que podrían ser útiles en estudios futuros más finos de potencial funcional y fenotípico.

Como parte de este trabajo, se sometió un artículo a publicación en la revista Microbiology Resource Announcements, perteneciente a la Sociedad Americana de Microbiología (ASM). El artículo se tituló: “Draft Genome Sequence of Methanobacterium paludis IBT-C12, Recovered from Sediments of the Apatlaco River, Mexico” (se añade al final de las referencias) y plantea la reconstrucción de un nuevo genoma para la especie metanogénica *Methanobacterium paludis*. Dicho artículo se encuentra aceptado y publicado.

REFERENCIAS

1. *Alberts - Molecular Biology Of The Cell*. (2003).
2. Alneberg, J., Bjarnason, B., de Bruijn, I. et al. Binning metagenomic contigs by coverage and composition. Nat Methods 11, 1144–1146 (2014). https://doi.org/10.1038/nmeth.3103
3. Barco, R. A., Garrity, G. M., Scott, J. J., Amend, J. P., Nealson, K. H., & Emerson, D. (2020). A genus definition for bacteria and archaea based on a standard genome relatedness index. *MBio*, *11*(1). https://doi.org/10.1128/MBIO.02475-19
4. Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2018). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4). https://doi.org/10.1093/bib/bbx120
5. Breton-Deval, L., Sanchez-Flores, A., Juárez, K., & Vera-Estrella, R. (2019). Integrative study of microbial community dynamics and water quality along The Apatlaco River. *Environmental Pollution*, *255*. https://doi.org/10.1016/j.envpol.2019.113158
6. Breton-Deval, L., Sanchez-Reyes, A., Sanchez-Flores, A., Juárez, K., Salinas-Peralta, I., & Mussali-Galante, P. (2020). Functional analysis of a polluted river microbiome reveals a metabolic potential for bioremediation. *Microorganisms*, *8*(4). https://doi.org/10.3390/microorganisms8040554
7. Breton-Deval, L., Sanchez-Reyes, A., Sanchez-Flores, A., Juárez, K., Salinas-Peralta, I., & Mussali-Galante, P. (2020). Functional analysis of a polluted river microbiome reveals a metabolic potential for bioremediation. *Microorganisms*, *8*(4). https://doi.org/10.3390/microorganisms8040554
8. Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, *73*(1). https://doi.org/10.1128/AEM.01177-06
9. Chun J, Rainey FA. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int J Syst Evol Microbiol. 2014 Feb;64(Pt 2):316-324. doi: 10.1099/ijs.0.054171-0. PMID: 24505069.
10. Cisterna, R. (2007). Microbiología. Más dermatología. https://masdermatologia.com/PDF/0006.pdf
11. Comisión Nacional del Agua & Secretaría de Medio Ambiente y Recursos Naturales. (2008). La cuenca del río Apatlaco recuperemos el patrimonio ambiental de los morelenses. http://centro.paot.org.mx/documentos/semarnat/cuenca_rio_apatlaco.pdf
12. Comisión Nacional del Agua, Secretaría de Medio Ambiente y Recursos Naturales, & Organismo de Cuenca Balsas. (2012). El saneamiento del río Apatlaco. De lo crítico a lo sustentable. México. https://www.gob.mx/cms/uploads/attachment/file/121857/El_saneamiento_del_r_o_Apatlaco._De_lo_cr_tico_a_lo_sustentable.pdf
13. Comisión Nacional del Agua. (2021). Calidad del agua en México. https://www.gob.mx/conagua/articulos/calidad-del-agua

