UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

LA VÍA BENZOIL-COA ES UN MARCADOR GENÉTICO PARA PREDECIR LOS REQUERIMIENTOS DE OXÍGENO PARA LA DEGRADACIÓN DE HIDROCARBUROS MONOAROMÁTICOS.

T E S I S

PARA OBTENER EL TÍTULO DE:

BIÓLOGA

P R E S E N T A:

CAMILA MONSERRAT GODÍNEZ PÉREZ

DIRECTORA DE TESIS:
Dra. ROSA MARÍA GUTIÉRREZ RÍOS
(2023) INSTITUTO DE BIOTECNOLOGÍA

CD. MX.

Este proyecto se realizó bajo la asesoría de la Dra. Rosa María Gutiérrez Ríos, en el laboratorio de Genómica Computacional adscrito al Departamento de Microbiología Molecular del Instituto de Biotecnología de la Universidad Nacional Autónoma de México campos Morelos. Agradezco al proyecto PAPIIT-DGAPA IN202821 por el apoyo económico.

Esta investigación fue financiada por PAPIIT-DGAPA IN202821 y por Ciencia Básica y/o Ciencia de Frontera. Modalidad: Paradigmas y Controversias de la Ciencia 2022, Proyecto 319234, ambos proyectos adjudicados a RMGR.

Índice

Índice de figuras

Índice de tablas.

Índice de tablas del anexo.

1. Agradecimientos

En el camino de la realización de este proyecto, he encontrado apoyo y aliento en muchas personas. Me gustaría expresar mi sincera gratitud a aquellos que han contribuido de manera significativa a este logro.

Agradezco a la Dra. Rosa María Gutiérrez Ríos por su guía, su gran determinación y su disposición constante para responder a mis preguntas y compartirme su conocimiento.

A mi familia, especialmente a mi madre; Sofia Pérez, por su amor incondicional y su apoyo constante a lo largo de los años. Sin duda, esto no sería posible sin la ayuda de mis padres y hermanos, que me han apoyado a lo largo de mi trayectoria académica.

¡Gracias!

2. Abstract

The enzyme Benzoate-CoA ligase (BCL) plays a central role in the anaerobic degradation of aromatic hydrocarbons. These compounds are present in a variety of environmental pollutants such as petroleum derivatives, industrial chemicals, and organic waste, which pose significant concerns for human health and ecological balance. BCL plays a key role in the anaerobic degradation of aromatic hydrocarbons by catalyzing the conversion of benzoate to benzoil-CoA. The anaerobic pathway is driven by a benzoyl-CoA reductase that initiates benzoyl-CoA breakdown, after which is oxidized. In several bacteria, however, the aerobic metabolism of aromatic acids occurs via the degradation box pathway. The pathway has been studied both experimentally, in some Alphaproteobacteria and Betaproteobacteria, and bioinformatically, in representative Betaproteobacteria. However, the distribution of the benzoyl-CoA pathway and the evolutionary forces driving its adaptation beyond representative bacteria have not been reported. In this study, we propose a series of bioinformatic steps for recognizing benzoate-CoA ligase (BCL) in fully sequenced genomes and metagenome-assembled genomes (MAGs) based on the recognition of protein architectures, the preservation of genomic context, and the conservation of ungapped motifs describing the catalytic properties of BCLs, two of which were exclusive to these enzymes. The established rules have proven to be highly effective in accurately distinguishing BCL from other aryl-CoA ligases involved in aerobic, anaerobic, or hybrid pathways commonly found in Betaproteobacteria. During this research, paralogs of BCL were identified that followed all the rules in genera belonging to the families Rhodocyclaceae, Zoogloeaceae, and Burkholderiaceae in fully sequenced genomes, where some Burkholderiaceae have multipartite genomes. As a result, a phylogenetic tree was constructed. The phylogenetic analysis of BCLs classified the paralogs into distinct clades, indicating that these BCL copies were acquired in separate evolutionary events. This distribution is replicated in multipartite genomes where the phylogenetic tree placed the BCLs from the main chromosome in an older clade and those encoded in a second chromosome in a second clade. The phylogenetic distribution of BCL in fully sequenced genomes is primarily observed in the Betaproteobacteria class, while in MAGs, it predominantly occurs in the Alphaproteobacteria class. Similarly, paralogs of BCL were identified in the MAGs. Furthermore, it was observed that high-quality MAGs retain a more complete genomic context due to their better assembly, which increases the accuracy of predictions. An analysis of the isolation sites of fully sequenced genomes exhibiting the BCL revealed that these organisms inhabit a variety of environments, predominantly terrestrial, thus allowing the exploration of the marine ecosystem through MAG analysis. The presence and redundancy of the BCL and benzoil-CoA degradation pathway genes in both sequenced genomes and MAGs demonstrate the versatility of these genomes to adapt to various environmental conditions, from terrestrial

ecosystems, ecosystems contaminated with HA, human hosts, animals, rhizosphere, to marine ecosystems. Therefore, the developed methods, along with the obtained results, provide information for future research related to degrader organisms.

3. Resumen

La enzima Benzoato-CoA ligasa (BCL) tiene un papel central en la degradación anaeróbica de los hidrocarburos aromáticos. Estos compuestos, están presentes en una variedad de contaminantes ambientales como los derivados del petróleo, los productos químicos industriales y los desechos orgánicos, que representan una preocupación significativa para la salud humana y el equilibrio ecológico. La BCL desempeña un papel fundamental en la degradación anaeróbica de los hidrocarburos aromáticos al catalizar la conversión del benzoato en benzoil-CoA. La vía subsecuente implica la degradación del benzoil-CoA y se encuentra en bacterias que habitan en ambientes tanto anaeróbicos como aeróbicos, lo que da lugar a diversas rutas de degradación. La vía anaeróbica incluye a la enzima benzoil-CoA reductasa encargada de la reducción del benzoil-CoA. Por otro lado, en la vía aeróbica, se encuentra la conocida vía "Box", que ha sido objeto de investigación tanto experimentalmente en ciertas Alphaproteobacterias y Betaproteobacterias, como a través de análisis bioinformáticos en Betaproteobacteria representativas. Sin embargo, la distribución de la vía del benzoil-CoA y los factores evolutivos que impulsan su adaptación más allá de las bacterias representativas aún no se han estudiado. Para abordar esta cuestión, hemos desarrollado un conjunto de pasos bioinformáticos que permiten la identificación de la benzoato-CoA ligasa en genomas completamente secuenciados y en genomas ensamblados de metagenomas (MAGs). Estos pasos se basan en el análisis de arquitecturas proteicas, la conservación del contexto genómico y la detección de motivos que describen las propiedades catalíticas de las BCLs. Las reglas establecidas han demostrado ser altamente efectivas al diferenciar con precisión las BCL de otras aril-CoA ligasas que participan en las vías aeróbicas, anaeróbicas o híbridas comúnmente presentes en las Betaproteobacterias. Durante esta investigación, se identificaron parálogos de la BCL que siguieron todas las reglas en los géneros pertenecientes a las familias Rhodocyclaceae, Zoogloeaceae y Burkholderiaceae en genomas completamente secuenciados, en donde algunas Burkholderiaceae presentan un genoma segmentado. En consecuencia, construyó un árbol filogenético de las BCLs que clasificó los parálogos en clados distintos, lo que indica que estas copias de la BCL se adquirieron en eventos evolutivos separados. Esta distribución se replica en los genomas segmentados en donde el árbol filogenético ubicó en un clado más antiguo a las BCLs del cromosoma principal y a los codificados en un crómido putativo en un segundo clado. La distribución filogenética de la BCL en los genomas completamente secuenciados se observa principalmente en la clase Betaproteobacteria, mientras que en los MAGs predomina en la clase Alphaproteobacteria. Igualmente, se identificaron parálogos de la BCL en los MAGS. Además, se observó que los MAGs de alta calidad conservan un contexto genómico más completo, dado su mejor ensamblaje, lo que aumenta la precisión de las predicciones. Un análisis de los sitios de aislamiento de los genomas totalmente secuenciados que exhiben la BCL reveló que estos organismos habitan una variedad de entornos, predominantemente

terrestres, por lo que el análisis de los MAGs permitió explorar el ecosistema marino. La presencia y redundancia de la BCL y los genes de la vía de degradación del benzoil-CoA en los genomas secuenciados y en los MAGs, demuestra la versatilidad de estos genomas para adaptarse a diversas condiciones ambientales, desde ecosistemas terrestres, ecosistemas contaminados con HA, huéspedes humanos, animales, rizosfera y ecosistemas marinos. Por lo que los métodos desarrollados, junto con los resultados obtenidos, proporcionan información para futuras investigaciones relacionadas con organismos degradadores de HA, así como para la predicción de ortólogos de diferentes proteínas.

4. Introducción

4.1 Problemática de los océanos contaminados por derrames de petróleo

La demanda mundial de energía y productos derivados del petróleo, tales como los combustibles, plásticos y una amplia gama de productos químicos, ha aumentado en las últimas décadas debido al crecimiento demográfico y al desarrollo económico (Scoma *et al.*, 2017). La economía mundial consume alrededor de 30 mil millones de barriles de petróleo cada año. Sin embargo, este consumo se considera insostenible ya que aumenta los riesgos de desastres ambientales debido a las actividades de exploración y transporte marítimo. En los últimos 50 años, se han vertido más de 5.8 millones de toneladas de petróleo en los océanos por diferentes causas, como colisiones entre transportes petroleros, fallas de los equipos, incendios y explosiones en plataformas petroleras, eventos climáticos, entre otros (de Melo *et al.*, 2022).

Algunos ejemplos de los desastres ecológicos asociados al petróleo son el caso Exxon Valdez en 1989, Prestige en 2002 y, por último, el desastre de Deepwater Horizon en 2010, el cual se considera como uno de los mayores derrames de petróleo reportados en la historia. Este último tuvo lugar en el Golfo de México debido a la explosión de la plataforma Deepwater Horizon, derramando 700.000 toneladas de petróleo en el Océano Atlántico durante 87 días (Scoma *et al.*, 2017).

Los derrames de petróleo en el medio marino han sido una importante amenaza para el ecosistema, incluida la vida oceánica y los seres humanos, (Parab & Phadke 2020). Los hidrocarburos aromáticos (HA) componen entre el 0.2 y el 7% del petróleo y se consideran una clase importante de marcadores de contaminación. La constante exposición de los seres humanos a los HA a través de los alimentos y en el medio ambiente, tiene efectos genotóxicos, mutagénicos y carcinógenos. Además, se consideran potentes inmunosupresores lo que representa un problema de salud pública (de Melo *et al.*, 2022).

Por su naturaleza recalcitrante los HA son un grave problema ambiental, ya que pueden permanecer décadas en el ecosistema sin poder ser degradados. Los métodos físicos y químicos como la volatilización, la foto oxidación, la oxidación química y la bioacumulación rara vez tienen éxito en la eliminación rápida de los HA en el ecosistema. Además, se considera que estos métodos no son seguros y rentables en comparación con la biorremediación microbiana. Las bacterias han sido consideradas durante mucho tiempo

como uno de los agentes degradadores de hidrocarburos predominantes en el medio ambiente (Dasgupta *et al.,* 2013). La recuperación de los océanos tras los derrames petroleros evidencia la presencia de microorganismos especializados, los cuales degradan los HA eficientemente al estar fisiológicamente adaptados a utilizar como fuente de energía moléculas orgánicas complejas (Meckenstock & Mouttaki 2011, Scoma *et al*., 2017).

A lo largo de su evolución, los organismos que habitan a grandes profundidades del océano han desarrollado rutas metabólicas que les permiten degradar hidrocarburos aromáticos que se encuentran de manera natural en el ecosistema, lo que los convierte en una herramienta ambiental para limpiar ecosistemas perturbados por derrames de petróleo (Scoma *et al.*, 2017).

4.1.2 Hidrocarburos aromáticos en ecosistemas marinos

Los HA son compuestos cíclicos que presentan uno o más anillos de benceno. La presencia de este anillo aromático permite que el compuesto sea muy estable y que se requiera una gran cantidad de energía para romper dicho enlace (Porter & Young 2014). Los HA están ampliamente distribuidos en el ambiente, comprende alrededor de un cuarto de la biomasa de la tierra (Scoma *et al.,* 2017). La principal fuente de HA de origen industrial es el petróleo y sus gases naturales asociados, los cuales son originados por procesos geoquímicos a partir de materia orgánica en condiciones de alta presión hidrostática y temperatura (Heider *et al.*, 1998, Scoma *et al.,* 2017). Sin embargo, los HA también se forman como producto de la degradación de lignina en plantas (Porter & Young 2014).

4.1.3 Biorremediación llevada a cabo por bacterias.

La biodegradación es un método eficaz y rentable para degradar HA, la cual puede ser llevada a cabo por bacterias y hongos que tienen la capacidad de utilizarlos como fuente de carbono y energía (Kumar *et al.,* 2018).

El estudio del fondo marino ha revelado que existe una diversidad microbiana diferente en una zona donde la exposición a HA es prolongada que con aquellas zonas donde no se da esta exposición. Por lo que se sugiere que dicha exposición genera que algunos microorganismos presenten la capacidad de degradar HA (Scoma *et al.,* 2017).

Se ha documentado que algunas especies bacterianas como *Alcaligenes odorans, Achromobacter* sp.*, Mycobacterium* sp.*, Sphingomonas paucimobilis, Mycobacterium flavescens, Pseudomonas* sp.*, Arthrobacter* sp.*, Bacillus* sp.*, Rhodococcus* sp.*, Xanthomonas* sp.*, Alcaligenes* sp.*,* y *Burkholderia cepacia* son capaces de degradar HA (Kumar *et al.*, 2018).

4.1.4 La metagenómica como herramienta para la biorremediación

Los microorganismos como las bacterias, hongos y protozoarios, además de ayudar en la biorremediación de ambientes contaminados, son responsables de la mayoría de los ciclos biogeoquímicos que conforman el medio ambiente de la Tierra y sus océanos. Sin embargo, el estudiar y comprender el potencial metabólico de estos microorganismos se ha visto obstaculizado por la incapacidad de generar cultivos en el laboratorio (Venter *et al.,* 2004).

Se sabe que algunas cepas son eficaces como agentes de biorremediación, pero solo en condiciones de laboratorio. El crecimiento bacteriano se ve limitado por diversos factores, como el pH, la temperatura, la disponibilidad de oxígeno, la estructura del suelo, la humedad, el nivel adecuado de nutrientes y la presencia de compuestos tóxicos. Aunque los microorganismos pueden existir en entornos extremos, la mayoría de ellos crecen en una condición óptima que es difícil de lograr fuera del laboratorio (Karigar, & Rao, 2011). El estudio de las comunidades microbianas hasta hace unos años progresaba lentamente. Sin embargo, los recientes avances en la comprensión de la ecología de los distintos ambientes, como es el caso de los ecosistemas marinos, se han visto facilitada en gran medida por la creciente disponibilidad de datos metagenómicos que proporcionan información sobre la identidad, la diversidad y el potencial funcional de la comunidad microbiana en un lugar y tiempo determinados (Biller *et al.,* 2018)

La metagenómica es un segmento de la genómica microbiana dedicado a la secuenciación y análisis del ADN proveniente del microbioma, obtenido directamente de una muestra ambiental, utilizando tecnología de secuenciación de nueva generación (Aguilar & Falquet 2015). Tras el aislamiento y secuenciación del ADN total de una comunidad microbiana, del cual no se sabe qué fragmento de ADN corresponde a cada organismo presente en dicha comunidad, el estudio metagenómico permite correlacionar el material genético y la posible identidad del organismo del que proviene. Lo que permite la identificación informática de varios microorganismos, en lugar de hacerlo de manera aislada en el laboratorio lo que sería un proceso muy complejo y en algunos casos imposible de realizar. De acuerdo con Lapidus y Korobeynikov, (2021) la metagenómica, también conocida como secuenciación "shotgun", representa un enfoque utilizado para el análisis de comunidades microbianas. A diferencia de métodos como el análisis de 16S rRNA, que se centra exclusivamente en la secuenciación del gen específico 16S rARN presente en bacterias y arqueas, la metagenómica "shotgun" se destaca al secuenciar y analizar todo el genoma. Esta diferencia es crucial, ya que el análisis de 16S rRNA proporciona información limitada sobre las funciones metabólicas de los microorganismos presentes. En cambio, el enfoque "*shotgun*" tiene la capacidad de ofrecer más información sobre las funciones metabólicas, al abordar todo el genoma de los microorganismos en la muestra. Este enfoque permite explorar ecosistemas que albergan microorganismos aún desconocidos (Prakash & Taylor, 2012).

4.1.5 Flujo de trabajo de un proyecto metagenómico

La metagenómica "*shotgun"* permite el ensamble de metagenomas con ayuda de genomas de referencia o también puede hacerse de *novo*. Existen ensambladores genómicos diseñados para el tratamiento de datos metagenómicos, pero presentan algunas limitaciones para lograr con éxito el análisis ya que depende de la complejidad de los datos, la abundancia de los miembros de la comunidad, la calidad de los datos y la disponibilidad de datos experimentales. Los principales objetivos del análisis funcional metagenómico; es determinar cuáles son los repertorios funcionales y metabólicos de los diferentes miembros de la comunidad que les permiten ejercer diferentes efectos, e identificar las variaciones, si las hay, dentro de las composiciones funcionales de las diferentes comunidades (Lapidus & Korobeynikov, 2021).

El flujo de trabajo para la realización de un estudio metagenómico consiste en:

- La obtención de la muestra ambiental a estudiar.

- La extracción y secuenciación del ADN utilizando secuenciadores de nueva generación como Illumina, que procesa secuencias cortas de 50 a 300 pares de bases, y Oxford Nanopore, que permite procesar secuencias largas de 50 a 2 Mb (Hon *et al*., 2020; Illumina, 2023; Oxford Nanopore Technologies, 2024).

- El preprocesamiento de datos una vez que se tienen las lecturas (*reads*), debido a que éstos pueden presentar contaminación por el adaptador, sesgos en el contenido de bases y secuencias sobrerrepresentadas. Por lo tanto, se debe realizar un análisis de control de calidad con programas como Fastp, el cual es una eficaz herramienta que ayuda a filtrar los datos para eliminar secuencias del adaptador, intrones y colas de poliA y poliG (Chen *et al.,* 2018), permitiendo obtener las secuencias de mejor calidad para los siguientes pasos.

- El ensamblaje de las lecturas de mayor calidad con programas como MEGAHIT (Dinghua *et al.,* 2015) o meta SPADES (Nurk *et al.,* 2017), en donde estas lecturas son ensambladas mediante gráficos de de Bruijn con el objetivo de reconstruir la secuencia original de cada organismo. Posteriormente, para evaluar la calidad del ensamble metagenómico se puede utilizar el programa metaQUAST (Mikheenko *et al.,* 2016).

- La predicción de genes para la identificación de secuencias de ADN de la muestra con programas como BLAST y MEGAN (Kunin, *et al.,* 2008).

- La anotación funcional de genes, donde se comparan los genes predichos con secuencias existentes previamente anotadas con el objetivo de obtener anotaciones precisas de genes homólogos de ser posible genes ortólogos, los cuales presentan la

misma función biológica y son derivados de eventos de especiación (Altenhoff *et al.*, 2019).

Sin embargo, cada etapa del análisis presenta dificultades debido a los problemas inherentes de los datos metagenómicos como la cobertura incompleta del ADN, lo cual impide obtener ensambles completos, lo que resulta en la predicción de genes fragmentados que no muestran ninguna coincidencia en las bases de datos de secuencias de referencia. Así como la representación desigual de los miembros de la comunidad microbiana y la presencia de microorganismos estrechamente relacionados con genomas similares. Para resolver estos problemas, se han creado y aplicado varios enfoques y líneas de análisis (National Research Council 2007, Prakash & Taylor 2012).