14. Cortazar-Martínez, A., González-Ramírez, C., Coronel-Olivares, C., Escalante-Lozada, J., Castro-Rosas, J., & Villagómez-Ibarra, J. (2012). Biotecnología aplicada a la degradación de colorantes de la industria textil. *Universidad y Ciencia*, *28*(2). https://doi.org/10.19136/era.a28n2.27
15. Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, Tak-Wah Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph, *Bioinformatics*, Volume 31, Issue 10, 15 May 2015, Pages 1674–1676, https://doi.org/10.1093/bioinformatics/btv033
16. Editorweb. (2018, 27 enero). Muestran vecinos dónde corre agua contaminada. Diario de Morelos. https://www.diariodemorelos.com/noticias/muestran-vecinos-d-nde-corre-agua-contaminada
17. Eren, AM, Kiefl, E., Shaiber, A. et al. Multi-ómicas integradas, reproducibles y dirigidas por la comunidad con anvi'o. Nat Microbiol 6, 3–6 (2021). https://doi.org/10.1038/s41564-020-00834-3
18. Federhen S. Type material in the NCBI Taxonomy Database. Nucleic Acids Res. 2015 Jan;43(Database issue):D1086-98. doi: 10.1093/nar/gku1127. Epub 2014 Nov 14. PMID: 25398905; PMCID: PMC4383940.
19. INEGI. (2020). *INEGI. Censo de Población y Vivienda 2020*. INEGI. Censo de Población y Vivienda 2020.
20. Instituto Mexicano de Tecnología del Agua. (2007). Plan estratégico para la recuperación ambiental de la cuenca del río Apatlaco. México. https://agua.org.mx/wp-content/uploads/2014/05/plan-cuenca-del-rio-apatlaco.pdf
21. Instituto Mexicano de Tecnología del Agua. (2018). Impacto del cambio climático para la gestión integral de la cuenca hidrológica del río Apatlaco. México. http://www.imta.gob.mx/biblioteca/libros_html/rio_apatlaco/cambio_climatico_rio_aplatlaco.pdf
22. Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. In *Journal of Clinical Microbiology* (Vol. 45, Issue 9). https://doi.org/10.1128/JCM.01228-07
23. Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. In *Nucleic Acids Research* (Vol. 28, Issue 1). https://doi.org/10.1093/nar/28.1.27
24. Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ, 7, e7359. https://doi.org/10.7717/peerj.7359
25. Kanyó, I., & Molnár, L. V. (2016). Procaryotic species and subspecies delineation using average nucleotide identity and gene order conservation. *Gene Reports*, *5*. https://doi.org/10.1016/j.genrep.2016.09.004
26. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 2005;102(7):2567-2572. doi:10.1073/pnas.0409727102

27. Lee, I., Kim, Y. O., Park, S. C., & Chun, J. (2016). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology*, *66*(2). https://doi.org/10.1099/ijsem.0.000760
28. Leopardi, C., & Duno, R. (2010). LA ESPECIE, SU CONCEPTO Y LA MÁS RECIENTE DE LAS PROPUESTAS. *Desde El Herbario CICY*, *2*.
29. Li, H. hong, Wang, Y. tao, Wang, Y., Wang, H. xia, Sun, K. kai, & Lu, Z. mei. (2019). Bacterial degradation of anthraquinone dyes. In *Journal of Zhejiang University: Science B* (Vol. 20, Issue 6). https://doi.org/10.1631/jzus.B1900165
30. Marco Pagni, C. Victor Jongeneel, Making sense of score statistics for sequence alignments, *Briefings in Bioinformatics*, Volume 2, Issue 1, March 2001, Pages 51–67, https://doi.org/10.1093/bib/2.1.51
31. Mende, D. R., Letunic, I., Maistrenko, O. M., Schmidt, T. S. B., Milanese, A., Paoli, L., Hernández-Plaza, A., Orakov, A. N., Forslund, S. K., Sunagawa, S., Zeller, G., Huerta-Cepas, J., Coelho, L. P., & Bork, P. (2020). ProGenomes2: An improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Research*, *48*(D1). https://doi.org/10.1093/nar/gkz1002
32. Mende, D. R., Sunagawa, S., Zeller, G., & Bork, P. (2013). Accurate and universal delineation of prokaryotic species. *Nature Methods*, *10*(9). https://doi.org/10.1038/nmeth.2575
33. Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms11257
34. Na, S. I., Kim, Y. O., Yoon, S. H., Ha, S. min, Baek, I., & Chun, J. (2018). UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *Journal of Microbiology*, *56*(4). https://doi.org/10.1007/s12275-018-8014-6
35. Ochoa-Sánchez, L. E., & Vinuesa, P. (2017). Evolutionary genetic analysis uncovers multiple species with distinct habitat preferences and antibiotic resistance phenotypes in the Stenotrophomonas maltophilia complex. *Frontiers in Microbiology*, *8*(AUG). https://doi.org/10.3389/fmicb.2017.01548
36. Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0997-x
37. Oren A, Garrity G. Notification of changes in taxonomic opinion previously published outside the IJSEM. List of changes in taxonomic opinion no. 35. Int J Syst Evol Microbiol. 2022 Feb;72(1). doi: 10.1099/ijsem.0.005164. PMID: 35113782.
38. Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, Donovan H Parks, GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database, Bioinformatics, Volume 36, Issue 6, 15 March 2020, Pages 1925–1927, https://doi.org/10.1093/bioinformatics/btz848
39. Press MO, Wiser AH, Kronenberg ZN, Langford KW, Shakya M, Lo C-C, Mueller KA, Sullivan ST, Chain PSG, Liachko I. Hi-C deconvolution of a