El enfoque más simple y comúnmente adoptado por la mayoría de los flujos de trabajo para la predicción funcional se basa en la búsqueda de homología a través de diversas bases de datos. Por ejemplo, el programa BLAST (Basic local alignment search tool) realiza una búsqueda de similitud de secuencias de proteínas predichas en las bases de datos que contienen secuencias de proteínas de referencia (She *et al.,* 2009); sin embargo, también se pueden usar bases de datos de dominios de proteínas como *Pfam* que son utilizadas para inferir dominios en las proteínas predichas. Por otro lado, el uso de programas como HMMER, el cual se fundamenta en el algoritmo basado en el modelo oculto de Markov, permite una búsqueda más sensible de perfiles Pfam (Prakash & Taylor 2012, Simon *et al.*, 2018).

Para solventar los problemas relacionados a la generación de secuencias cortas que generalmente exhiben pocas similitudes con las secuencias en las bases de datos existentes o la presencia de especies nunca reportadas, se pueden utilizar métodos alternos para la predicción funcional. Por ejemplo, un enfoque basado en la búsqueda de motivos o patrones estructurales de aminoácidos que son necesarios para la estructuración y funcionalidad biológica de la proteína. Otro método para superar estas limitaciones consiste en el enfoque basado en el contexto genético o vecindad genómica, el cual consiste en observar la similitud de los genes que rodean al gen de la proteína predicha. Sin embargo, debido a la escasez de genomas completos en los conjuntos de datos metagenómicos y la falta de conocimiento sobre el verdadero origen de las secuencias, este enfoque tiene sus limitaciones. Estos problemas se pueden mejorar aumentando la profundidad de la secuencia, generando *contigs* más largos y mejorando la asignación taxonómica de las secuencias (Figura 1) (Kunin, *et al.*, 2008, Prakash & Taylor 2012).

Posteriormente, se realiza la clasificación y agrupamiento de lecturas en unidades taxonómicas operacionales (OTUs). En este proceso se da la asociación entre las secuencias y la identidad de las especies presentes, se trata de generar un "ensamblaje" o cierre del *cluster* (Figura 1) (National Research Council 2007, Biller *et al.*, 2018, Lapidus & Korobeynikov, 2021).

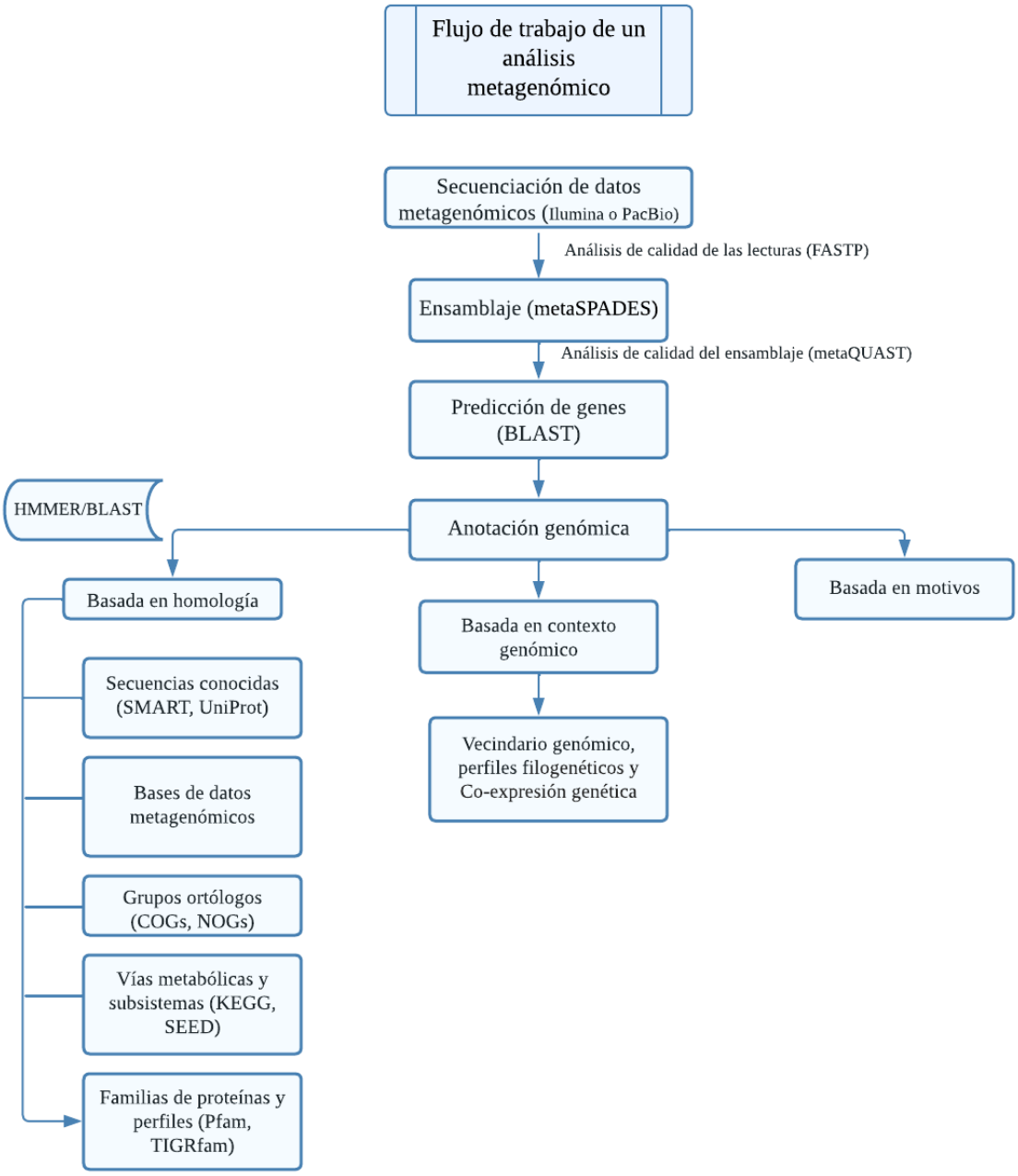


Figura 1. Ejemplo de flujo de trabajo para la realización de un análisis metagenómico.
Se muestra el flujo de trabajo básico hasta el paso de predicción de genes. Posteriormente, en la parte de anotación genómica, se observan los posibles enfoques que se han utilizado para genomas secuenciados y que se sugiere se pueden emplear y probar para la anotación en metagenomas. El enfoque del contexto genómico se ha implementado en conjuntos de datos metagenómicos, (modificado de Prakash & Taylor 2012).

4.1.6 La inferencia de ortología es fundamental para la anotación de los genomas y los metagenomas.

En un estudio genético evolutivo se lleva a cabo la identificación, dentro o entre especies, de regiones homólogas de ascendencia común. La homología se define como la relación existente entre genes que comparten un ancestro común. Por lo tanto, al alinear los metagenomas secuenciados con genomas de referencia se busca comprender su historia evolutiva. Sin embargo, los genomas experimentan cambios estructurales a gran escala, tales como duplicaciones y reorganizaciones, por lo que la tarea de deducir su historia evolutiva se hace más compleja. Las relaciones evolutivas de los genomas se pueden explicar a través de las tres principales subclases de homología: ortología, paralogía y xenología (Altenhoff *et al.,* 2019).

Las secuencias ortólogas son homólogos que divergieron de su ancestro común más reciente debido a un evento de especiación, mientras que las secuencias parálogas son homólogos que divergieron del ancestro común más reciente debido a un evento de duplicación. Por otro lado, las secuencias xenólogas son homólogos que divergieron a partir de un evento de transferencia horizontal. Las secuencias ortólogas son de interés primario porque son útiles para aplicaciones como la predicción de funciones y la inferencia de árboles de especie. La diferenciación entre genes ortólogos y parálogos resulta crucial para la predicción de la función génica. Los genes ortólogos, al haber sido el mismo gen en el último ancestro común de las especies implicadas, se considera que probablemente desempeñan funciones biológicas similares. Por el contrario, los genes parálogos, derivados de duplicaciones conservadas, suelen diferir en sus funciones (Altenhoff *et al.,* 2019). Un método para inferir ortología en un grupo de secuencias dado se basa en la construcción de árboles filogenéticos, en los cuales las divisiones se anotan como un evento de duplicación o especiación con base en la filogenia de las especies relevantes (Altenhoff *et al.,* 2019).

4.1.7 Genomas ensamblados de metagenomas

La mayoría de los estudios metagenómicos y genómicos han dependido de la disponibilidad de genomas de referencia. Sin embargo, cuando los genomas de referencia están incompletos o no existen, como en el caso de organismos nuevos, se requiere de la construcción de genomas ensamblados a partir de los genomas ensamblados de metagenomas (MAGs Metagenome-assembled genomes), lo cual facilita la reconstrucción de genomas de las especies individuales de ambientes naturales (Lin & Liao 2016). Un MAG se refiere a un grupo de *scaffolds* con características similares provenientes de un metagenoma, que se agrupan y en conjunto representan al genoma microbiano. Es decir, las lecturas obtenidas de la secuenciación se ensamblan en *scaffolds* que son secuencias más largas que los *reads* y luego éstos se agrupan en MAGs candidatos basados en frecuencias de trinucleótidos (TNFs), abundancias, genes marcadores complementarios, alineaciones taxonómicas y el uso de codones. Los MAGs con alta integridad y bajos niveles de contaminación se utilizan para la anotación taxonómica adicional y predicción de genes (Yang *et al.,* 2021).

4.1.8 Aplicación de la metagenómica para estudiar ecosistemas marinos perturbados

El avance en el entendimiento de los sistemas marinos microbianos ha sido facilitado por el incremento de la disponibilidad de datos metagenómicos, los cuales proveen información sobre la identidad y diversidad funcional de las comunidades microbianas (Biller *et al.*, 2018). Los microorganismos oceánicos desempeñan un papel fundamental en los ciclos biogeoquímicos, ya que su metabolismo colectivo tiene efectos globales en los flujos de energía y materia en el mar y en la composición de la atmósfera de la Tierra (National Research Council 2007).

Actualmente, es de gran interés investigar la distribución de genes y organismos en los océanos, especialmente aquellos relacionados con la degradación de HA. Se han llevado a cabo diversos estudios que se centran en el análisis de metagenomas marinos con el propósito de caracterizar organismos presentes en los océanos (Venter *et al.*, 2004). En el estudio realizado por Venter y colaboradores en 2004, se empleó un enfoque de secuenciación metagenómica *shotgun* en poblaciones microbianas recolectadas en el Mar de los Sargazos, (https://www.ncbi.nlm.nih.gov/nuccore/AACY00000000.1?report=genbank). Como resultado, obtuvieron información sobre el contenido genético y la abundancia relativa de los organismos presentes en las muestras. Posteriormente, en 2007, ampliaron la expedición y recolectaron muestras de la costa este de América del Norte, el Golfo de México y el Océano Pacífico (Rusch *et al.*, 2007).

En el proyecto dirigido por Anderson (*et al.*, 2014) realizan la colecta de muestras de agua en cruceros GEOTRACES a través del océano global (Anderson *et al*., 2014; Biller *et al.*, 2018). Por otro lado, en la expedición de Tara Oceans se recolectaron muestras de la parte epipelágica y mesopelágica en el océano global, generando un catálogo de 40 millones de genes provenientes de virus, bacterias, arqueas y protistas (Pesant *et al.*, 2015). Este estudio proporcionó también datos fisiológicos como el pH, la cantidad de oxígeno disponible, la temperatura y la cantidad de clorofila, lo que permitió correlacionar los hábitats con los organismos que habitan en ellos.

La creciente información sobre microorganismos marinos generada por estas expediciones ha permitido realizar nuevos estudios utilizando la información sobre los metagenomas existentes. Tal como en el estudio realizado por Loza (*et al.*, 2022), donde utilizan y analizan datos metagenómicos para obtener un perfil taxonómico y el potencial metabólico de las poblaciones que habitan a distintas profundidades en localidades perturbadas o contaminadas en el Golfo de México. De la comparación de las enzimas anotadas a partir de dos muestras de dicho estudio, contra las enzimas de referencia provenientes del proyecto de Biller (*et al.*, 2018), se utilizaron las enzimas que presentan abundancias muy diferentes a las de las enzimas de referencia para realizar una reconstrucción metabólica, en la cual identificaron como sobrerrepresentadas a un conjunto de enzimas involucradas en la degradación de hidrocarburos. Entre las vías metabólicas mejor representadas se encontraron a las de aminobenzoato, benzoato, caprolactam y tolueno.

De manera particular identificaron una enzima clave implicada en la degradación anaeróbica de HAs a partir de la vía del benzoil-CoA, denominada benzoato CoA ligasa (BCL) (EC 6.2.1.25), la cual participa en la reacción de transformación del benzoato a benzoil-CoA (Figura 2). A través de esta vía se da la degradación de hidrocarburos monoaromáticos, como benzoato, fenol, tolueno y etilbenceno. Sin embargo, los siguientes pasos para la conversión del benzoil-CoA a succinil-CoA o acetil pueden proceder a través de diferentes vías. Algunas de estas vías han sido descritas en bacterias como *Thauera aromatica (T. aromatica)* y *Azoarcus* sp. CIB, (Porter & Young, 2014). El hallazgo de las enzimas implicadas en las vías de degradación de hidrocarburos indica que la distribución de las comunidades microbianas, dependiendo de la profundidad donde se localizan, puede ser específica de ese sitio, es decir, su actividad depende de las condiciones fisicoquímicas del ambiente donde habitan.

Otro gran esfuerzo por analizar los metagenomas marinos de diferentes partes del mundo fue el realizado por el equipo de Nishimura y Yoshizawa en 2022, donde colectaron 2057 metagenomas marinos provenientes de muestras de agua de distintas profundidades, así como de sedimento y biopelícula. A partir de estos datos se reconstruyeron 52,325 MAGs. Las secuencias genómicas resultantes fueron depositadas en el proyecto denominado OceanDNA MAGs. Este estudio permitió descubrir linajes que participan en ciclos biogeoquímicos importantes para el desarrollo de la vida en la tierra, así como la caracterización de los potenciales metabólicos de especies no cultivables y la reconstrucción filogenética de microorganismos. Cada uno de estos proyectos abre el camino a nuevas investigaciones con el objetivo de descubrir cómo funcionan las comunidades de microorganismos marinos que habitan en sitios contaminados por HAs.

4.2 La vía de la benzoil CoA.

4.2.1 Enzima de interés Benzoato CoA ligasa

La enzima BCL pertenece a la clase I de la superfamilia de enzimas formadoras de adenilato (Adenylate forming enzyme, ANL). Su función principal radica en catalizar la formación de un enlace tioéster entre compuestos aromáticos y la coenzima A (CoA o CoASH), utilizando ATP como cofactor (Figura 2). El proceso se inicia con la reacción en la que el oxígeno con carga negativa presente en el benzoato ataca al grupo α-fosfato del ATP, generando así un producto intermedio denominado aril-adenilato y liberando pirofosfato (PPi). En una segunda etapa, este intermediario es objeto de un ataque por parte del grupo tiol de la CoA, lo que da lugar a la liberación de AMP y la formación de benzoil CoA (Figura 3) (Kawaguchi *et al.*, 2006, Arnold *et al.*, 2021).
En las bacterias que llevan a cabo la degradación de compuestos aromáticos, como es el caso de *Rhodopseudomonas palustris* (*R. palustris*), la formación de tioésteres CoA cumple una función crucial ya que facilita la acumulación de sustratos dentro de la célula al reducir el gradiente de concentración entre el citoplasma y el entorno externo. Además, su tamaño voluminoso y su estructura polar evitan la retrodifusión de estos compuestos fuera de la célula (Arnold *et al.* 2021).

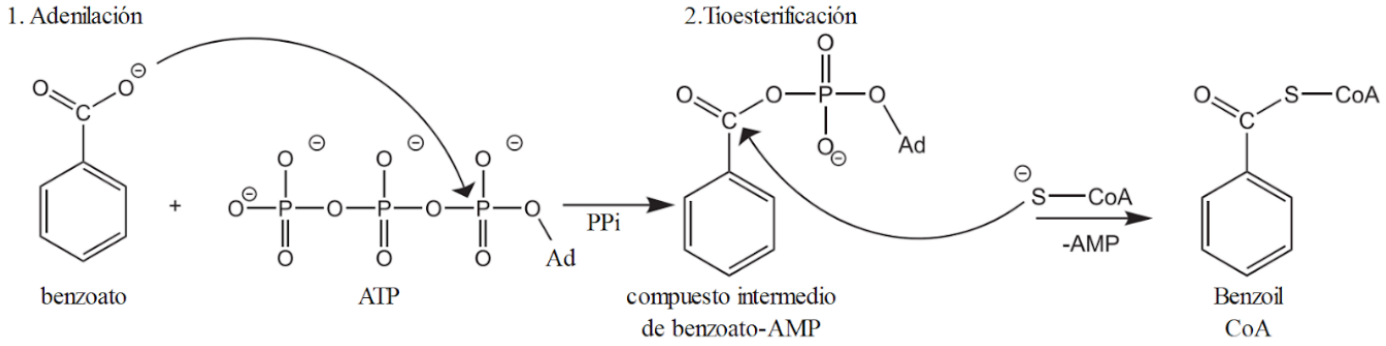


Figura 2. Mecanismo de acción de la Benzoato CoA ligasa. La primera parte involucra una reacción de adenilación, donde el grupo carboxilo del benzoato realiza un ataque nucleófilo al α-fosfato del ATP, liberando fosfato y un intermediario benzoato-AMP. En la segunda parte del proceso, se da una reacción de tioesterificación, donde el grupo tiol de la CoA ataca al intermediario benzoato AMP, dando lugar a la formación de benzoil CoA, el cual podrá ingresar a diferentes vías para poder ser degradado (Modificado de Arnold *et al.*, 2021).

4.2.2 Catabolismo aeróbico del benzoil CoA.

La biodegradación aeróbica de compuestos aromáticos ha sido ampliamente estudiada en las últimas décadas. Para llevar a cabo el catabolismo aeróbico del benzoil CoA se requiere de la presencia de oxígeno molecular, el cual actúa como último aceptor de electrones y, a su vez, como co-sustrato para activar el anillo aromático del compuesto (Brzeszcz & Kaszycki, 2018). Las enzimas encargadas de llevar a cabo este proceso se denominan oxigenasas (Figura 3). De manera general el mecanismo que llevan a cabo dichas enzimas consiste en la hidroxilación y la escisión del anillo aromático (Díaz *et al.*, 2013). Las oxigenasas a su vez son agrupadas en monooxigenasas y dioxigenasas de acuerdo con la cantidad de átomos de oxígeno que agreguen a la reacción (Laczi *et al.*, 2020).

El primer paso del catabolismo consiste en la adición de átomos de oxígeno y grupos hidroxilo al compuesto aromático, generando la desestabilización electrónica en el anillo de benceno provocando que se vuelva más reactivo (Karigar & Rao, 2011). El segundo paso consiste en la oxigenación del compuesto aromático, en el cual las dioxigenasas utilizan al compuesto aromático resultante, que comúnmente es catecol, como sustrato y catalizan la escisión o ruptura del anillo mediante la adición de oxígeno (Díaz *et al.*, 2013). El resultado es un compuesto alifático que pasará por reacciones de oxidación progresiva, terminando con la incorporación de los metabolitos resultantes al ciclo de Krebs (Díaz *et al.*, 2013).

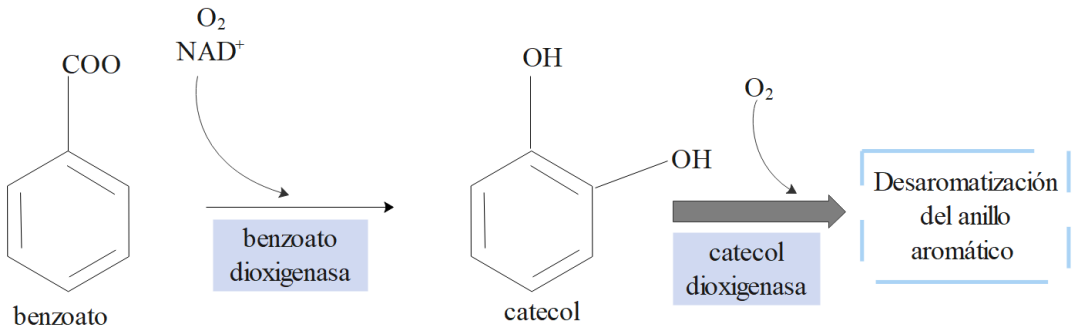


Figura 3. Mecanismo del catabolismo aeróbico del benzoato. En el catabolismo aeróbico, las enzimas oxigenasas y dioxigenasas llevan a cabo la hidroxilación (activación) y escisión (desaromatización) del anillo aromático, generando como producto intermediario la molécula de catecol, al cual la enzima catecol dioxigenasa incorpora átomos de oxígeno para llevar a cabo la desaromatización del anillo aromático (Modificado de Valderrama *et al.*, 2012).

4.2.3 Catabolismo anaeróbico del benzoil CoA

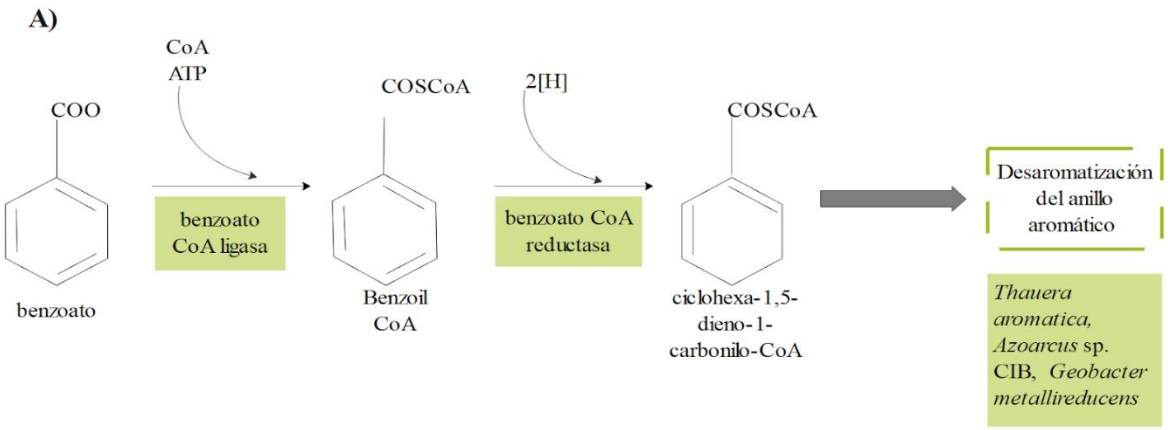
Las herramientas genómicas han permitido un incremento en el conocimiento del catabolismo anaeróbico de compuestos aromáticos en organismos que habitan en ambientes anóxicos y que utilizan dichos compuestos como única fuente de carbono. La metagenómica ha permitido la identificación de organismos degradadores de hidrocarburos aromáticos debido a que normalmente habitan en ambientes perturbados generados por la dispersión de sustancias químicas derivadas del petróleo generando así condiciones imposibles de replicar en el laboratorio, (Laczi *et al.*, 2020).

En ausencia de oxígeno el benzoato se puede degradar por dos vías catabólicas alternativas, una de ellas la llevan a cabo organismos aeróbicos facultativos y la otra, por organismos anaeróbicos estrictos. Sin embargo, ambos metabolismos son reductivos, (más adelante se tratará a fondo la vía híbrida). Sabemos que en la degradación aeróbica se utiliza una gran cantidad de energía para llevar a cabo la escisión del anillo aromático a través de la adición de átomos de oxígeno. En cambio, en el catabolismo anaeróbico los organismos aerobios facultativos, fotótrofos y anaeróbicos estrictos utilizan un metabolismo aromático reductor (Fuchs, 2008). En los organismos anaeróbicos obligados el benzoil-CoA se degrada a acetil-CoA y CO_2 después de una serie de reacciones que constituyen la vía de biodegradación del benzoil-CoA (Carmona *et al.*, 2009). Por otra parte, la reducción del anillo aromático se cataliza por la enzima benzoil-CoA reductasa utilizando una variedad de aceptores de electrones (Figura 4, panel A). Por ejemplo, el nitrato, el sulfato, el hierro (III), el manganeso (II), y el selenato, cada uno conservando diferentes rendimientos de energía (Fuchs, 2008).

4.2.4 Catabolismo híbrido del benzoato

En presencia de bajas concentraciones de oxígeno, los organismos aerobios facultativos utilizan un tipo de metabolismo denominado híbrido en el cual hacen uso de oxígeno para introducir grupos hidroxilo, como en la ruta aeróbica clásica. Al mismo tiempo, reducen el anillo aromático y utilizan tioésteres de CoA, como en el metabolismo anaeróbico. En este caso la escisión del anillo tampoco requiere oxígeno. En la vía híbrida aeróbica (genes *box*) se inicia la activación del benzoato a benzoil CoA por acción de la BCL. Posteriormente, una benzoil CoA 2,3-epoxidasa (BoxAB), una benzoil dihidrodiol liasa (BoxC), y una 3,4-deshidrogodifenil-CoA semialdehído deshidrogenasa (BoxD) son responsables de los pasos de desaromatización y escisión del anillo, respectivamente (Figura 4, panel B) (Valderrama *et al.,* 2012). Sin embargo, se conocen pocos organismos con este tipo de metabolismo. También se ha reportado la presencia de enzimas reductasas, como la benzoil-CoA reductasa (Bcr, gen *bcr*), para desaromatizar el anillo aromático. Esta reducción es impulsada por la hidrólisis de dos moléculas de ATP. En el catabolismo híbrido del benzoato ocurre una sustitución de los pasos dependientes de oxígeno por un conjunto alternativo de reacciones y la formación de diferentes intermediarios centrales. Cabe destacar que la reducción de dos electrones del anillo aromático del benzoil-CoA es impulsada por la hidrólisis de dos moléculas de trifosfato de adenosina (ATP). El producto cíclico no aromático formado se abre hidrolíticamente y finalmente se oxida a tres moléculas de acetil-CoA (Fuchs, 2008).

Las especies bacterianas representativas incluyen a la bacteria fototrópica *R. palustris,* y a las bacterias desnitrificantes *T. aromática*, *Azoarcus evansii* renombrada *Aromatoleum evansii (A. evansii)*, *Aromatoleum* sp. y *Paraburkholderia xenovorans* LB400 (*P. xenovorans*). El rendimiento de energía (ATP) por el metabolismo en estas bacterias es alto, en comparación con anaerobios. Las vías dependientes de tioéster CoA pueden ser ventajosas en condiciones fluctuantes óxicas/anóxicas, ya que permiten flexibilidad y una rápida adaptación a diferentes niveles de oxígeno (Fuchs, 2008, Carmona *et al.*, 2009).



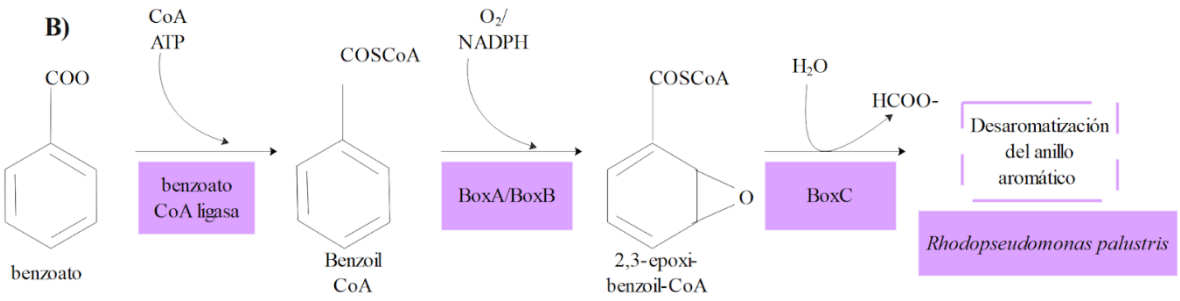


Figura 4. Estrategias anaeróbica e híbrida para la degradación del benzoil CoA. A) Se muestra el metabolismo anaeróbico estricto, en el cual participan enzimas reductasas. En el catabolismo anaeróbico, la activación de anillo aromático es dependiente de la coenzima A (CoA), seguido de una desaromatización reductiva y luego una escisión del anillo hidrolítico. B) En este panel se observa el metabolismo aeróbico híbrido, en el cual participan los genes Box ABC. Esta ruta híbrida hace uso de oxígeno para introducir grupos hidroxilo, como en las rutas aeróbicas clásicas. Al mismo tiempo, se reduce el anillo aromático y se utilizan tioésteres de CoA, como en el metabolismo anaeróbico. La escisión del anillo tampoco requiere oxígeno (Modificado de Fuch 2008 y Valderrama *et al.*, 2012).

4.2.5 Regiones conservadas en las CoA ligasas

El estudio de las regiones estructuralmente conservadas (RCS) consiste en el modelado de estructuras y la alineación de secuencias conservadas, para identificar relaciones de homología entre organismos que se consideran provenientes de un ancestro común. Generalmente se considera que un alto nivel de semejanza entre secuencias de aminoácidos implica homología en una proteína. Sin embargo, la deducción de relaciones de homología mediante el uso de estructuras tridimensionales suele ser más precisa, ya que las estructuras tienden a ser más conservadas que las secuencias (Huang *et al.*, 2013). La búsqueda de similitudes es efectiva y confiable porque las secuencias que comparten similitudes significativas pueden inferirse como homólogas (Pearson, 2013). Por otro lado, cuando las proteínas presentan similitud en su secuencia y función se considera que han evolucionado a partir de un ancestro común, en este caso se considera que dichos organismos son ortólogos (Snyder & Champness 2007).

La enzima benzoato-CoA ligasa muestra una estructura de dos dominios. El dominio N-terminal, el cual contiene casi todos los residuos que se unen al sustrato carboxilado y también se une al grupo adenosil del ATP. El dominio C-terminal coordina a los grupos ribosa y fosfato (Figura 7) (Thornburg, *et al.*, 2015). Una rotación de 140° del dominio C-terminal permite a la enzima catalizar las dos reacciones distintas de adenilación y tioesterificación. De acuerdo con el estudio de Arnold *et al.*, 2021, se han detectado nueve dominios conservados ampliamente distribuidos en enzimas pertenecientes a la clase I de la superfamilia (ANL).

5. Justificación

En contraste con el catabolismo aeróbico de HA para el cual se dispone de mucha información, aún se sabe poco acerca de la distribución de los genes involucrados en el catabolismo anaeróbico e híbrido de dichos compuestos en bacterias, en donde la vía de benzoil-CoA ha sido poco explorada. En bacterias marinas, de ecosistemas perturbados con HA, se ha observado la presencia de esta vía de degradación (Loza *et al.*, 2022). Estudios previos de nuestro grupo sugieren la conservación de la enzima BCL en metagenomas del Golfo de México en donde existe una constante exposición a hidrocarburos derivados del petróleo. No obstante, la degradación de HA puede también ocurrir en sistemas no perturbados como los yacimientos naturales donde se produce el petróleo en el fondo marino o en ecosistemas donde se ha dado la dispersión de los HA. En este trabajo se explorarán genomas totalmente secuenciados de diversos ambientes y MAGs provenientes de distintas localidades marinas del mundo, para evaluar la distribución de las vías de degradación del benzoato, tomando como punto de partida a la BCL, reconocida por llevar a cabo un paso enzimático indispensable de estas rutas (transformación de benzoato a benzoil-CoA). Los resultados obtenidos nos permitirán identificar qué especies presentan el gen que codifica para la enzima BCL, así como las vías hibridas, aeróbicas y anaeróbicas por las que procede la degradación del benzoil-CoA, producto de la catálisis llevada a cabo por la BCL.

6. Hipótesis

La distribución de los genes de la BCL y los genes implicados en la vía del benzoil-CoA son determinantes en la adaptación de diferentes linajes bacterianos a condiciones ambientales específicas, lo que influiría en las estrategias de degradación de sustratos aromáticos en diversos ecosistemas.

7. Objetivos

7.1 Objetivo general

Determinar si la enzima BCL es un marcador de degradación de HAs, en organismos totalmente secuenciados y MAGs derivados de muestras oceánicas de agua.

7.2 Objetivos específicos

1. Plantear una metodología que nos permita identificar ortólogos de la BCL.
2. Identificar ortólogos de la BCL en genomas y MAGs.
3. Definir la vía de degradación en función de los requerimientos de oxígeno (aeróbico, anaeróbico e híbrido)
4. Evaluar la distribución de la enzima BCL en genomas totalmente secuenciados (modelos).
5. Evaluar la distribución de la BCL y la vía de degradación del benzoil-CoA en MAGs marinos.
6. Analizar la correlación entre los ambientes y los organismos completamente secuenciados y MAGs.
7. Evaluar la conservación de las enzimas involucradas en la vía de degradación del benzoil-CoA.

8. Métodos

8.1 Búsqueda de organismos que presentan la vía del Benzoil-CoA

Se investigo acerca de los organismos que presentan la enzima BCL. La búsqueda se realizó en bases de datos como Google Académico, Connected Papers, KEGG (Kanehisa & Goto, 2000) y BRENDA Enzyme (Chang A. *et al.,* 2021). Para realizar la búsqueda se utilizaron palabras clave como: metabolismo anaeróbico del benzoato, vía de degradación de benzoil-CoA, BCL y organismos capaces de degradar hidrocarburos aromáticos. Una vez identificados los organismos, se buscó a cada uno de ellos en la base de datos KEGG y se comprobó la presencia en su genoma de los genes que codifican para la enzima BCL, realizando una búsqueda por su número enzimático (6.2.1.25).

8.2 Construcción de arquitecturas de proteína

Para realizar la búsqueda de ortólogos de la BCL se utilizó el enfoque desarrollado por nuestro grupo (Martinez-Amador *et al.,* 2019; Soto *et al.,* 2021), en el que se emplea, como primer paso para la identificación de ortólogos, la construcción de arquitecturas de proteínas. Una arquitectura de proteína se compone de un dominio o una combinación de dos o más dominios Pfam no traslapados que al ser identificados por el programa *hmmerscan* presentan un valor esperado (e) < = 0.001, y que cubren la longitud de la proteína tanto como sea posible. Para identificar secuencias homólogas de la BCL en los proteomas de bacterias, se construyó la arquitectura de proteína que consta de los dominios Pfam *AMP-binding* (PF00501*) y AMP-binding_C* (PF13193), considerando a los dominios encontrados en los organismos con evidencia experimental de presentar la vía. Estas secuencias presentaron una longitud promedio de 530 aa. En la base de datos KEGG (Kanehisa & Goto, 2000) se identificaron los genes de los organismos modelos que contaban con un identificador de ortología KEGG (K04110), el cual sirvió para etiquetar a los ortólogos predichos.

Para optimizar la identificación de ortólogos, se analizó el contexto genómico del gen que codifica para la BCL. En este caso, varios genes vecinos codifican las vías metabólicas que actúan río abajo y catalizan la transformación del sustrato benzoil CoA. Los genes involucrados en estas vías se organizan frecuentemente en operones (Egland *et al.*, 1997; Porter & Young, 2014; Arnold *et al.,* 2021). Por esta razón, en lugar de utilizar tres genes anteriores y tres posteriores, como se sugiere en trabajos anteriores (Soto *et al.,* 2021; Martinez-Amador *et al.*, 2019), se amplió la inspección del contexto genómico a diez genes anteriores y posteriores desde la posición en la que se encontraba el gen de la BCL. Posteriormente, se identificaron en KEGG los dominios Pfam de los diez genes presentes en el contexto genómico de la BCL. Las matrices de todos los dominios Pfam asociados a los genes del contexto se descargaron de la base de datos Pfam-A (Mistry *et al.,* 2020); posteriormente, se construyó la arquitectura de proteína que presentaba los dominios Pfam tanto de la BCL como de las proteínas vecinas.

8.3 Búsqueda de homólogos de la Benzoato-CoA ligasa en genomas secuenciados.

La arquitectura de proteínas se empleó para escanear los 6536 proteomas bacterianos presentes en la base de datos KEGG comprada por nuestro grupo en junio de 2022, utilizando *hmmscan* de HMMR suite (Simon *et al.,* 2018), el cual realiza una búsqueda de secuencias contra una base de datos de perfiles curados y almacenados en la base de datos Pfam-A (Mistry *et al.,* 2020). *hmmscan* requiere como datos de entrada a las secuencias de aminoácidos en formato fasta y las matrices Pfam de interés. Se corrió el programa *hmmscan* utilizando los parámetros predeterminados. Se consideran homólogos a todas aquellas secuencias con un valor esperado <= -E 0.001 HMMR suite (Simon *et al.,* 2018). Las matrices utilizadas incluyeron a los dominios Pfam de la BCL y de las proteínas del vecindario (Tabla A1 del anexo). Con los resultados obtenidos, se reconstruyó el contexto genómico de cada gen que codificaba para una BCL predicha que presentaba los dominios *AMP-binding* y *AMP-binding_C* en este orden.

8.4 Identificación de ortólogos con base en el contexto genómico

Para discernir entre homólogos de la BCL, se seleccionaron aquellas proteínas que presentan en su genoma genes vecinos que codifican para la subunidad de la 2,3-epoxidasa benzoil-CoA (*boxA*, *boxB*) o todas las subunidades de la benzoil-CoA reductasa (*bcr*). En esta etapa, también se incluyen las secuencias de genes de BCLs que tienen como vecinos genes que codifican para un regulador transcripcional de la familia XRE (*xenobiotic response element*), o la 3,4-dehidrocatil-CoA semialdehído deshidrogenasa (*boxD*) o a la benzoil-CoA-dihidrodiol liasa (*boxC*). Si existían, se asociaron los identificadores de KEGG (KO) de la BCL y de las proteínas codificadas en el contexto genómico (Tabla A1, anexo). Los datos complementarios muestran los identificadores de Pfam relacionados a cada proteína, las descripciones de los dominios y los KOs asociados disponibles (Tabla A1, anexo).

8.5 Corroboración de ortólogos predichos por motivos conservados.

Para determinar si los genes ortólogos de la BCL comparten motivos conservados que las distinguiera de otras aril CoA se utilizó el programa MEME (Bailey y Gribskov, 1998). Al programa MEME se proporcionaron las secuencias de los homólogos no redundantes de las secuencias de BCL para construir motivos conservados, que conservaron al menos uno de los genes vecinos encontrados en los organismos experimentalmente caracterizados. Las BCLs no redundantes se seleccionaron ejecutando CD-HIT con los parámetros predeterminados (Fu *et al.,* 2012). Si CD-HIT no seleccionó una BCL caracterizada experimentalmente en un *cluster,* se sustituyeron las secuencias sugeridas por CD-HIT por aquellas reportadas experimentalmente (Tabla A1, del anexo). Se corrió el programa MEME, estableciendo motivos con una longitud mínima de seis aminoácidos y una longitud máxima de 30, pidiendo al menos ocho matrices. Estas longitudes, así como el número de motivos solicitados, representaron los motivos reportados encontrados en la clase I, subclase Ib de las aril-CoA ligasas (Arnold *et al.,* 2021; Clark *et al.,* 2018; Muroski *et al.,* 2022). Las matrices resultantes fueron utilizadas por el programa MAST (Bailey & Gribskov, 1998), que escaneó

los homólogos de BCL con arquitectura de proteína conservada (*AMP-binding AMP-binding_C*) y el contexto genómico sugerido.

8.6 Porcentaje de GC por gen.

Se descargaron de la base de datos KEGG comprada por nuestro grupo en junio de 2022 las secuencias de los 6536 genomas. El GC% se calculó de la siguiente manera: Count(G + C)/Count(A + T + G + C) * 100%.

8.7 Obtención de las secuencias de MAGS marinos

Se seleccionaron las secuencias de las especies representativas a partir de los MAGs marinos obtenidos en la investigación del grupo de Nishimura y Yoshizawa (2022), en la cual utilizan la información de los metagenomas registrados en los proyectos antes mencionados y proponen una metodología mejorada para realizar la reconstrucción de MAGs.