human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. 2017. https://doi.org/10.1101/198713. https://www.biorxiv.org/content/early/2017/10/05/198713

40. Qin, Q. L., Xie, B. bin, Zhang, X. Y., Chen, X. L., Zhou, B. C., Zhou, J., Oren, A., & Zhang, Y. Z. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of Bacteriology*, *196*(12). https://doi.org/10.1128/JB.01688-14
41. Rodriguez-R, L.M., & Konstantinidis, K.T. (2014). Bypassing Cultivation To Identify Bacterial Species: Culture-independent genomic approaches identify credibly distinct clusters, avoid cultivation bias, and provide true insights into microbial species. Microbe Magazine, 9, 111-118.
42. Rosselló-Mora, R., & Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiology Reviews*, *25*(1). https://doi.org/10.1111/j.1574-6976.2001.tb00571.x
43. Sánchez-Reyes, A., & Luis Folch-Mallol, J. (2020). Metagenomics-Based Phylogeny and Phylogenomic. In *Metagenomics - Basics, Methods and Applications*. https://doi.org/10.5772/intechopen.89492
44. Sánchez-Reyes, A., Bretón-Deval, L., Mangelson, H., Salinas-Peralta, I., & Sanchez-Flores, A. (2020). Hi-C deconvolution of a textile-dye degrader microbiome reveals novel taxonomic landscapes and link phenotypic potential to individual genomes. *BioRxiv*. https://doi.org/10.1101/2020.06.18.159848
45. Sánchez-Salazar EA, Hernández-Jaimes L, Breton-Deval L, Sánchez-Reyes A. Draft Genome Sequence of Methanobacterium paludis IBT-C12, Recovered from Sediments of the Apatlaco River, Mexico. Microbiol Resour Announc. 2022 Feb 3:e0090621. doi: 10.1128/mra.00906-21. Epub ahead of print. PMID: 35112899.
46. Sedlar, K., Kupkova, K., & Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. In *Computational and Structural Biotechnology Journal* (Vol. 15). https://doi.org/10.1016/j.csbj.2016.11.005
47. Seemann T (2018), barrnap 0.9: rap. id ribosomal RNA predictionhttps://github.com/tseemann/barrnap/0.9
48. Seemann, T. (2013). barrnap 0.9 : rapid ribosomal RNA prediction. *Github.Com*.
49. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8). https://doi.org/10.1038/nmeth.2066
50. SEMARNAT, CEAGUA, & CONAGUA. (2017). Estadísticas del agua en el estado de Morelos, 2017. México. https://ceagua.morelos.gob.mx/node/108
51. Shanker, K., Vijayakumar, S. P., & Ganeshaiah, K. N. (2017). Unpacking the species conundrum: philosophy, practice and a way forward. In *Journal of Genetics* (Vol. 96, Issue 3). https://doi.org/10.1007/s12041-017-0800-0
52. Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E., & Edwards, R. A. (2014). FOCUS: An alignment-free model to identify organisms in metagenomes

using non-negative least squares. *PeerJ*, *2014*(1). https://doi.org/10.7717/peerj.425