Obtuvimos la lista de los MAGs resultantes, proporcionada en los datos suplementarios de Nishimura y Yoshizawa (2022). Posteriormente, se extrajo del archivo original el identificador de la base de datos de NCBI, así como el identificador SRA (Sequence Read Archive) de cada muestra. A continuación, se desarrolló un script en Perl que toma como valor de entrada el identificador de la base de datos NCBI assembly. Utilizando este dato, se obtuvo el identificador de descarga del servidor FTP de la base de datos GENOMES (ftp.ncbi.nlm.nih.gov/genomes/) mediante el procesamiento de cadenas de texto (parseo). Con esta información, se descargaron los archivos genómicos de la lista de especies representativas desde el repositorio correspondiente, se descomprimieron y se organizaron en directorios basados en el identificador único de MAG. Además, se analizaron 11 binns (conjuntos binarios de lecturas de secuenciación) ensamblados del Golfo de México, los cuales se obtuvieron del material suplementario de Loza (*et al.*, 2022). Estos binns representan agrupaciones específicas de información genómica y proteómica derivadas de secuencias metagenómicas (Mardanov *et al*., 2018). Es importante destacar que los proteomas del Golfo de México incluidos en estos binns ya habían sido previamente trabajados por miembros del laboratorio.

8.8 Identificación de ortólogos de la enzima Benzoato-CoA ligasa en los MAGs

Los archivos fasta de 1526 MAGs del estudio de Nishimura y Yoshizawa (2022), ahora almacenados en el servidor de nuestro equipo (Agora), dichos archivos fueron escaneados usando el programa *hmmscan* con las matrices tomadas de la base de datos Pfam-A de los dominios *AMP-binding* y *AMP-binding C*. Así mismo, se usaron matrices que describen a las arquitecturas de las proteínas del contexto genómico. Los parámetros utilizados fueron los mismos que los utilizados para los genomas secuenciados, mencionado anteriormente. Una vez identificadas las secuencias, se seleccionaron a todas aquellas que tuvieran para cada dominio un valor esperado (e-value <= 0.001). El orden fue verificado usando las coordenadas de inicio y fin reportadas por *hmmerscan.* Finalmente, se aplicó un último filtro usando las matrices MEME obtenidas de los genomas totalmente secuenciados para identificar, con el programa MAST (Timothy *et al.*, 2015), al subconjunto de secuencias que

conservó los motivos en el número y orden identificado en los genomas de los organismos modelos.

8.9 Curación manual de los ortólogos identificados en los MAGs.

Para seleccionar de manera manual los ortólogos más probables dentro de los MAGs, se realizó una curación basada en que el probable ortólogo presentara el contexto genómico identificado en los modelos antes descritos, y que éstos fueran consecutivos o ubicados a una distancia similar a la reportada para los genes descritos en la literatura en relación con su posición dentro del MAG.

8.10 Construcción de árbol filogenético.

Las secuencias se alinearon con MUSCLE 5 (Edgar, 2022) y se recortaron utilizando los parámetros predeterminados de TRIMAL (Capella-Gutiérrez *et al.*, 2009). Para reconstruir el árbol, utilizamos IQ-TREE multinúcleo versión 1.6.12 para Linux 64-bits (máxima probabilidad) con corrección aLRT (una aproximación de prueba estándar de probabilidad) con 1000 réplicas para identificar el modelo de distancia de arranque óptimo (Nguyen *et al.*, 2015). Seis secuencias fueron seleccionadas como un grupo de árboles. Las secuencias fueron tomadas de aril-CoA ligasas utilizadas en el trabajo original de Arnold *et al.*, 2021, que utilizó estas secuencias para construir un árbol filogenético de secuencias de aminoácidos de la superfamilia ANL. El árbol definitivo agrupa 148 secuencias y fue visualizado utilizando la herramienta iTol (Letunic & Bork, 2021).

9. Resultados

9.1 Identificación de los modelos experimentales

Con el objetivo de identificar los organismos que poseen la enzima BCL se revisaron los trabajos de Gescher *et al.,* (2002), Valderrama *et al.*, (2012), Porter & Young (2014), Thornburg *et al.,* (2015) Arnold *et al.,* (2021) en los cuales se reporta de manera experimental que las especies como *T. aromatica*, *R. palustris*, *Paraburkholderia xenovorans* LB400, anteriormente *Burkholderia xenovorans* LB400 *(P. xenovorans* LB400*)* y *A. evansii* presentan a la enzima BCL en sus genomas. Estos organismos se conocen por poseer la vía de degradación del Benzoil-CoA, por lo que se consideran como organismos degradadores de hidrocarburos aromáticos.

9.2 Predicción de ortólogos de la benzoato-CoA ligasa

La identificación de los ortólogos de la BCL se llevó a cabo mediante varios pasos. El primer paso consistió en la búsqueda de homólogos realizando la construcción de la arquitectura de dominios Pfam que presentan las secuencias de la BCL provenientes de bacterias para las cuales se cuenta con evidencia experimental y que comprenden dos dominios Pfam conservados; AMP-binding (PF00501) y AMP-binding_C (PF13193) en este orden. Esta

arquitectura de proteína se usó para escanear, con el software *hmmscan*, 6536 proteomas de bacterias y arqueas completamente secuenciados obtenidos de la base de datos KEGG (Kanehisa *et al.*, 2017). Se obtuvieron un total de 232 secuencias que presentaron la arquitectura de dominio propuesta. El segundo paso se basó en la identificación de ortólogos con base en el uso del contexto genómico de los homólogos, con el objetivo de seleccionar aquellos que preservaron un contexto similar al reportado en los organismos modelo. Después de analizar el contexto genómico de los 232 homólogos obtenidos anteriormente, se observó que 135 secuencias mostraron un contexto completo ya sea aeróbico, anaeróbico o híbrido; mientras que 13 secuencias presentaron un contexto incompleto por lo que no fue posible clasificar el tipo de metabolismo correspondiente

En el tercer paso para reforzar la predicción de ortólogos de la BCL se buscó, utilizando el software MEME (Timothy *et al.*, 2015), que los motivos conservados en la clase I de las aril-CoA ligasas de la superfamilia ANL estuvieran conservados. Estos dos últimos pasos se explicarán más a detalle en las siguientes dos secciones. A partir de esto, 132 secuencias distribuidas en 118 especies mostraron la organización propuesta de motivos MEME, así como el contexto genómico completo. Estas secuencias fueron categorizadas como FC_FM (full context, full motifs) (Tabla A2, anexo). Por otro lado, las 13 secuencias que conservaban la organización de motivos MEME, pero no contenían alguna de las subunidades *boxA* o *boxB*, es decir, presentaban un contexto incompleto (Tabla A2, anexo), se categorizaron como IC_FM (incomplete context, full motifs). Por último, 3 secuencias que presentaban el contexto genómico completo, pero mostraban alguna duplicación en los motivos MEME predichos, se categorizaron como FC_DM (full context, duplicate motifs).

Sin embargo, una de las limitaciones del método utilizado es que, en el caso donde se presentan dos secuencias de la BCL codificadas dentro del mismo genoma, cuyo contexto genómico es igual al de los modelos y que comparten los motivos encontrados por MEME, nuestra propuesta no logra discernir cuál de los homólogos puede definirse como la secuencia ortóloga. Este es el caso de las secuencias parálogas encontradas en algunos representantes de los géneros *Cupriavidus*, *Aromatoleum*, *Thauera*, y *Ralstonia* (Tabla A3, anexo) de la categoría FC_FM, donde 6 secuencias presentan una copia con un contexto aeróbico y otra con un contexto anaeróbico. Por otro lado, se observó un segundo grupo de parálogos pertenecientes a la categoría IC_FM. En estos casos se consideraron como ortólogos de la BCL, a los genes que preservan el contexto genómico más completo; cup-BKK80_29555, cuu-BKK79_35510 y ccup-BKK81_26760, almacenados en los genomas de *Cupriavidus* sp. USMAA2-4, *Cupriavidus malaysiensis* USMAA1020 y *Cupriavidus* sp. USMAHM13, respectivamente. En la tercera categoría FC_DM, la secuencia paráloga es la del gen eba-ebA5301 contenido en el genoma de *Aromatoleum aromaticum EbN1 (A. aromaticum EbN1)*, que presenta una duplicación en el Motivo BCL-3 y la cual, en el sentido estricto de las reglas, deberíamos considerar la secuencia paráloga (Tabla 1, Figura 7). No obstante, un estudio reciente que reportó la presencia de parálogos de la BCL en *A. aromaticum EbN1*, encontró que ambas copias son funcionales y participan en la vía de degradación del benzoato (Suvorova & Gelfand 2019).

9.3 Identificación de ortólogos a partir del contexto genómico

La inspección del contexto genómico del gen que codifica para la BCL en organismos modelo mostró tres tipos de organización genómica (Figura 5). El grupo clasificado como cluster aeróbico presenta los genes que codifican para las enzimas epoxidasas, las cuales utilizan oxígeno para llevar a cabo la escisión del anillo aromático del benzoato (Figura 4B). Estos genes también son conocidos como genes *box*, los cuales están involucrados en la degradación híbrida aeróbica del benzoato, ya que la vía incorpora características de las vías aeróbica y anaeróbica (Valderrama *et al.*, 2012). El *cluster* de las enzimas box está constituido por los genes *boxA* y *boxB* que codifican para las subunidades BoxA (KO: K15511) y BoxB (KO: K15512) de la benzoil-CoA 2,3-epoxidasa, el gen *boxC* que codifica para la benzoil CoA-dihidrodiol liasa (KO: K15513), el gen *boxD* que codifica para la 3,4-deshidrogodifenil-CoA semialdehído deshidrogenasa, y el gen que codifica para el regulador transcripcional del catabolismo aeróbico/anaeróbico del benzoato perteneciente a la familia XRE (KO: K15546). Por otro lado, el *cluster* clasificado como anaeróbico presenta los genes que codifican para las subunidades de la benzoil-CoA reductasa (KOs: K04112, K04113, K04114, K04115). Mientras que el *cluster* clasificado como catabolismo híbrido presenta tanto los genes que codifican para las epoxidasas como las reductasas. Los *clusters* presentan un conjunto de genes que codifican para los transportadores ABC, los cuales son clasificados como sistemas de transporte de aminoácidos de cadena ramificada (KOs: K01995, K01996, K01997, K01998, K01999) (Tabla A1, anexo). Sin embargo, no se utilizaron para elegir los genes ortólogos, ya que están ampliamente distribuidos en los genomas bacterianos y no son parte de la vía metabólica río abajo y únicamente se utilizan en la representación de los *clusters*.

Se observó en los resultados obtenidos que la organización de estos genes en el genoma de *A. evansii* KB740, presenta dos *clusters* en ubicaciones distintas en el genoma (Tabla A2 y A3, anexo). El primero está relacionado con la degradación aeróbica del benzoato, es decir, presenta en su contexto a los genes *box* (Figura 6). Mientras que el segundo está relacionado con la degradación anaeróbica del benzoato y presenta en su contexto los genes *bcr* de la benzoil-CoA reductasa. Por otro lado, la bacteria *T. aromatica*, muestra un *cluster* que presenta dos copias de la BCL, la primera copia (tak-Tharo_1132) presenta en su contexto a los genes *box* y los transportadores ABC. Mientras que la segunda copia (tak-Tharo_1138) presenta los genes *bcr*, los cuales están situados a 179,207 pares de bases río abajo de la BCL.

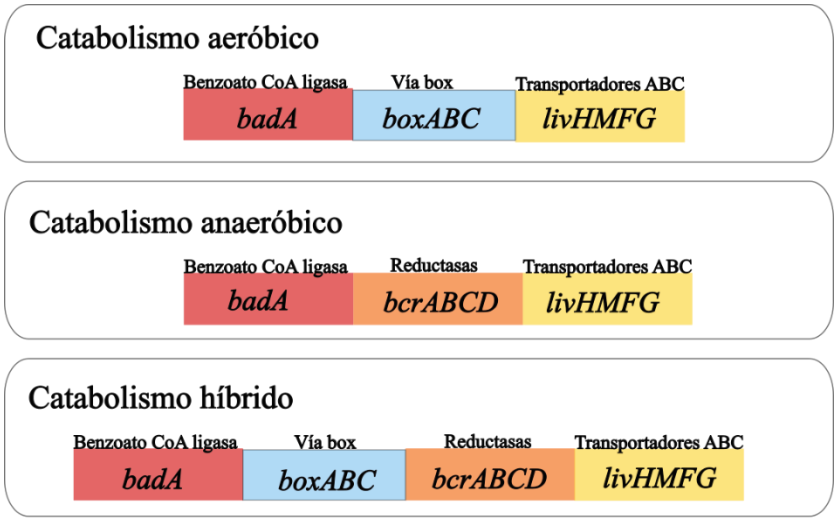


Figura 5. Representación general de *clusters* implicados en la degradación de HA. Los *clusters* fueron clasificados de acuerdo con la presencia de los genes en el contexto que codifican a las proteínas estudiadas. Es decir, el metabolismo aeróbico presenta enzimas epoxidasas, transportadores ABC y la enzima BCL. El metabolismo anaeróbico presenta enzimas reductasas, transportadores ABC y la enzima BCL. Mientras que en el metabolismo híbrido se presentan tanto epoxidasas como reductasas, transportadores ABC y la enzima BCL.

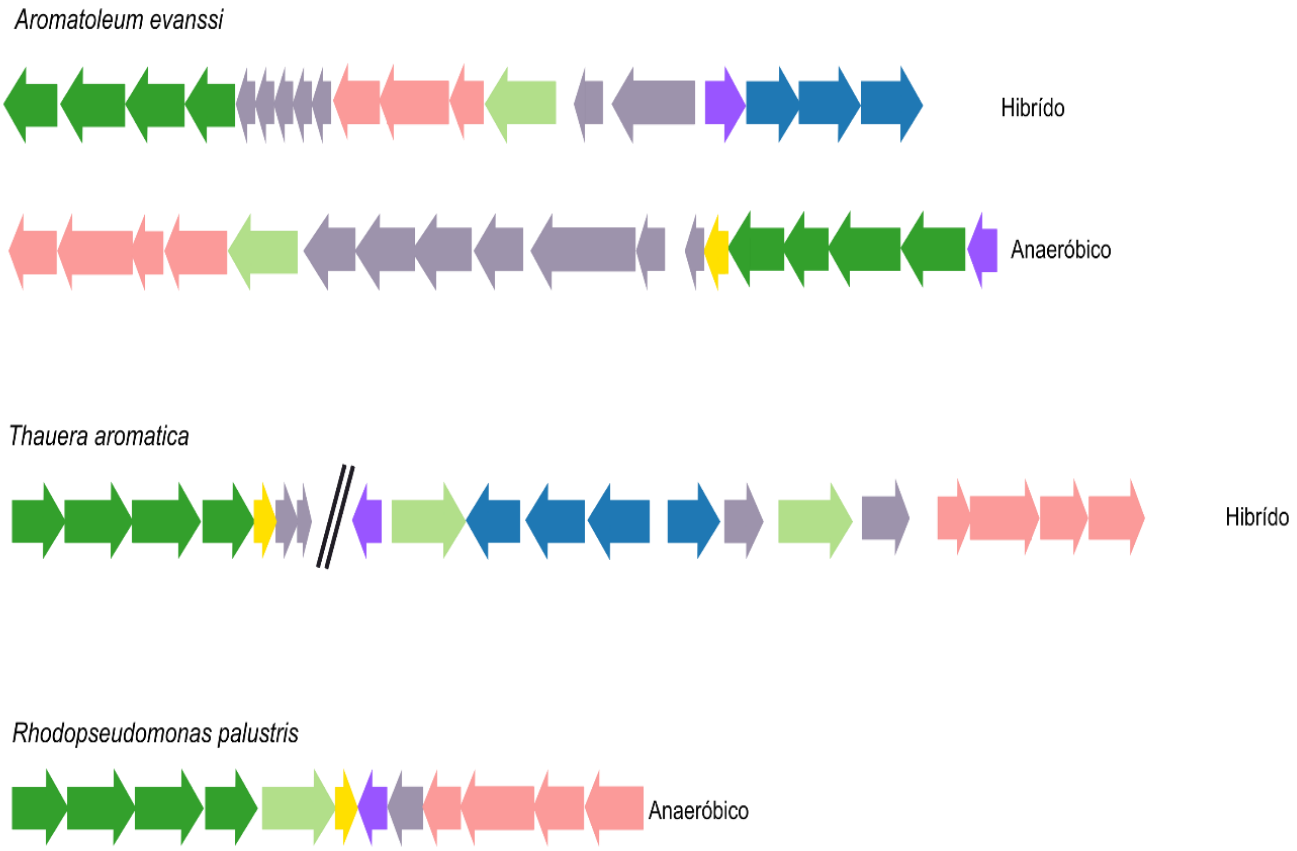


Figura 6. Organización de grupos de genes implicados en el catabolismo anaeróbico o/y híbrido de la BCL. Organización genómica de los *clusters* de la BCL encontrados en *A. evanssii*, *R. palustris* y *T. aromatica*. Los genes están representados por flechas: verde claro, genes que codifican las BCLs; verde oscuro, genes que codifican las subunidades de las reductasas para la degradación anaeróbica; amarillo, genes que codifican la ferredoxina; violeta, genes que codifican; rosa, posibles ortólogos de genes de transporte ABC; azul, genes que codifican el *cluster box* (epoxidasas) y gris, genes que codifican otras funciones conocidas o desconocidas. Dos líneas diagonales indican que los genes no son adyacentes en el genoma.

9.4 Identificación de motivos conservados en las benzoil-CoA ligasas para mejorar la búsqueda de proteínas ortólogas de la BCL

Para respaldar la selección de ortólogos predichos, se llevó a cabo un análisis de búsqueda de motivos conservados en las secuencias de aminoácidos utilizando el software MEME (Timothy *et al.*, 2015). Este análisis resultó en la identificación de ocho motivos altamente conservados. Estudios anteriores de Marahiel *et al.*, (1997), Thornburg *et al.*, (2015), Clark *et al.*, (2018) y Arnold *et al.*, (2021) han documentado la presencia de motivos MEME en secuencias de enzimas que pertenecen a la Clase I de la superfamilia ANL. Algunos de estos motivos exhiben residuos conservados y características estructurales que concuerdan con los hallazgos obtenidos en el presente estudio.

En particular, al comparar el Motivo-BCL-1 predicho en esta investigación con los trabajos previamente reportados, se revela que se encuentra en la región N-terminal y está asociado con una estructura conocida como P-loop (Tabla 1). La secuencia consenso del motivo identificado en las aril-CoA ligasas se describe a continuación: ΨΨx(S/T)(S/T/G)G(S/T)TGxPK, correspondiente a lo reportado por Arnold (*et al.,* 2021). En el Motivo-BCL-1, este consenso se encuentra completamente conservado en las posiciones Ser {181}, Ser{182}, Gly{183}, Ser{184}, Thr{185}, Gly{186}, Pro{188}, y Lys{189}, utilizando como referencia el gen rpa:RPA066 de *R. palustris* BisB (Tabla 1).

El Motivo-BCL-2, ha sido observado en otras aril-CoA ligasas y desempeña un papel fundamental en la unión de ATP y Mg^{2+}. Al analizar el consenso de este motivo (X(G/W) x(A/T)E), se destaca que los residuos Gly{327}, Ser{328}, Thr{329} y Glu{330} están completamente conservados en todas las secuencias de la enzima BCL (indicado con * en la Tabla 1). En la superfamilia ANL se ha reportado un motivo que presenta a los residuos PTIYR completamente conservados (Tabla 1), no obstante, la matriz etiquetada como Motivo-BCL-8 obtenida a partir de las secuencias no redundantes de este trabajo, muestra que las BCL solo conservan al residuo Pro {269} en el 100% de las secuencias, no siendo el caso de la Thr {270} y Val {271}, que se observan parcialmente conservadas. A diferencia de la Arg que se observa conservada en el motivo reportado para otras aril-CoA ligasas (Clark *et al.,* 2018), nuestras matrices muestran un sustitución de la Arg por un residuo de Phe en la posición 272. Por otro lado, el Motivo-BCL-3 se asemeja al motivo Rx(D/K) x6G (Thornburg *et al.,* 2015), en el que se observa que los residuos Arg {421}, Asp {423}, y Gly {430} están completamente conservados. Como se puede observar en la tabla 1, el motivo encontrado conserva también a la Lys {427}, la cual en *R. palustris* BisB5 se ha descrito que establece contacto con el borde exterior de la cavidad de unión al benzoato, facilitando su posicionamiento en el sitio activo mediante una interacción de carga entre el carboxilo del benzoato y la de la BCL. Otros residuos que encontramos 100% conservados alrededor de la Lys {427} en las secuencias analizadas, son la Asp {424}, Met {425}, y Val {428}.