53. Sun, D. L., Jiang, X., Wu, Q. L., & Zhou, N. Y. (2013). Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. *Applied and Environmental Microbiology*, *79*(19). https://doi.org/10.1128/AEM.01282-13
54. Thomas V, Herrera-Rimann K, Blanc DS, Greub G. Biodiversity of amoebae and amoeba-resisting bacteria in a hospital water network. Appl Environ Microbiol. 2006 Apr;72(4):2428-38. doi: 10.1128/AEM.72.4.2428-2438.2006.
55. Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. Nucleic acids research, 43(14), 6761-6771
56. Wayne, L.G. & Brenner, D.J. & Colwell, Rita & Grimont, Patrick & Krichevsky, Micah & Moore, L.H. & Moore, W.E.C. & Murray, R.G.E. & Stackebrandt, Erko & Starr, M.P. & Truper, H.G.. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. International Journal of Systematic Bacteriology. 37. 10.1099/00207713-37-4-463.
57. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010 Jul;38(12):e132. doi: 10.1093/nar/gkq275. Epub 2010 Apr 19. PMID: 20403810; PMCID: PMC2896542.
58. Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3). https://doi.org/10.1186/gb-2014-15-3-r46
59. Wu, Y. W., Tang, Y. H., Tringe, S. G., Simmons, B. A., & Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome, 2, 26. https://doi.org/10.1186/2049-2618-2-26
60. Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. In *Cell* (Vol. 178, Issue 4). https://doi.org/10.1016/j.cell.2019.07.010
61. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, Chun J. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol. 2017 May;67(5):1613-1617. doi: 10.1099/ijsem.0.001755. Epub 2017 May 30. PMID: 28005526; PMCID: PMC5563544.

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Draft Genome Sequence of *Methanobacterium paludis* IBT-C12, Recovered from Sediments of the Apatlaco River, Mexico

Erick Alejandro Sánchez-Salazar,[a] Lizbeth Hernández-Jaimes,[b] Luz Breton-Deval,[c] Ayixon Sánchez-Reyes[c]

[a]Universidad Autónoma Metropolitana, Unidad Iztapalapa, Mexico City, Mexico
[b]Centro de Investigación en Dinámica Celular, Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos, Mexico
[c]Cátedras Conacyt-Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

ABSTRACT *Methanobacterium paludis* is a hydrogenotrophic archaea first described in 2014 and isolated from a peatland area. So far, there is only one sequenced genome of this taxon. Here, we report the draft genome sequence of *M. paludis* IBT-C12, a metagenome-assembled genome (MAG) from sediments in the Apatlaco River, Mexico.

Methanobacterium paludis is an anaerobic archaeon initially described in 2014 and isolated from a peatland area. It uses a hydrogenotrophic pathway to reduce CO_2 to methane on methanogenesis (1). There is little information on the genomic complement of this methanogenic species, with only one available assembly for the type strain, SWAN1, on the National Center for Biotechnology Information (NCBI) (accession number GCA_000214725.1) (2). Other genomic composites would be crucial to gain information on the metabolism, ecology, taxonomy, and other traits of environmental relevance. Here, we report the draft genome sequence of *M. paludis* IBT-C12, a metagenome-assembled genome (MAG) recovered from freshwater sediments in a highly polluted river in Mexico (3, 4). The MAG was isolated from an environmental sample. Full information about the sample nature, collection site, deconvolution processes, and data were reported before (3). Briefly, sediments were collected at 10-cm depth at a rate of 0.5 kg per site and were stored in coolers at 4°C until further processing. Metagenomic DNA was extracted using the DNeasy PowerWater kit (Qiagen, Hilden, Germany) according to the manufacturer's recommendations. The *M. paludis* IBT-C12 genome was deconvolved from two sets of short reads obtained from freshwater and sediment in the Apatlaco River, Mexico. One read set was acquired from freshwater on a shotgun NextSeq 500 sequencing platform (Illumina, Inc., San Diego, CA) (108,785,988 paired-end reads, 75 bp) and a set of Hi-C reads from sediments (72,007,031 paired-end reads, 150 bp) from a HiSeq 4000 Illumina-compatible sequencing library. The shotgun library was prepared using the TruSeq kit version 2 (Illumina, Inc.) according the manufacturer's protocol. The ProxiMeta microbiome kit (Phase Genomics, Seattle, WA) was used to prepare the Hi-C library, and the DNA was digested with Sau3AI and MluCI enzymes before proximity ligation. Read quality was assessed and filtered under the Phase Genomics cloud-based bioinformatics portal by using the Fastp tool version 0.23.1 (5, 6). A metagenome assembly was constructed with the shotgun reads (only the NextSeq data were used) using MEGAHIT software version 1.2.9 (7) (size 550,428,717 bp and 746,031 contigs). The Hi-C reads were subsequently mapped with the Burrows-Wheeler Aligner MEM algorithm (BWA-MEM) version 0.7.17 (-5SP and -t 8 options) to create Hi-C-based contact probability maps (8). The binning was performed with ProxiMeta software (5), resulting in 97 MAGs (3). The MAG corresponding to IBT-C12 was identified by comparing genome relatedness indexes (mash distance [9] and average nucleotide identity [ANI] [10, 11]) against a custom database (12). Default parameters were used for all software unless otherwise noted. The total

Editor Frank J. Stewart, Montana State University

Address correspondence to Ayixon Sánchez-Reyes, ayixon.sanchez@ibt.unam.mx.

The authors declare no conflict of interest.

Received 21 September 2021
Accepted 12 January 2022
Published 3 February 2022

TABLE 1 Average nucleotide identity and Mash mutational distance (D) of *Methanobacterium paludis* IBT-C12 compared with the closest NCBI type strains of the genus *Methanobacterium*

GenBank assembly accession no.	Reference organism	OrthoANI (%)	Mash D
GCA_000214725.1	*Methanobacterium paludis* SWAN1	95.20	0.05
GCA_017874455.1	*Methanobacterium aggregans* DSM 29428	78.02	0.22
GCA_002287175.1	*Methanobacterium bryantii* M.o.H	71.07	0.26
GCA_000745485.1	*Methanobacterium veterum* MK4	70.64	0.26
GCA_000746075.1	*Methanobacterium arcticum* M2	70.60	0.26
GCA_001316325.1	*Methanobacterium formicicum* JCM 10132	69.41	0.30
GCA_017873625.1	*Methanobacterium petrolearium* DSM 22353	74.38	0.30
GCA_002813695.1	*Methanobacterium subterraneum* A8p	69.33	1.00
GCA_000016525.1	*Methanobrevibacter smithii* DSMZ 861	65.91	1.00

length of the *M. paludis* IBT-C12 genome was 2,696,644 bp, with a G+C content of 35.95%. The number of contigs was 374, with an average shotgun sequencing coverage depth of 16× and an N_{50} value of 1,849,042 bp. The completeness and contamination were estimated with miComplete version 1.1.1 (13) at 61.83% and 1.13%, respectively. The genome size was close to that of the reference strain, SWAN1 (2,546,541 bp), so IBT-C12 MAG must be moderately complete.