El Motivo-BCL-5, localizado en la región C-terminal, conserva el residuo clave catalítico Lys {512}, que está involucrado en la reacción de adenilación que lleva a cabo la enzima, interactuando con el alfa-fosfato durante el ataque nucleofílico al ATP. Este residuo, Lys {512}, se encuentra altamente conservado en todas las aril-CoA ligasas (Arnold *et al.,* 2021).

El Motivo-BCL-4 comparte el residuo conservado Asp {406} con otras aril-CoA ligasas, mientras que el Gly {405} se conserva exclusivamente en las BCLs. Ejemplos de aril-CoA ligasas que presentan el residuo Gly {405} incluyen a las enzimas 4-hidroxibenzoato-CoA ligasa, 3-hidroxibenzoato-CoA ligasa en los organismos *T. aromatica* y *R.palustris*, así como la 3-hidroxibenzoato-CoA, aminobenzoato-CoA ligasa en *A. evansii* (Arnold *et al.,* 2021).

Por último, también se identificaron los motivos BCL-7 y BCL-6, los cuales se localizan en la región N-terminal de la proteína (Figura 7). La función de estos motivos aún no se ha determinado. Sin embargo, se observó que el motivo BCL-6 exhibe una región altamente conservada con la secuencia (FA/..YGLGN), mientras que el motivo BCL-7 presenta la conservación de un residuo de ácido aspártico, como se muestra en la Tabla 1, figura 7.

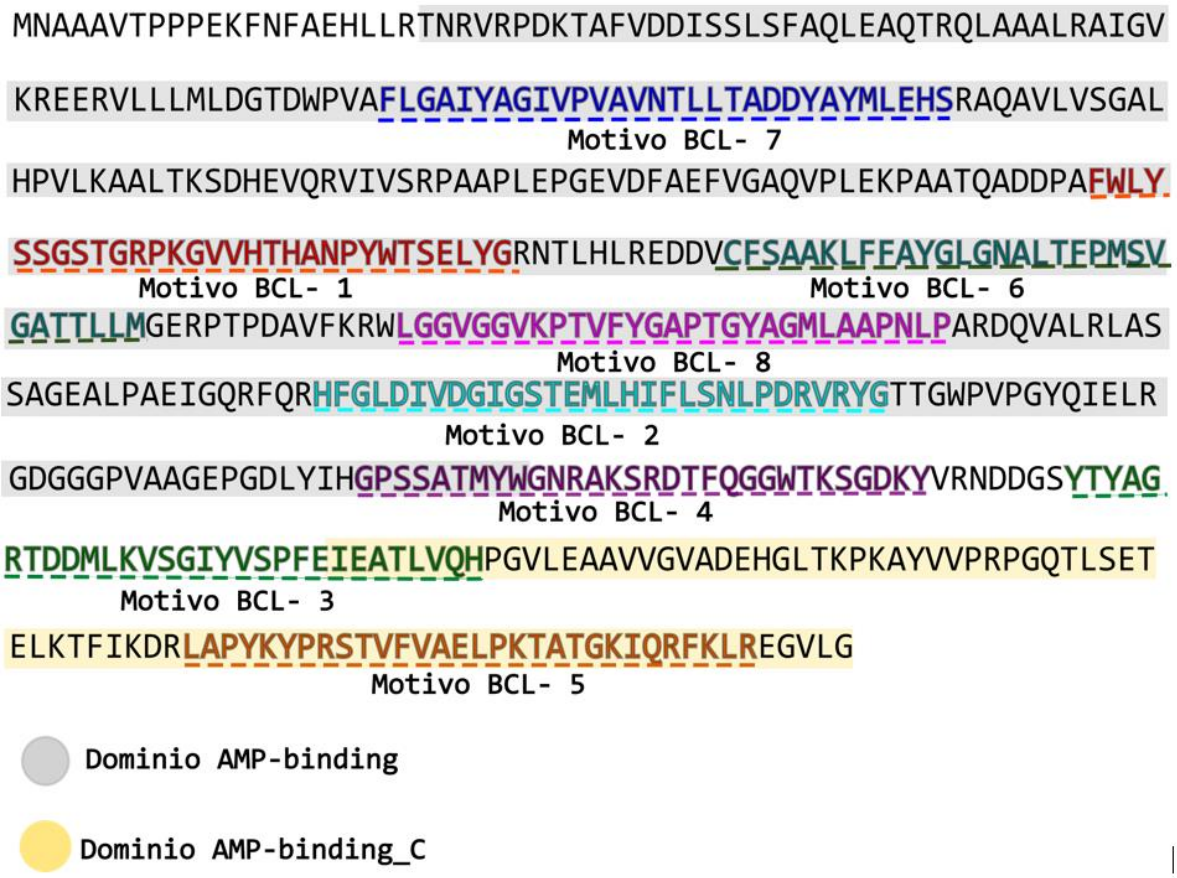
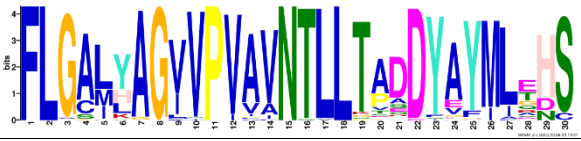
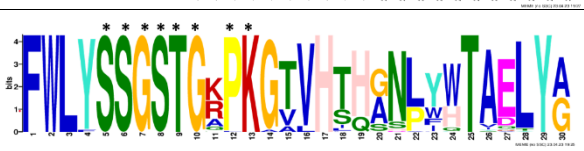
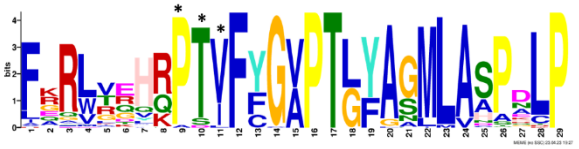
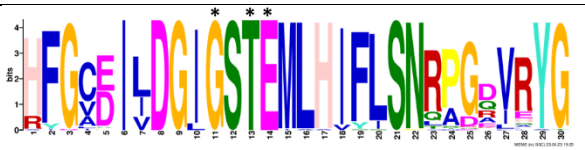
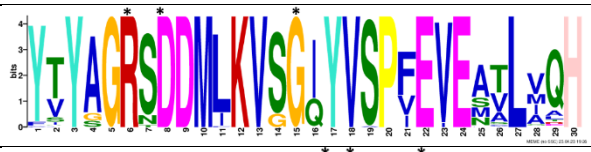
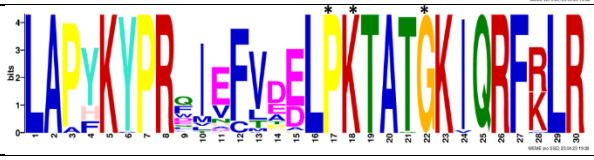


Figura 7. Motivos BCL y dominios Pfam de las BCL de *R. palustris* CGA009 (rpa-TX73_003425). Se muestra la ubicación de los ocho motivos MEME identificados en este estudio en la secuencia de la BCL de *R. palustris* CGA009, así como la longitud de los dominios Pfam, AMP-binding (área gris) y AMP-binding_C (área amarilla).

Tabla 1. Motivos conservados en los ortólogos predichos de la Benzoato-CoA ligasa. Los ortólogos predichos para la BCL presentan ocho motivos conservados. De los cuales, cinco motivos interactúan con la molécula de benzoato y el sitio activo de la enzima. Un asterisco en el logo muestra los residuos consistentes con motivos encontrados en otros trabajos (Marahiel *et al.*, 1997, Thornburg *et al.*, 2015, Clark *et al.*, 2018, Arnold *et al.*, 2021), o dichos motivos se encuentran resaltados en negrita. "Ψ", aminoácido aromático (F, Y, H o W); "Ω", aminoácido alifático (A, V, L, I o M); x, cualquier aminoácido.

Del N al C-terminal	Motivo MEME	Función	Motivos reportados (Aril-CoA-ligasas	Referencia
Motivo BCL-7		Función desconocida (propuesta del N terminal	----------	Este trabajo
Motivo BCL-1		P-loop; regula la interacción y unión del fosfato con el GDP	ΨΨx**(S/T)(S/T/G)G(S/T)***TG*x *PK*	Arnold *et al.*, 2021
Motivo BCL-6		Mutantes indican un función en el reconocimiento del sustrato	Ala227	Este trabajo
Motivo BCL-8		Pro278; Posicionamiento de ATP y unión de Mg21	(I/L)(E/Q)K(Y/E)(K/R)(V/I)Tx(L/F) xG(V/A)PTIYR(F/A)L(L/A)(K/Q)	Clark *et al.*, 2018
Motivo BCL-2		Posicionamiento de ATP y unión de Mg	**Ω(G/W)x(A/T)E**	Arnold *et al.*, 2021
Motivo BCL-4		Función desconocida	**(S/T)GD**	*Clark et al., 2018, Arnold et al., 2021*
Motivo BCL-3		Posicionamiento del ATP, unión/interacción	**Rx(D**/K)x6*G* (Lys427) de *R. palustris*	Thornburg *et al.*, 2015
Motivo BCL-5		Interactúa en la conformación de adenilación.	*P*x4*GK*Ψx(*R*/K) (Lys512)	Arnold *et al.*, 2021

9.5 Los ortólogos de la benzoato-CoA ligasa y la vía del benzoil CoA funcionan como un marcador que sugiere el tipo de requerimiento de oxígeno durante el catabolismo.

El análisis del contexto genómico demostró ser una herramienta efectiva para la identificación de posibles ortólogos de la enzima BCL y, al mismo tiempo, permitió determinar si la presencia de reductasas y/o epoxidasas en las proximidades era indicativa de reacciones subsiguientes en la catálisis del compuesto benzoil-CoA a través de una vía aeróbica, anaeróbica o híbrida. Con los resultados obtenidos se establecieron las tres categorías anteriormente mencionadas, que son FC_FM, IC_FM y FC_DM.

En la categoría FC_FM, se presentaron un total de 132 probables ortólogos de la BCL (POs), de los cuales 124 (94%) pertenecían a la clase Betaproteobacteria (Figura 8). De estos, 111 POs albergaban genes relacionados con la vía aeróbica (*box*), 7 presentaban genes asociados a las vías anaeróbicas (*brc*) y 6 poseían genes de ambas vías, caracterizado aquí como catabolismo híbrido. Adicionalmente, se detectaron 6 POs dentro de la clase Alphaproteobacteria, y todos ellos presentaban genes vinculados al catabolismo aeróbico (Figura 8 y 9). Por último, las clases Delta y Gammaproteobacteria presentaron un PO cada una, ambos con genes implicados en la vía aeróbica (Figura 8 y 9).

La categoría FC_FM engloba un total de 32 géneros, que albergan 29 proteínas distribuidas principalmente en la clase Betaproteobacteria. Dentro de esta clase, varios géneros se destacan por presentar la mayoría de los POs identificados. En particular, los géneros *Bordetella* (6, 4.55%), *Hydrogenophaga* (6, 4.55%), *Achromobacter* (8, 6.06%), *Azoarcus* (9, 6.82%), *Pandoraea* (13, 9.85%), Paraburkholderia (16, 11.35%), y *Cupriavidus* (20, 13.64%), como se muestra en la Figura 9.

El género *Thauera* presenta cuatro POs de la BCL, dos de ellos asociados con vías aeróbicas y los otros dos están asociados a vías híbridas. El género *Ralstonia* muestra dos POs, con un contexto genómico que indica una posible degradación aeróbica del benzoil CoA (Tabla A2, anexo). El género *Azoarcus*, presenta nueve POs que representan el 6,82% de las secuencias totales analizadas, exhibió diversos contextos genómicos: cinco proteínas con un contexto aeróbico, tres con un contexto anaeróbico y una con un contexto híbrido (azi-AzCIB_4632), (Tabla A2, anexo). Cabe destacar que tres genomas de *Azoarcus* (aza, azi y azd) muestran dos copias de la BCL, lo que se predijo como proteínas parálogas indistinguibles ya que todos los POs poseen uno de los contextos genómicos propuestos. Específicamente, las cepas *Azoarcus* sp. KH32C, *Azoarcus* sp. DN11 y *Azoarcus* sp. CIB han demostrado que cada una posee un PO de la enzima BCL con un contexto anaeróbico completo. Estos POs se identifican como aza-AZKH_2151, azd-CDA09_12090 y azi-AzCIB_1616, respectivamente. Sin embargo, es importante señalar que el gen azd-CDA09_22715 de *Azoarcus* sp. DN11 tiene en su contexto a los genes *box* y presenta únicamente las subunidades *bcrB* (K04112) y *bcrC* (K04113) del *cluster* anaeróbico. En contraste, el gen azi-AzCIB_4632 de *Azoarcus* sp. CIB exhibe tanto los genes aeróbicos como una copia del regulador transcripcional *boxR*, junto con un conjunto completo de genes relacionados con la degradación anaerobia (Tabla A2 del anexo). También se observaron parálogos indistinguibles de la BCL en los genomas de ciertas bacterias de los géneros *Cupriavidus*, *Thauera*, *Ralstonia* y *Aromatoleum* (Tabla A3, anexo).

Los dos grupos que no se adhieren a las reglas propuestas, IC_FM y FC_DM, presentan un total de 13 y 3 genes, respectivamente (Figuras 8 y 9). En la categoría IC_FM, se encuentran los posibles ortólogos que poseen un contexto genético menos conservado, pero aún presentan todos los motivos esenciales. En esta categoría, se destaca la presencia de la cepa *Ferrovibrio terrae* K5, que pertenece a la clase Alphaproteobacteria. Además, se han identificado 12 bacterias pertenecientes a la clase Betaproteobacteria, distribuidas en los géneros *Cupriavidus, Pandoraea, Paraburkholderia, Polaromonas, Ramlibacter y Thauera* (Figura 9, Tabla A2, anexo). Es relevante señalar que el análisis del contexto genético de los genes en esta categoría ha revelado la falta de algunos genes *box.*

Por otro lado, en las cepas *Cupriavidus malaysiensis* USMAA1020, *Cupriavidus* sp. USMAA2-4 y *Cupriavidus* sp. USMAHM13, se ha identificado únicamente la presencia del regulador transcripcional *bdzR/boxR* en su contexto (Tabla A2, anexo). En contraste, los genes encontrados en la categoría FC_FM, específicamente cup-BKK80_29555, cuu-BKK79_35510 y ccup-BKK81_26760 en *Cupriavidus malaysiensis* USMAA1020, *Cupriavidus* sp. USMAA2-4 y *Cupriavidus* sp. USMAHM13, respectivamente, muestran la presencia de los genes relacionados con la vía aeróbica completa. Por lo tanto, estas proteínas han sido clasificadas como ortólogas de la enzima BCL.

Las secuencias pertenecientes a la categoría FC_DM exhibieron un contexto genómico completamente conservado, pero presentaron motivos MEME duplicados (Tabla A8, anexo). Dentro de este grupo, destaca *Aromatoleum aromaticum* EbN1, que presenta dos copias funcionales de la enzima BCL (Arnold *et al*., 2021). Una de estas copias, eba-ebA2757, mostró una duplicación del Motivo BCL-8 en la región N-terminal para la unión de ATP y Mg^{2+}, aunque este motivo presentó una conservación deficiente, la prolina catalítica Pro {269}, se conserva en esta secuencia. Por otro lado, el gen shd-SUTH_01659 de *Sulfuritalea hydrogenivorans* exhibió una una duplicación del Motivo-BCL-2. Por último, snn-EWH46_00425 de *Sphaerotilus natans* subsp. *sulfidivorans* D-507 mostró una duplicación del Motivo BCL-3 en la proximidad de su región N-terminal, (Tabla A8, anexo).

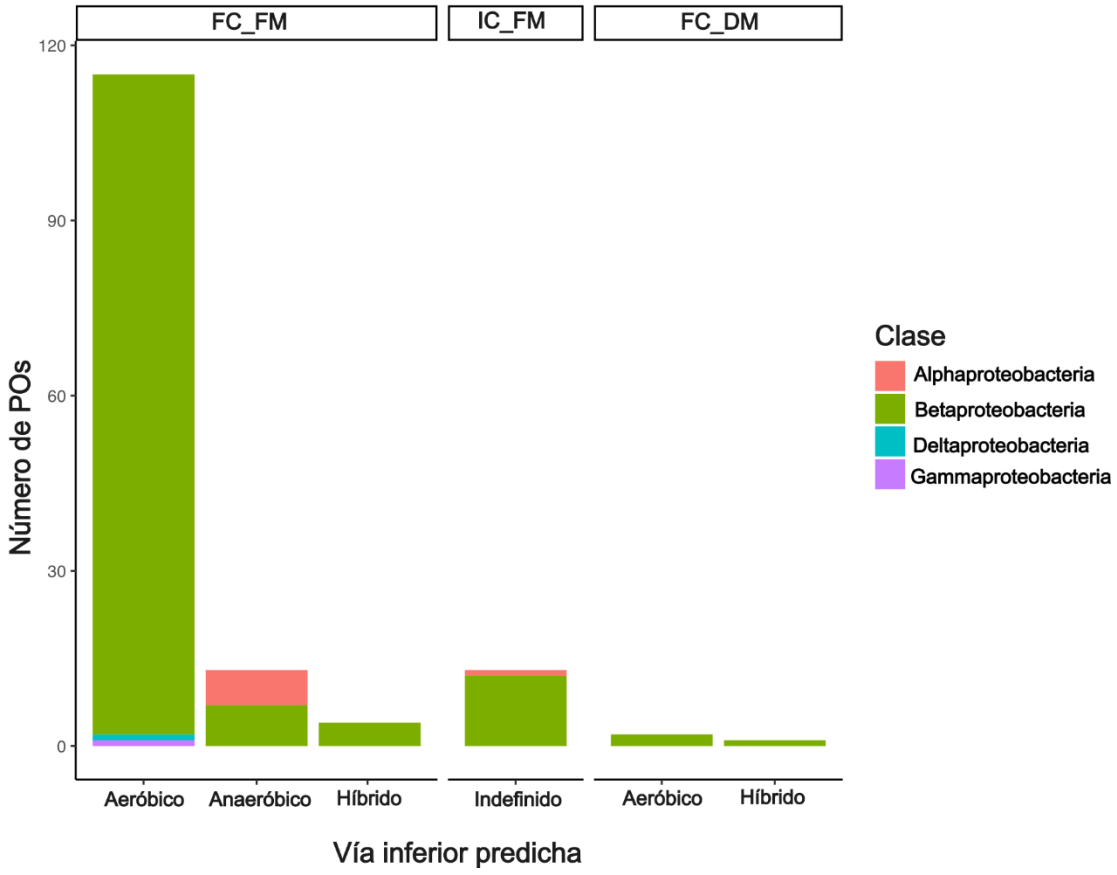


Figura 8. Organización de los POs de la BCL y vías río abajo predichas en clases. Se observa el número de POs de la BCL predichos y su distribución en cada clase filogenética dentro de las categorías consideradas en este estudio. Basado en la predicción de ortólogos de las vías río abajo. En la categoría FC_FM se observa la presencia de los pres tipos de catabolismos. Para la categoría IC_FM no se puede definir un tipo de catabolismo. En la categoría FC_DM se muestran únicamente los catabolismos aeróbico e híbrido. FC_FM (full context_full motifs), IC_FM (incomplete context_full motifs) y FC_DM (full context_ double motifs).

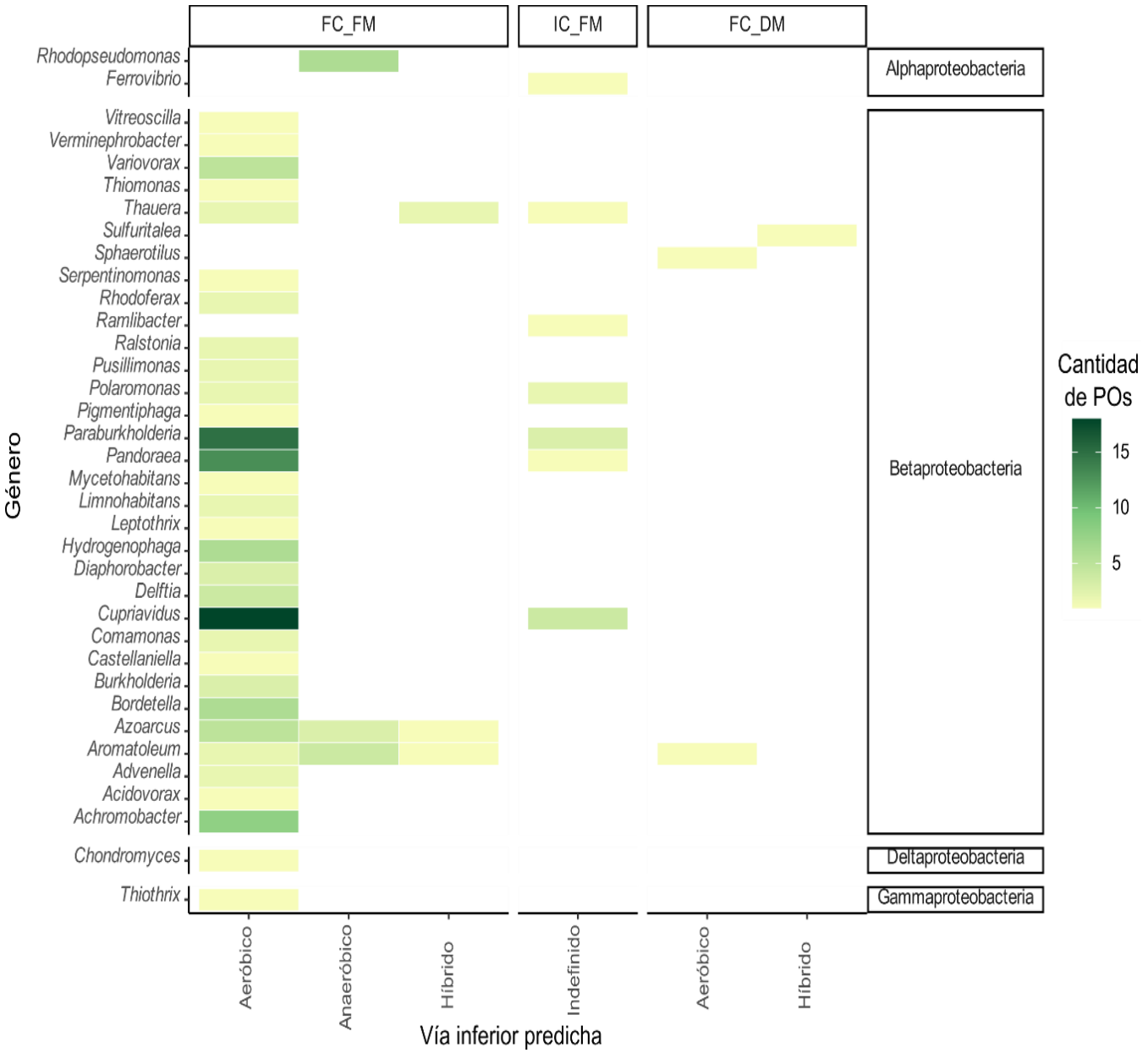


Figura 9. Distribución a nivel género de los POs de la BCL y vías río abajo predichas para cada categoría. Representación de la distribución y abundancia de las proteínas implicadas en las vías aerobias, anaerobias e híbridas de la degradación del benzoil-CoA dentro de las categorías establecidas. FC_FM (full context_full motifs), IC_FM (incomplete context_full motifs) y FC_DM (full context_ double motifs)

9.6 Distribución filogenética de los POs con metabolismo aeróbico, anaeróbico e híbrido.

Dada la dificultad para distinguir entre ortólogos y parálogos en ciertas especies, se optó por la construcción de un árbol filogenético con el propósito de evaluar si los parálogos se agrupaban en clados separados. Para llevar a cabo este análisis, se emplearon las secuencias pertenecientes a las tres categorías previamente mencionadas (FC_FM, IC_FM y FC_DM), junto con un grupo externo compuesto por otras aril-CoA ligasas de la clase I, subclase Ib, tal como se definió en el estudio de Arnold (*et al.*, 2021). Los miembros de estas categorías se organizaron en grupos de familias, y se observó que las familias Burkholderiaceae y Comamonadaceae se destacaron como las más abundantes en este análisis filogenético. El árbol fue dividido en 12 clados, (Figura 10).

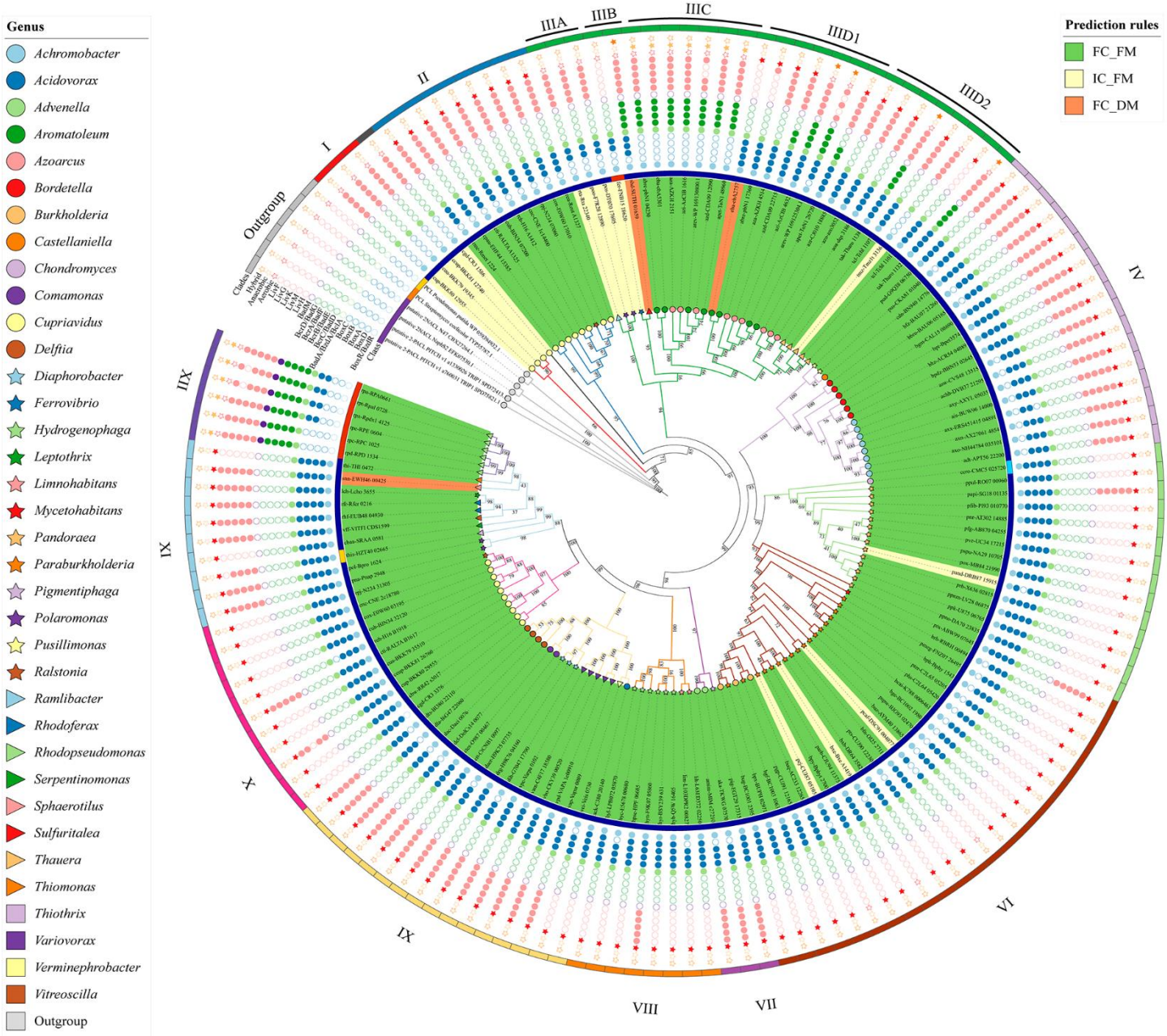


Figura 10. Distribución filogenética de ortólogos y parálogos de la BCL. El árbol se divide en doce clados marcados con números romanos. Los perfiles filéticos (puntos) que describen a los fenotipos (estrellas) rodean el árbol. Las BCLs dentro de los géneros se indican mediante figuras geométricas en cada punta de las ramas. La barra al lado de los nombres de los genes codificando a las BCL indica las clases filogenéticas Beta, Alfa, Gamma y Deltaproteobacteria. También se muestran los BCLs agrupados en FC_FM, IC_FM y FC_DM.

9.7 Distribución filogenética de los parálogos

La clasificación basada en el contexto genómico y la presencia de motivos reveló una separación notoria del grupo FC_FM, que comprende 16 proteínas parálogas procedentes de diversos géneros, como *Aromatoleum, Azoarcus, Cupriavidus, Paraburkholderia, Ralstonia* y *Thauer*a, (Tabla A3, anexo), (Figura 10).

El árbol filogenético propuesto muestra que las proteínas de la categoría IC_FM, pertenecientes al género *Cupriavidus* (cuu-BKK79_19345, cup-BKK80_12955 y ccup-BKK81_12740), se ubicaban cerca de la raíz del árbol, sugiriendo un origen ancestral. Estas proteínas exhiben en su perfil filético los genes que codifican para el regulador transcripcional *boxR/bzdR* y para el grupo de transportadores ABC. Adicionalmente, se identificó otro grupo de estas copias, posicionadas cerca de la punta del árbol en el clado X, compuesto por cuu-BKK79_35510, ccup-BKK81_26760 y cup-BKK80_29555, debido a la presencia del *cluster* de proteínas Box. Es relevante mencionar que los genes que codifican estas copias de la BCL se encuentran en cromosomas secundarios más pequeños, conocidos como crómidos putativos (Dicenzo *et al.,* 2019), (Figura 11, Tabla A3, anexo).

Otras copias pertenecientes a los géneros *Cupriavidus* y *Ralstonia*, clasificadas como FC_FM, también se localizan en un crómido putativo (Figura 11). La especie *P. xenovorans* LB400 presenta dos copias de la BCL, una ubicada en cromosoma principal (bxe-Bxe_A1419) y la siguiente ubicada en el crómido putativo. En el caso del grupo IC_FM, una de las copias se sitúa adyacente a la raíz del árbol filogenético (clado II), mientras que las segundas copias se encuentran en el clado X, en las ramas exteriores, (Figura 10). Las proteínas del clado II muestran un perfil filético que sugiere su participación en el catabolismo aeróbico, y esta característica se repite en el perfil observado en el grupo X (IIA).

Por otro lado, los clados IIIC y IIID1 albergan un segundo conjunto de parálogos, en el cual las copias se encuentran en un solo cromosoma. Ambos clados engloban bacterias de los géneros *Aromatoleum* y *Azoarcus* (Figura 11). Los POs de las BCLs agrupadas en el clado IIIC presentan un perfil filético compuesto por un conjunto completo de reductasas y un grupo de transportadores ABC, lo que sugiere que estas enzimas están involucradas en la transformación de benzoil CoA mediante el catabolismo anaeróbico.

Por otro lado, el clado IIID1 incluye dos copias (eba-ebA2757 y abre-pbN1_17360) capaces de metabolizar el benzoil CoA a través de la vía aeróbica (*box*). Es importante destacar que el PO eba-ebA2757 exhibe una duplicación del Motivo-BCL-8 en su región N-terminal. El perfil filético en las copias restantes del clado IIID1 sugiere un metabolismo híbrido en el que se encuentran presentes los genes de las proteína Box, y las subunidades de la benzoil CoA reductasa. Esto se refleja en los perfiles de azi-AzCIB_1616 y are-WP_169125304. Por otro lado, el perfil filético de azd-CDA09_22715 y apet-ToN1_26720 carece de las

subunidades D y A de la benzoil-CoA reductasa codificadas en los genes *bcrD/badG* y *bcrA/badF*, lo que sugiere un metabolismo aeróbico. Las copias de la BCL encontradas en *Thauera* se dividen en dos subclados, una de las copias muestra un perfil filético híbrido y la otra un perfil filético aeróbico, ya que ambas comparten los genes *box* en su contexto genómico, (Figura 10, Tabla A2 anexo).

El análisis del porcentaje de GC en los genomas que albergan copias parálogas de la BCL (Figura 12) revela que los POs ubicados en el cromosoma tienden a mostrar una mayor similitud en el contenido de GC en comparación con aquellos que residen en crómidos putativos. Sin embargo, existen excepciones a esta tendencia, como las copias cti-RALTA_A1325 (ubicada en el cromosoma) y cti-RALTA_B1617 (ubicada en el crómido putativo) de *Cupriavidus taiwanensis*, que presentan un contenido de GC similar. Es importante señalar que estas copias pertenecen a la categoría FC_FM, lo que significa que poseen un contexto genómico similar a los propuestos y los ocho motivos predichos, a diferencia de las copias agrupadas en el clado I del género *Cupriavidus*, que carecen de las proteínas Box (Tabla A3, anexo). Por otro lado, en el caso de especies como *Cupriavidus necator* NH9 y *Cupriavidus necator* H16 (anteriormente conocida como *Ralstonia eutropha* H16), el porcentaje de GC de los POs sin genoma segmentado se encuentra dentro del rango intercuartílico, lo que sugiere que los POs están adaptados de manera más eficiente a cada genoma del hospedero (Figura 12).

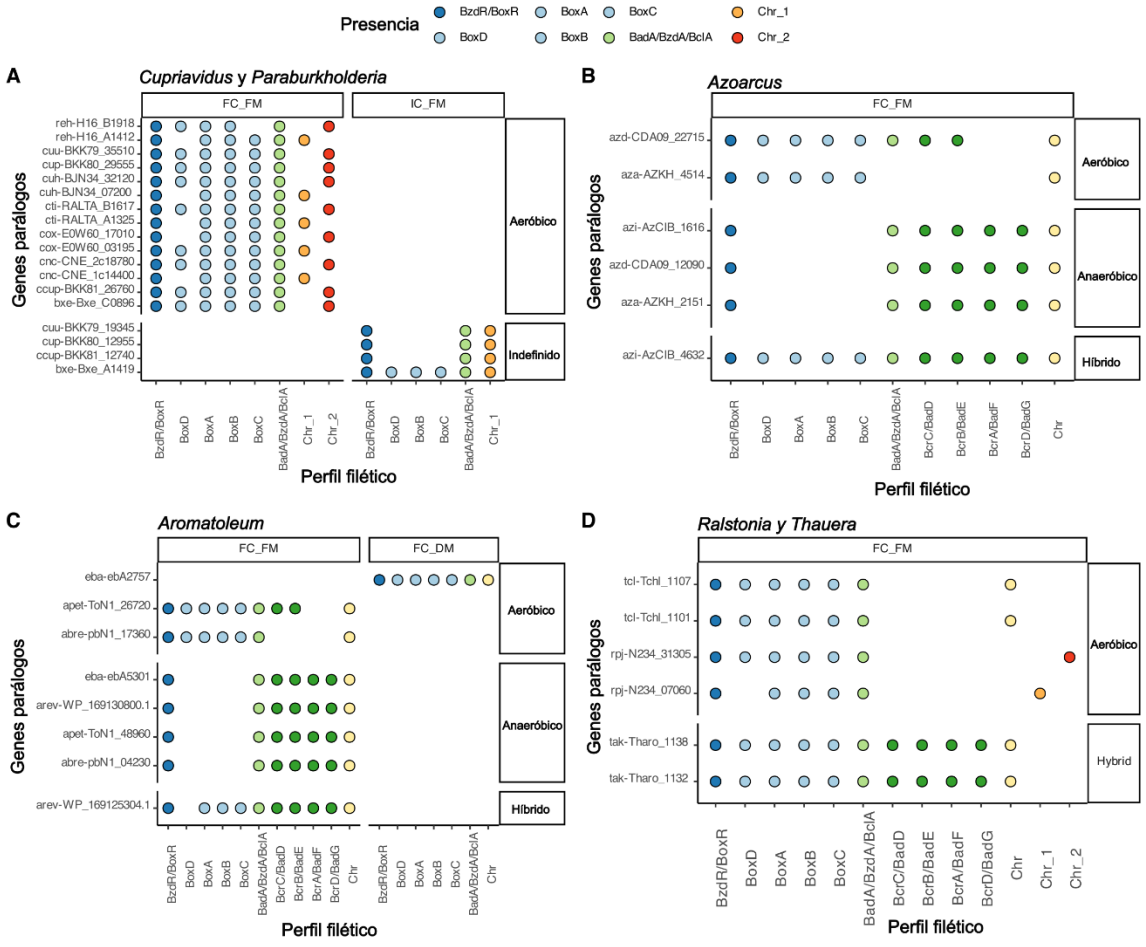


Figura 11. Perfil filético de parálogos de la BCL y sus ubicaciones dentro y fuera de genomas multipartidos. A) Distribución de las BCLs asociadas a una vía aeróbica o catabolismo indefinido en los géneros *Cupriavidus* y *Paraburkholderia*. B) Distribución de BCLs asociadas con vías aeróbicas, anaeróbicas e híbridas en el género *Azoarcus*. C) Distribución de BCLs asociadas con catabolismo aeróbico, anaeróbico e híbrido en el género *Aromatoleum* D) Distribución de BCLs asociadas con una vía aeróbica o híbrida río abajo en los géneros *Ralstonia y Thauera*.