The IBT-C12 strain is closely related to *M. paludis* SWAN1, as shown by a FastANI result of 94.34% (95.2% with OrthoANI version 0.5.0) and the phylogenomic analysis (Table 1; Fig. 1) (10, 14). The annotation of the MAG sequence was performed with the DFAST service version web with default parameters (15). A total of 2,550 genes were identified, with 2,514 protein-coding sequences and 36 RNA genes. The RNA genes comprised 1 partial 16S rRNA, 2 partial 23S rRNAs, and 33 tRNAs; 2 CRISPR arrays were also identified.

The MAG described in this work could provide valuable information regarding the ecology, metabolism, phylogeny, and evolution of the *M. paludis* clade.

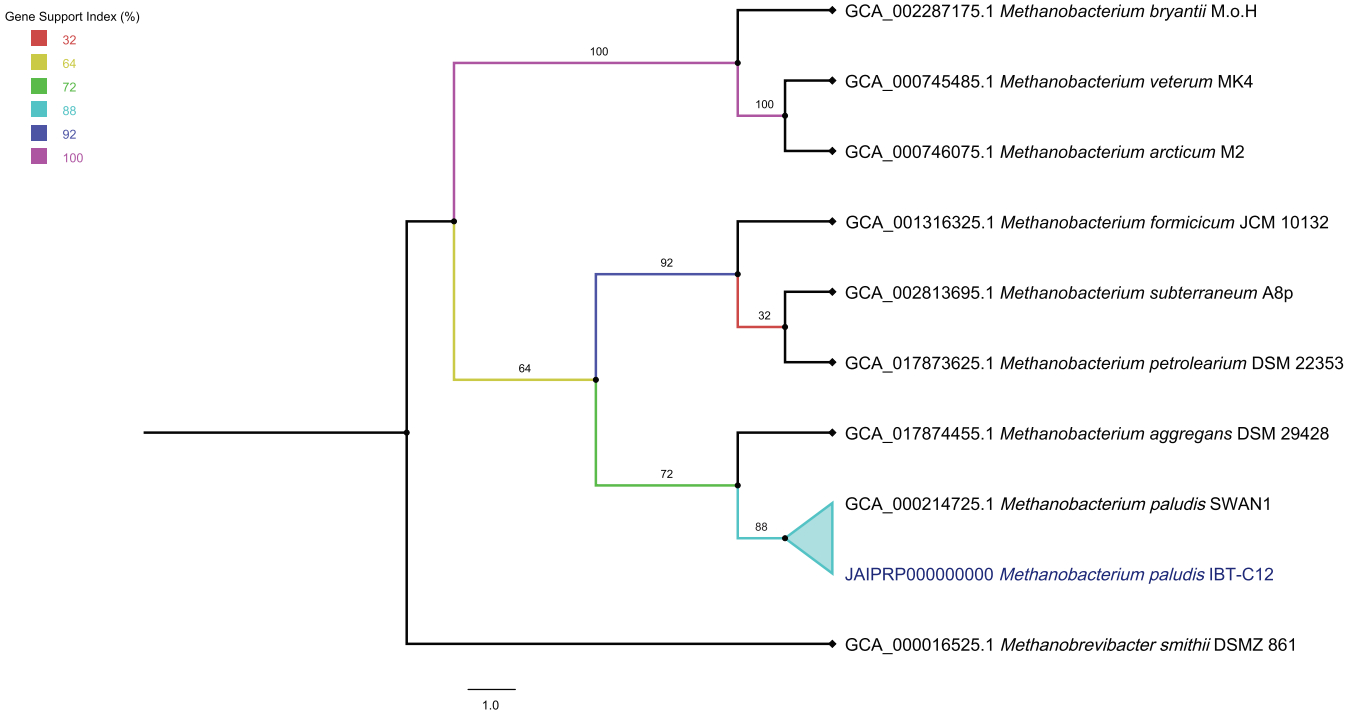


FIG 1 FastTree phylogenomic tree using 25 core genes inferred with the UBCG pipeline (14, 16). The 25 core genes were automatically aligned with MAFFT and concatenated (17). Poorly conserved regions were curated with Gblocks (18). Gene support indices (GSIs) as percentages are shown at branching points. Bars indicate substitution per site.