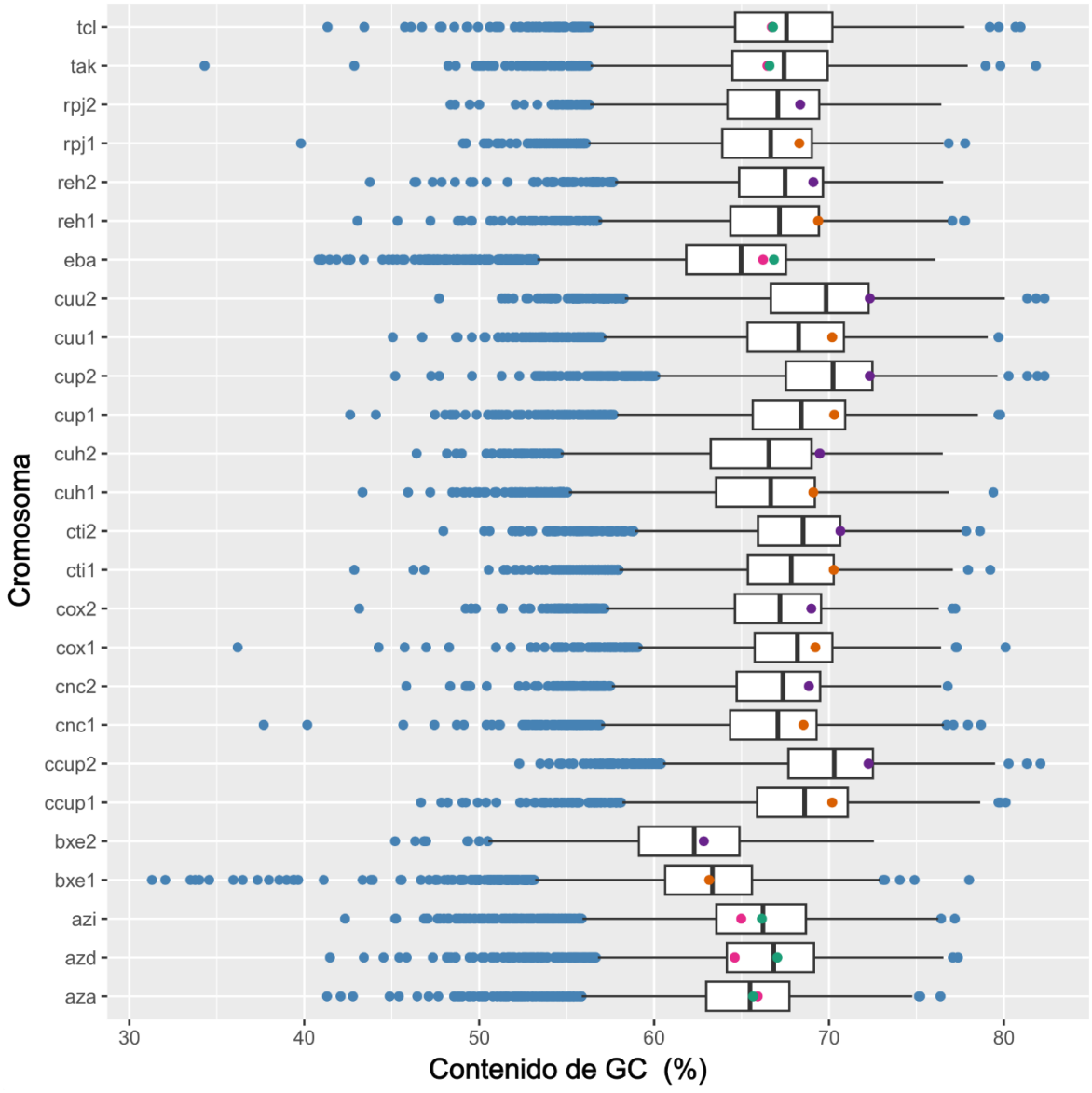


Figura 12. Distribución del %GC en copias paralelas. Se mostró la distribución del %GC de cada gen, destacando la posición del BCL %GC en la distribución. Los puntos anaranjados representan genes en el cromosoma principal, el púrpura representa genes en los crómidos putativos, y las BCLs en puntos verdes están presentes en genomas con un cromosoma.

9.8 Distribución filogenética de la BCL en genomas con una sola copia del gen.

De las 132 proteínas clasificadas en la categoría FC_FM, 101 se codifican en una única copia del gen de la BCL, lo que sugiere que estas proteínas son ortólogas de la BCL según las predicciones obtenidas en esta investigación. Los clados II y III agrupan principalmente proteínas parálogas (Figura 10). En particular, el Clado II está compuesto por tres proteínas pertenecientes al género *Cupriavidus* (cpau-EHF44_15385; *Cupriavidus pauculus* FDAARGOS_614, rme-Rmet_1224; *Cupriavidus metallidurans* CH34 y reu-Reut_A1327; *Cupriavidus pinatubonensis* JMP134), cuyos genes se ubican en el cromosoma I y todas ellas presentan un perfil filogenético aeróbico.

Los ortólogos ubicados en el clado IIID2 pertenecen al género *Azoarcus* y forman un subclado pequeño que guarda una estrecha relación con los POs del género *Thauera*. Dos de estos ortólogos son de la misma especie (aoa-dqs_3186; *Azoarcus olearius* azo-azo3052; *Azoarcus olearius* BH72), mientras que la tercera proteína, azr-CJ010_18885, pertenece a la especie *Azoarcus* sp. DD4. Los tres ortólogos predichos exhiben un perfil filético de tipo aeróbico.

Los ortólogos pertenecientes a la familia Alcaligenaceae y agrupados en el clado IV comprenden proteínas de los géneros *Achromobacter* (8 ortólogos), *Bordetella* (6 ortólogos), *Pusillimonas* (2 ortólogos) y *Castellaniella* (1 ortólogo). Cada género presenta un perfil filético característico (Figura 10, Tabla A2, anexo). Es importante señalar que la proteína axx-ERS451415 de *Achromobacter xylosoxidans* NCTC10807 carece de las subunidades C y D del *cluster* Box, mientras que la proteína pus-CKA81_01040 de *Pusillimonas thiosulfatoxidans* carece de la subunidad D del *cluster* Box. En consecuencia, el perfil filético de estas especies no sigue el patrón observado en los representantes típicos de sus respectivos géneros. Sin embargo, es destacable la ausencia de la subunidad D en pus-CKA81_01040 que es una característica compartida con las proteínas predichas del género *Achromobacter*, (Figura 10, Tabla A2, anexo).

El primer grupo de ortólogos pertenecientes a la familia Burkholderiaceae se ha agrupado en el clado V junto con un representante del género *Chondromyces* y doce del género *Pandoraea* (Figura 10). En ambos géneros, se encontró un *cluster* de proteínas Box completo en sus perfiles filéticos. Sin embargo, la treceava proteína en este clado (pand-DRB87 15915; *Pandoraea* sp. XY-2) carece de la subunidad A de la 2,3-epoxidasa benzoil-CoA (BoxA), por lo que no se considera como ortóloga. Cabe destacar que este clado incluye el único ortólogo predicho de la clase Deltaproteobacteria, que pertenece al género *Chondromyces*: *Chondromyces crocatus* Cm c5.

Un segundo grupo de ortólogos encontrados en la familia Burkholderiaceae se agrupa en el clado VI. Este clado está representado por miembros de los géneros *Paraburkholderia* (18 ortólogos), *Burkholderia* (3 ortólogos), *Mycetohabitans* (1 ortólogo, brh-RBRH 00494;

Mycetohabitans rhizoxinica HKI 454, anteriormente *Burkholderia rhizoxinica* HKI 454) y *Pandoraea* (ptx-ABW99 07645; 1 ortólogo). La proteína ptx-ABW99 07645, al igual que el resto de las proteínas en el clado VI, comparte un ancestro común con las proteínas del clado V, (Figura 10). Sin embargo, en el clado VI se identificaron tres proteínas no ortólogas (bxe-Bxe_A1419; *P. xenovorans* LB400, pcj-CUJ87_05105; *Paraburkholderia caledonica* PHRS4 y pcaf-DSC91_004077; *Paraburkholderia caffeinilytica* CF1). En los genomas de *P. xenovorans* LB400 y *Paraburkholderia caledonica* PHRS4, no se encontraron ortólogos de las subunidades benzoil-CoA 2,3-epoxidasa (BoxA y BoxB), respectivamente. Por otro lado, el ortólogo de *Paraburkholderia caffeinilytica* CF1 careció de ambas subunidades de la benzoil-CoA 2,3-epoxidasa. Los clados V y VI conservaron el *cluster* Box, y el 99,05% de las especies carecen del *cluster* de los transportadores ABC en el contexto genómico.

El clado VII está compuesto por tres miembros de la familia Alcaligenaceae (Tabla A2, anexo), con dos bacterias clasificadas en el género *Advenella* y una del género *Pigmentiphaga*. Todas estas bacterias presentaron un perfil filético aeróbico en el que se observa la presencia del *cluster* de transportadores ABC.

El clado VIII agrupa a ocho representantes de la familia Comamonadaceae, incluyendo organismos clasificados en los géneros *Hydrogenophaga* (6 ortólogos) y *Limnohabitans* (2 ortólogos), todos con un perfil filogenético aeróbico. El ortólogo de *Hydrogenophaga* sp. BPS33, hyn-F9K07_05860, mostró un *cluster* Box completo y un *cluster* que incluye a los transportadores ABC en el contexto genómico. Sin embargo, los ortólogos hyc-E5678_00680 y lim-L103DPR2_00827, encontrados en *Hydrogenophaga* sp. PAMC20947 y *Limnohabitans* sp. 103DPR2, respectivamente, carecen de la subunidad BoxD del *cluster* Box, (Figura 10, Tabla A2, anexo).

La familia Comamonadaceae comprende bacterias de los géneros *Variovorax* (5), *Delftia* (4), *Diaphorobacter* (3), *Comamonas* (2), *Acidovorax* (1) y *Verminephrobacter* (1), ubicados en el clado IX. Al igual que en otros clados, los miembros de cada género muestran un perfil filético aeróbico. Este perfil se conserva a nivel de género, con la excepción de la proteína drg-H9K76_0416 de *Diaphorobacter ruginosibacter* DSM_27467, que presenta un *cluster* Box completo. Por otro lado, los ortólogos daer-H9K75_07735 (*Diaphorobacter aerolatus* KACC_16536) y dih-07735G7047_17790 (*Diaphorobacter* sp. HDW4A) carecen de la subunidad BoxD del *cluster* Box (Figura 10).

Los clados XI y XII comparten un ancestro común reciente, aunque fueron separados en dos clados para facilitar la explicación. El clado XII alberga ortólogos de la BCL pertenecientes a la clase Alphaproteobacteria. En cambio, el clado XI incluye cinco especies de la familia Comamonadaceae, un ortólogo predicho en la familia Thiotrichaceae y otro en la familia Neisseriaceae, de la clase Betaproteobacteria. También se encuentran dos ortólogos de la BCL no asociados a una familia en particular, pero pertenecientes a los géneros *Leptothrix* y *Thiomonas*. Destaca en este clado el único ortólogo relacionado con la clase Gammaproteobacteria (*Candidatus Thiothrix singaporensis* SSD2), perteneciente al género *Thiothrix*. Todas las BCLs en el clado XI muestran un perfil filético aeróbico. Además, en este grupo se identifica a *Sphaerotilus natans* subsp. sulfidivorans D-507, la tercera especie con un motivo duplicado (snn-EWH46 00425) (Figura 10, Tabla A8, anexo).

Como se observa en el árbol, muy pocas BCLs muestran un perfil filético anaeróbico. Los ortólogos de la BCL de *R. palustris* se encuentran presentes en el clado XII y exhiben esta característica. De estas *R. palustris* CGA009, es la única BCL experimentalmente caracterizada (Geissler *et al.*, 1988). Dentro del clado XII, se observan dos perfiles filéticos distintos. *R. palustris* BisB18 y *R. palustris* BisB5 muestran todas las subunidades de la benzoil-CoA reductasa. Por otro lado, los perfiles de cuatro *Rhodopseudomonas* muestran las subunidades BoxC y BoxD (Figura 10).

9.9 Distribución filogenética de las secuencias consideradas ortólogas no probables.

Se observa que los clados IIIA y IIIB agrupan tres y una secuencia, respectivamente, con un perfil filético que carece de una o ambas subunidades BoxA/BoxB de la benzoil-CoA 2,3-epoxidasa (Figura 10). Se considera que las enzimas que carecen de una subunidad no podrían catalizar la reacción correspondiente, y dado que esta reacción es la primera de la vía, se optó por un enfoque conservador, que sugiere que estas secuencias no son ortólogas. Un caso particular es el de fer-FNB15_18620, encontrado en el genoma de *Ferrovibrio terrae* K5, clasificado como Alphaproteobacteria, el cual se agrupa en el clado IIB junto a la secuencia shd-SUTH_01659 presente en la especie *Sulfuritalea hydrogenians*, clasificada como Betaproteobacteria. Esta última presenta una copia duplicada del Motivo-BCL-2 en el extremo amino-terminal de la proteína, (Figura 10, Tabla A8, anexo).

9.10 Distribución filogenética de los probables ortólogos del regulador transcripcional de la familia XRE.

En esencia, todos los clados que muestran la presencia del *cluster* Box presentan un PO del regulador transcripcional *BdzR/BoxR* (K15546), independientemente de su clasificación en las categorías FC_FM, IC_FM y FC_DM. Sin embargo, se encuentran algunas excepciones; el perfil filético de la proteína cgd-CR3_3376 de *Cupriavidus gilardii* CR3 ubicada en el clado X, es el único ortólogo de la BCL que carece de un ortólogo probable de *BdzR/BoxR*. Otras BCLs agrupadas en el clado XII, en el que todas las bacterias pertenecen al género *Rhodopseudomonas,* tienen un conjunto completo de reductasas que carece de un regulador *BdzR/BoxR*. Sin embargo, se ha documentado que el factor de transcripción BadM regula al operón *bad* involucrado en la degradación anaeróbica del benzoato en *R. palustris*. Este regulador está presente en el contexto de todas las BCLs del clado XII, (Hirakawa *et al.*, 2015). Por otro lado, cuatro *Rhodopseudomonas* tienen POs de BoxD y BoxA que no deberían ser funcionales de acuerdo con las reglas propuestas ya que carecen de la subunidad B de benzoil CoA 2,3-epoxidasa codificada por el gen *BoxB*. Estas bacterias también carecen de *BdzR/BoxR*, (Figura 10). En la Figura 10, se observa que las BCLs que se presentan en el perfil filético *BdzR/BoxR* pueden estar ubicadas dentro de los grupos FC_FM, IC_FM o FC_DM.

9.11 Las bacterias que degradan los hidrocarburos aromáticos a través de la vía del benzoil CoA no se limitan a un entorno específico.

Los organismos con evidencia experimental de presentar la vía de degradación del benzoil CoA incluyen a la especie *R. palustris* CGA009, la cual se aisló de aguas y sedimentos marinos aparentemente no contaminados, pero también de lagunas de desechos porcinos, excrementos de gusanos de tierra y agua de estanque (Larimer *et al.,* 2004). Por otro lado, los genomas restantes con ortólogos y parálogos predichos se aislaron mayoritariamente de ambientes terrestres no marinos. Concretamente, 57 de ellos se obtuvieron de muestras de agua, suelo o sedimentos aparentemente no contaminados (agua/tierra), mientras que 37 se aislaron de organismos o de sus secreciones (Figura 13A, Tabla A4, anexo). A partir de estos genomas, se identificaron diez bacterias con un PO de la BCL que fueron aisladas de sitios contaminados con hidrocarburos derivados del petróleo, mientras que otras 25 se aislaron de sitios contaminados con otros tipos de contaminantes, como metales (Figura 13A). Los géneros *Paraburkholderia, Pandoraea, Bordetella* y *Achromobacter* destacan como los más representativos.

Cabe señalar que no se encontraron especies de *Pandoraea* aisladas de sitios contaminados con hidrocarburos del petróleo. Sin embargo, *Achromobacter, Azoarcus, Paraburkholderia* y *Cupriavidus* cuentan con al menos un representante encontrado en un sitio contaminado por hidrocarburos. Estos géneros también están asociados con sitios de aislamiento reportados como no contaminados o vinculados a un organismo. De estos géneros, *Cupriavidus* es el que presenta copias de la BCL en crómidos putativos (Figura 13, Tabla A3, anexo). Se observa que un porcentaje significativo de los géneros tiene menos de 6 miembros, lo que dificulta la evaluación estadística. No obstante, las observaciones sugieren que las vías de degradación de hidrocarburos están ampliamente distribuidas entre las diferentes especies estudiadas.

Dada la preocupación por la contaminación generada por la extracción y el transporte de petróleo en el suelo y el agua, así como por derrames como el ocurrido en la plataforma Deepwater Horizon en el Golfo de México, se esperaba encontrar una mayor cantidad de organismos en muestras oceánicas contaminadas por hidrocarburos. Sin embargo, debido a los pocos ortólogos predichos de la BCL en ambientes marinos en las secuencias inspeccionadas, se decidió ampliar la búsqueda y explorar genomas ensamblados a partir de MAGs marinos para evaluar la presencia de la BCL y las vías descendentes en ecosistemas marinos, (Figura 14).

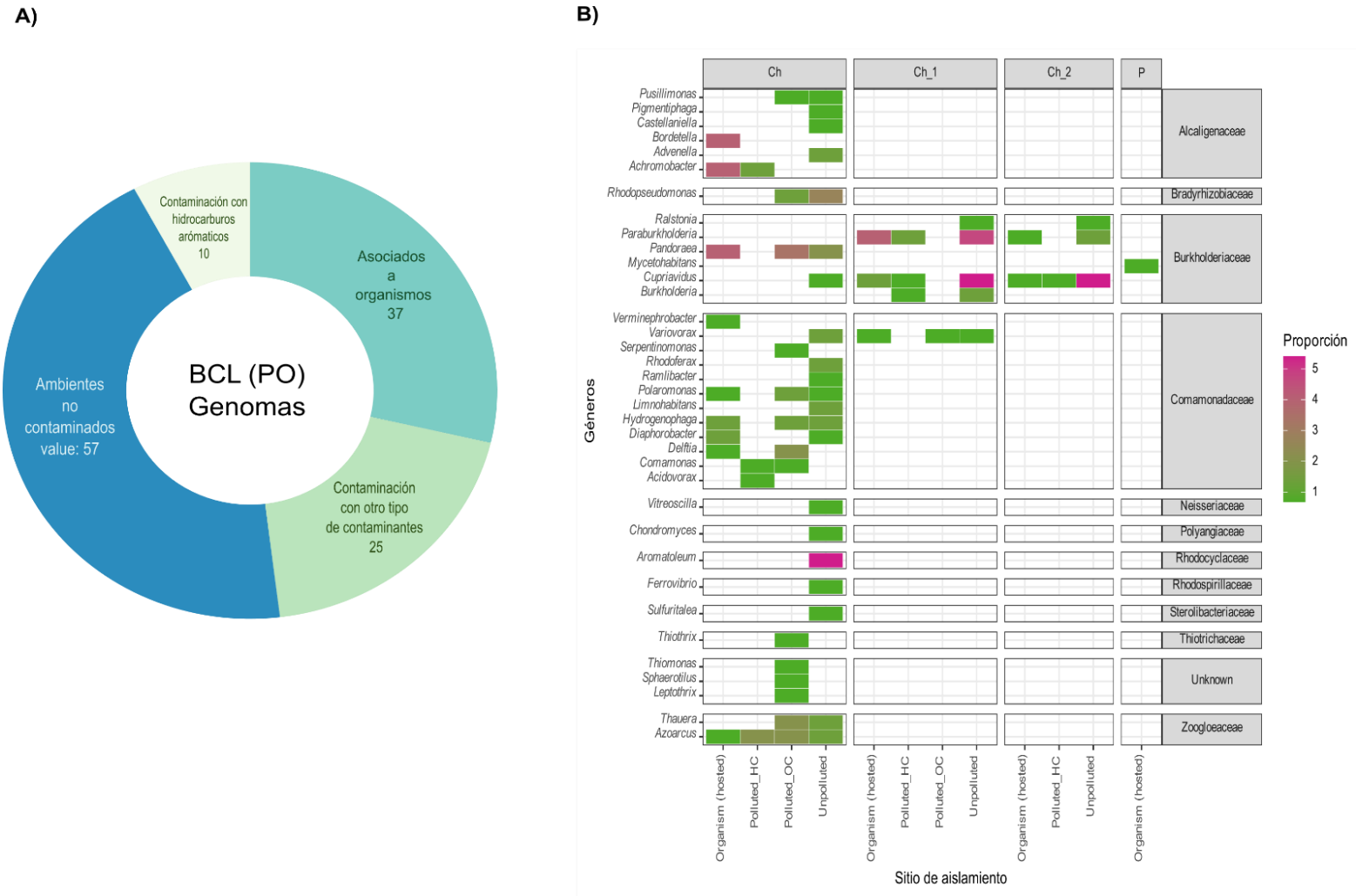


Figura 13. A) Distribución de la BCL por sitio de aislamiento. El gráfico de pastel representa bacterias con alguna BCL predicha en muestras contaminadas con hidrocarburos y otros contaminantes, así como bacterias aisladas de sitios no contaminados o asociadas con un organismo (vegetal o animal, incluidos los humanos). **B) Proporciones de las BCLs por sitio de aislamiento codificadas dentro de cromosomas y crómidos putativos.** Bacterias aisladas de diferentes ambientes agrupadas por familia y género. El mapa de calor también presenta la distribución de las BCLs en cromosomas.

9.12 Identificación de ortólogos de benzoato-CoA-ligasa en genomas secuenciados de MAGs.

Se utilizaron las matrices de arquitectura Pfam de la BCL, así como las matrices generadas por MEME, para escanear los proteomas de 1553 MAGs del catálogo OceanDNA MAG (Nishimura & Yoshizawa 2022) y en 11 binns ensamblados del Golfo de México derivados del estudio de Loza (*et al.,* 2022). En el transcurso de esta búsqueda exhaustiva, se logró identificar un total de 53 POs de la BCL únicamente del proyecto OceanDNA, como se detalla en la Tabla A6 del anexo.