Data availability. This whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the version number JAIPRP000000000.1. The version described in this paper is version JAIPRP000000000.2. The BioProject accession number is PRJNA759916. The BioSample accession number is SAMN21209357. The draft shotgun metagenome assembly is available on https://figshare.com/ndownloader/files/28075893. The Hi-C reads and NextSeq data are available under accession numbers SRR11481801 and SRX6045636 to SRX6045639, respectively, in the SRA.

ACKNOWLEDGMENTS

This work was funded in part by the Fondo Institucional para el Desarrollo Científico, Tecnológico y de Innovación FORDECYT-PRONACES, under the project code CF 2019 265222. The funders had no role in study design, data processing, or interpretation.

We thank UNAM-IBT and the program "Investigadoras e Investigadores por México" from CONACYT for support of this study.

REFERENCES

1. Cadillo-Quiroz H, Brauer SL, Goodson N, Yavitt JB, Zinder SH. 2014. Methanobacterium paludis sp. nov. and a novel strain of Methanobacterium lacus isolated from northern peatlands. Int J Syst Evol Microbiol 64:1473–1480. https://doi.org/10.1099/ijs.0.059964-0.
2. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, DiCuccio M, Murphy TD, Pruitt KD, Kimchi A. 2016. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res 44:D73–D80. https://doi.org/10.1093/nar/gkv1226.
3. Sánchez-Reyes A, Bretón-Deval L, Mangelson H, Salinas-Peralta I, Sanchez-Flores A. 2021. Hi-C deconvolution of a textile dye–related microbiome reveals novel taxonomic landscapes and links phenotypic potential to individual genomes. Int Microbiol 25:99–110. https://doi.org/10.1007/s10123-021-00189-7.
4. Breton-Deval L, Sanchez-Reyes A, Sanchez-Flores A, Juárez K, Salinas-Peralta I, Mussali-Galante P. 2020. Functional analysis of a polluted river microbiome reveals a metabolic potential for bioremediation. Microorganisms 8:554. https://doi.org/10.3390/microorganisms8040554.
5. Press MO, Wiser AH, Kronenberg ZN, Langford KW, Shakya M, Lo C, Mueller KA, Sullivan ST, Chain PSG, Liachko I. 2017. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. bioRxiv https://doi.org/10.1101/198713.
6. Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
7. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, Yamashita H, Lam TW. 2016. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102:3–11. https://doi.org/10.1016/j.ymeth.2016.02.020.
8. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
9. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:1–14. https://doi.org/10.1186/s13059-016-0997-x.
10. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9.
11. Lee I, Kim YO, Park SC, Chun J. 2016. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol 66:1100–1103. https://doi.org/10.1099/ijsem.0.000760.
12. Sánchez-Reyes A, Fernández-López MG. 2021. Mash sketched reference dataset for genome-based taxonomy and comparative genomics. Preprints https://doi.org/10.20944/PREPRINTS202106.0368.V1.
13. Hugoson E, Lam WT, Guy L. 2019. miComplete: weighted quality evaluation of assembled microbial genomes. Bioinformatics 36:936–937. https://doi.org/10.1093/bioinformatics/btz664.
14. Na SI, Kim YO, Yoon SH, Ha S-M, Baek I, Chun J. 2018. UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. J Microbiol 56:281–285. https://doi.org/10.1007/s12275-018-8014-6.
15. Tanizawa Y, Fujisawa T, Nakamura Y. 2018. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics 34:1037–1039. https://doi.org/10.1093/bioinformatics/btx713.
16. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2: approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
17. Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286–298. https://doi.org/10.1093/bib/bbn013.
18. Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577. https://doi.org/10.1080/10635150701472164.