Los POs de la BCL identificados, presentan en su contexto genómico al menos una de las enzimas aeróbicas o anaeróbicas propuestas, como se detalla en la Figura 13 y la Tabla A5 del anexo. En contraste con los genomas secuenciados, donde la mayoría de los POs pertenecen a la clase Betaproteobacteria, en los MAGs, la clase Alphaproteobacteria es la que presenta la mayoría de los POs, abarcando un 77.3%. Las clases Delta y Betaproteobacteria tienen proporciones iguales de POs, cada una con un 7.5%, mientras que Gammaproteobacteria contribuye con un único PO, y la clase Epsilonproteobacteria, que no se observó en genomas completamente secuenciados, presenta un PO en este conjunto (Figura 14).

Se lograron identificar siete POs en MAGs que cumplen con estándares aceptables, distribuidos entre las clases Alfa (4), Beta (2) y Gammaproteobacteria (1). Estos MAGs tienden a mantener un perfil filético bien conservado, a excepción de un par de parálogos encontrados en RII83223.1|GCA_003576595 y RII83731.1|GCA_003576595, ambos presentes en *Pusillimonas marítima (P. maritim*a), una Betaproteobacteria de la familia Alcaligenaceae. No obstante, en el proteoma de *P. maritima* se identificó un tercer parálogo con un *cluster* Box completo (RII84310.1|GCA_003576595), lo que sugiere que podría ser el ortólogo de la BCL siguiendo las reglas propuestas.

Asimismo, la proteína ALD92563.1|GCA_001281465, encontrada en el proteoma de *Cupriavidus gilardii* CR3, exhibe un *cluster* Box completo y pertenece a la familia Burkholderiaceae (Betaproteobacteria), (Figura 14, Tabla A5). Otros MAGs de alta calidad también presentan el *cluster* Box, aunque carecen de la subunidad BoxD. Estos genes fueron clasificados en las familias Roseobacteraceae, Rhodobacteraceae y Alcanivoracaceae. Al igual que en algunos de los genomas completamente secuenciados, se encontraron parálogos en los MAGs. Todos estos parálogos, excepto los de GCA_003576595 (Alcaligenaceae) y GCA_002715265 (familia no reportada ND; Alphaproteobacteria), se predijeron en la familia Rhodospirillaceae. Estas copias mostraron una conservación parcial de las proteínas Box, y en el caso de los proteomas de GCA_002725625 y GCA_002690215, se predijeron POs de las subunidades de la benzoil CoA reductasa (Figura 14).
Es importante señalar que los POs de las subunidades de la benzoil-CoA reductasa en los MAGs explorados no siempre estuvieron codificados en el contexto genómico. No obstante, en el genoma completamente secuenciado de *T. aromática*, se ha informado que los ortólogos de las subunidades de la benzoil CoA reductasa no están adyacentes a la BCL en el genoma (Carmona *et al.,* 2009).

Respecto al proteoma etiquetado como GCA_002687515, se detectaron siete POs relacionados con un contexto anaeróbico. Sin embargo, se debe considerar con precaución la interpretación de estos resultados. En siete perfiles predichos en la familia Rhodospirillaceae y en un grupo de genes encontrados en MAGs de una familia no determinada (ND), se identificó la presencia del regulador transcripcional de la familia XRE (*BzdR/BoxR*). Nuevamente, en Rhodospirillaceae, se identificaron dos POs que presentan en su contexto tanto las proteínas Box como reductasas o subunidades. En ambos proteomas (GCA_002724505 y GCA_002720855), se predijeron parálogos que sugieren un catabolismo anaeróbico del benzoil CoA. Es importante mencionar que la identificación de estos POs en los MAGs representó un desafío significativo, principalmente debido a la calidad deficiente de varios de estos MAGs, que presentaban problemas de contaminación e integridad. Esta limitación, en efecto, tuvo un impacto notorio en las predicciones realizadas, lo cual se refleja en la baja conservación del contexto genómico y, por consiguiente, en la variabilidad de los perfiles filéticos observados (Tabla A6, anexo).

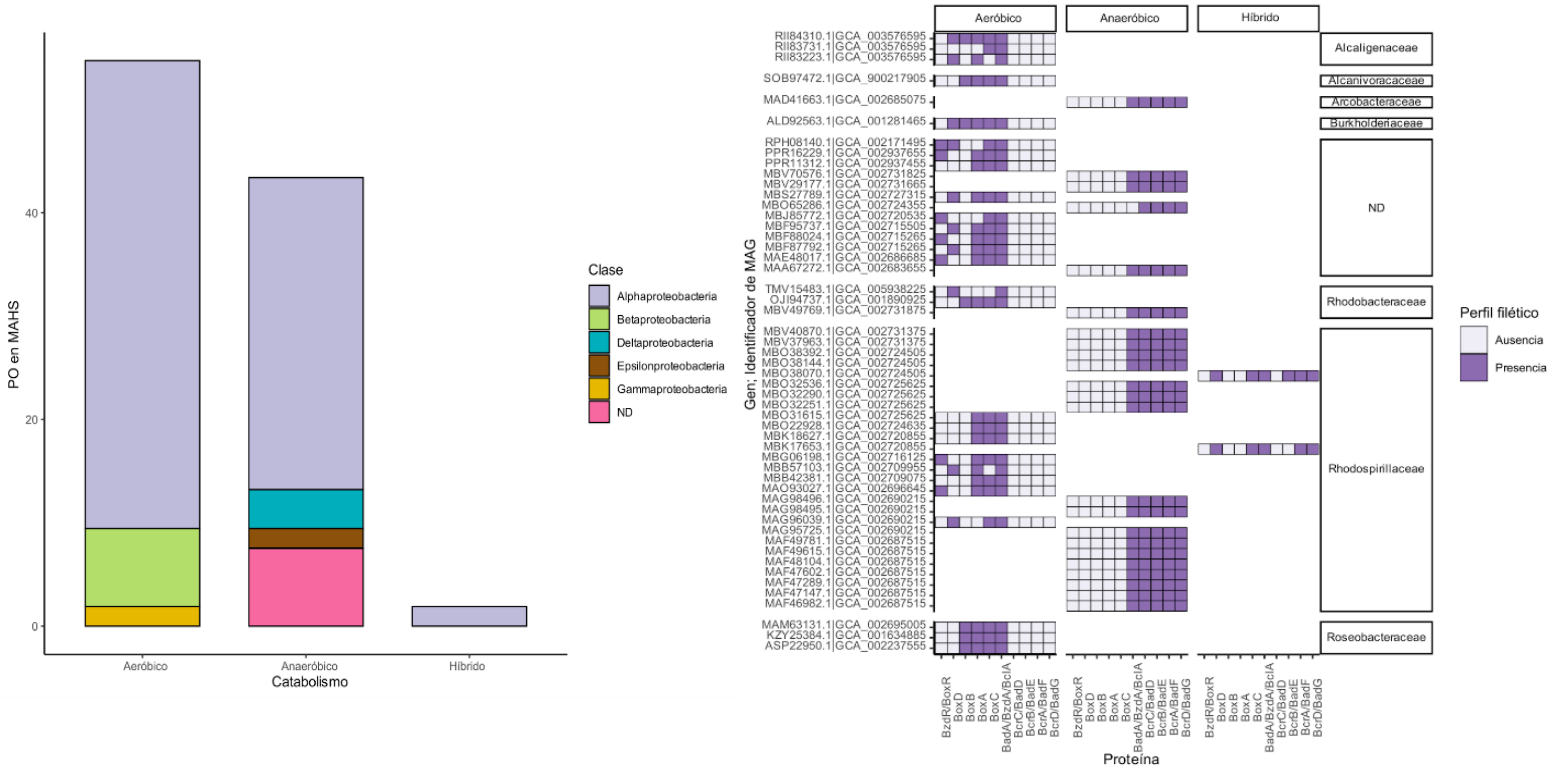


Figura 14. **Distribución de los POs identificados en genomas ensamblados de metagenomas. Panel A) Porcentaje de POs de acuerdo con la clase, así como la clasificación de acuerdo con las vías río abajo.** La mayor distribución de POs en MAGs se presenta en la clase Alphaproteobacteria. **Panel B) Representación de la presencia o ausencia de los genes implicados en las vías río abajo en los POs predichos.** Para las subunidades de la benzoil CoA reductasa, no siempre se encontraron dentro del contexto genómico.

10. Discusión

En este estudio, se investigó la distribución de la enzima BCL y sus vías metabólicas asociadas a la degradación del benzoil CoA, en un conjunto de 6536 genomas completamente secuenciados y en 1526 MAGs marinos. Las rutas metabólicas predichas se emplearon para categorizar el catabolismo en aeróbico, anaeróbico o híbrido según la presencia o ausencia de los genes involucrados en dichas vías. Para llevar a cabo este análisis, se realizó una predicción de ortólogos basada en el empleo de una serie de técnicas bioinformáticas centradas en la conservación de los dominios de proteínas y la preservación del contexto genómico. Estos métodos, utilizados por nuestro equipo y otros grupos, han demostrado ser eficaces en la predicción de marcadores genómicos para otros procesos biológicos, como los involucrados en diversas etapas del proceso de endoesporulación en Firmicutes (Davidson *et al.*, 2018; Kelly y Salgado, 2019; Martinez-Amador *et al.*, 2019; Soto-Avila *et al.,* 2021). Recientemente, esta estrategia, complementada con técnicas de aprendizaje automático, se aplicaron para identificar proteínas de la familia de enzimas formadoras de adenilato (ANL) para definir las principales clases funcionales (*Robinson et al.*, 2020). Estas clases comparten los motivos (Y/F)(G/W)X(A/T)E y (S/T)GD críticos para la unión y catálisis del ATP (Arnold *et al.,* 2021; Clark *et al.*, 2018; Marahiel *et al.,* 1997). Las clases de la superfamilia ANL predichas en el estudio de aprendizaje automático se utilizaron para construir un árbol filogenético de máxima probabilidad, revelando que el grupo de las aril CoA-ligasas estaba más estrechamente relacionado con subfamilias de proteínas distintas que con aquellos dentro de ellas.

Otro estudio reconstruyó un árbol filogenético que alineó 374 secuencias de proteínas de la superfamilia ANL, de las cuales 49 eran aril-CoA ligasas; estas proteínas conservaron cinco aminoácidos en todos los grupos: Glu328, Gly384, Asp418, Arg433, y Lys524 (Arnold *et al.,* 2021; Clark *et al*., 2018). Estos residuos se encontraron principalmente alrededor del sitio activo de las enzimas que contienen los sitios de unión a AMP y CoA en lugar de la superficie de la proteína. La reconstrucción filogenética de las 374 secuencias reveló nueve clados que exhibieron la conservación de diez motivos específicos del grupo, nueve de los cuales se pueden encontrar en la superfamilia ANL (Gulick, 2009; Marahiel *et al.,* 1997). Al igual que el estudio presentado por Robinson (*et al.,* 2020), en este trabajo enfrentamos desafíos en la clasificación precisa de las BCLs, destacando la naturaleza no trivial de la predicción de ortología en la familia de las enzimas ANL. Sin embargo, en este estudio también nos basamos en la presencia de los motivos y residuos específicos característicos de esta familia de enzimas para realizar una identificación computacional de ortólogos más precisa.

Nuestro análisis de los motivos reveló ocho motivos conservados, de los cuales dos no estaban previamente reportados. Dichos motivos denominados BCL-7 y BCL-6, se identificaron en el extremo N de los ortólogos predichos de la BCL. La inspección detallada de la composición del motivo BCL-6 reveló la presencia del residuo Ala227. En el estudio de Thornburg (*et al.,* 2015) se realizaron mutaciones, entre ellas la de Ala227Gly en la BCL (BadA) de *R. palustris*. Dicho estudio arrojó resultados significativos: la mutante Ala227Gly,

diseñada para reducir el choque estérico durante la unión del sustrato y la rotación al formar el intermediario benzoato-AMP, demostró un aumento notable en la actividad de BadA en relación con los orto-sustratos. Esto lleva a la conclusión de que el Motivo BCL-6 desempeña un papel crucial en el reconocimiento del sustrato. Por otro lado, para el Motivo BCL-7, no se encontraron procedimientos experimentales que arrojaran luz sobre su función específica. Por lo tanto, se requieren estudios adicionales para comprender a fondo la función de sus residuos. Es importante destacar que los motivos BCL-6 y BCL-7 rodean al Motivo BCL-1 en la secuencia primaria. El Motivo BCL-1, a su vez, alberga al motivo P-loop, que regula la interacción y unión del fosfato con el GDP. En el dominio N-terminal, se encuentran presentes prácticamente todos los residuos que se unen al sustrato carboxilado y al grupo adenosil del ATP. En contraste, los residuos del dominio C-terminal se encargan de coordinar los grupos ribosa y fosfato, (Thornburg *et al*., 2015). Adicionalmente, el estudio de Thornburg también destacó que Lys427 desempeña un papel esencial en la reacción de tíoesterificación, y se identificó que está completamente conservado en el Motivo BCL-3 de esta investigación. Como se mencionó previamente, los motivos restantes están relacionados en el reconocimiento y la conversión de sustratos.

La incorporación de estos motivos mejoró fuertemente el reconocimiento de las BCLs de otras aril-CoA ligasas. Sin embargo, identificamos que un pequeño grupo de genomas codifica parálogos de la BCL, algunos de los cuales presentan un contexto genómico incompleto (IC_FM) o una duplicación de un motivo en el N terminal (FC_DM). Dichos parálogos se presentaron principalmente en las familias Burkholderiaceae, Zoogloeaceae y Rhodocyclaceae. Por otro lado, aquellos parálogos de la BCL que presentaban un contexto incompleto no se asignaron inicialmente como ortólogos ya que carecían de una o más subunidades de una enzima, lo que indica que la vía de degradación del benzoil CoA no debe ser funcional. Sin embargo, ambos preservaron una copia del factor de transcripción BzdR/BadR en su contexto genómico, lo que puede asegurar su transcripción. No se ha determinado la razón por la cual ciertas bacterias conservan estas copias funcionales. Algunas hipótesis sugieren que estas vías redundantes pueden reflejar una estrategia biológica para aumentar la aptitud celular de los organismos para sobrevivir en ambientes sujetos a concentraciones de oxígeno cambiantes (Valderrama *et al.,* 2012).

Un fenómeno de gran importancia observado en los parálogos de la familia Burkholderiaceae es la presencia de un genoma segmentado en organismos de los géneros *Cupriavidu*s y *Ralstonia* (Dicenzo & Finan., 2017; Dicenzo *et al.,* 2019). Los cuales poseen un cromosoma y un crómido putativo, en donde el último contiene genes esenciales para la supervivencia del organismo y, por lo tanto, exhibe características tanto de cromosoma como de plásmido. En la familia Burkholderiaceae, se ha informado que las copias de las enzimas oxigenasas involucradas en la biodegradación de HAs se encuentran en diferentes cromosomas. Una copia reside en el cromosoma principal, mientras que la segunda se ubica en el crómido putativo (Pérez-Pantoja *et al.,* 2012). Los resultados de esta investigación sugieren que en algunos casos las subunidades de la benzoil CoA 2,3-epoxidasa (*Box*) pueden estar distribuidas en un cromosoma adicional (crómido putativo). Se han propuesto dos teorías para explicar el origen de los crómidos putativos: la hipótesis de la división (Schism), que plantea la escisión del cromosoma en dos replicones, dando como resultado un cromosoma y un crómido putativo con una distribución de genes esenciales casi idéntica. Por otro lado, la hipótesis del plásmido sugiere que el crómido putativo se formó a partir de un mega

plásmido adquirido mediante transferencia horizontal, (Dicenzo & Finan., 2017; Dicenzo *et al.*, 2019). El árbol filogenético propuesto en este estudio revela que la BCL encontrada en el cromosoma de las especies de los géneros *Cupriavidus* y *Ralstonia*, agrupados en los clados I y II, se encuentra cerca de la raíz o base del árbol, lo que sugiere que las enzimas en el cromosoma son ancestrales en comparación con las que se encuentran en el crómido putativo (clado X), (Figura 9). Se ha observado que los POs presentes en el clado X poseen un conjunto completo de proteínas Box que no se encuentran en el clado I. Estos resultados sugieren que la incorporación de la vía aeróbica río abajo es un evento posterior que podría atribuirse a la transferencia horizontal de genes. Al comparar el contenido de GC (%GC) entre la BCL en el crómido putativo y el %GC en el cromosoma principal, se observa que las BCL en el crómido putativo presentan valores atípicos, aunque no se alejan significativamente del rango intercuartílico.

En el caso particular de *P. xenovorans* LB400, se identificó la presencia de parálogos de la vía Box, donde una copia esta codificada en un megaplásmido no considerado como un crómido putativo y la segunda copia en el cromosoma. Esta configuración es única, ya que, en otras especies, dichos parálogos suelen ubicarse en un cromosoma y un crómido putativo (Bains & Boulanger 2007). La conservación de estas copias es un elemento fundamental para la adaptación a entornos caracterizados por fluctuaciones en los niveles de oxígeno, lo cual puede inducir la activación de vías catabólicas tanto aeróbicas como anaeróbicas.

Se observo que la mayoría de los POs de la BCL, que presentan, en su contexto, ortólogos de los genes *box* también incluyen al regulador transcripcional BzdR/BoxR perteneciente a la famlia XRE. De acuerdo con Valderrama (*et al.*, 2012), la presencia del gen *boxR* es una característica común en los *clusters box*. Observamos que, en el caso de los POs de la BCL pertenecientes al género *Rhodopseudomonas* agrupados en el clado XII, presentan en su contexto genes relacionados con el catabolismo anaeróbico del benzoil CoA, pero no incluyen al regulador transcripcional BzdR/BoxR. Sin embargo, Hirakawa (*et al.*, 2015) reportaron experimentalmente la regulación de la degradación anaerobia del benzoato en *R. palustris*. En su estudio, demostraron que el operón *bad*, relacionado con la benzoil CoA reductasa, está controlado por el factor de transcripción BadM. Esto concuerda con nuestros resultados, ya que el regulador *badM* está presente en el contexto de los POs pertenecientes al género *Rhodopseudomonas* (Figura 9).

Una hipótesis alternativa propone que la preservación de copias parálogas puede funcionar como una estrategia bacteriana, que pudiera permitir concentraciones variables de enzimas durante diferentes fases de crecimiento. Este fenómeno se detectó en *P. xenovorans* LB400, donde el análisis proteómico reveló que las proteínas Box son abundantes en el cromosoma durante la fase de crecimiento cuando el bifenilo y el benzoato se utilizan como fuentes de carbono. Simultáneamente, las proteínas Box en el megaplásmido se detectaron solo en presencia de benzoato durante la transición a la fase estacionaria (Denef *et al.*, 2005). Otro estudio planteó la hipótesis de que la eficiencia catalítica de la BCL sería diferente para sustratos específicos (Bains & Boulanger, 2007). Estos resultados fueron inconsistentes con los de un estudio proteómico en el que la BCL presente en el crómido putativo fue un 60% más eficiente en la degradación del benzoato que la BCL codificada en el cromosoma (Bains & Boulanger, 2007). Los autores argumentan que una posible explicación para la elevada actividad catalítica de la BCL codificada en el crómido putativo puede ser que la maquinaria

de transcripción/traducción es probablemente menos activa por lo que para compensar de una enzima más eficiente para metabolizar el mismo nivel de sustrato aromático. Sin embargo, ninguna de las hipótesis presentadas aborda completamente por qué algunos organismos tienen dos copias de estas enzimas. El análisis realizado en este trabajo reveló que la copia en el cromosoma de *P. xenovorans* LB400 pertenece al grupo IC_FM, en el que está ausente el PO de la subunidad BoxA, lo que sugiere un impedimento para la degradación posterior a través de esta vía. Considerando este resultado, suponemos que la ausencia de BoxA en el cromosoma puede imponer una presión selectiva sobre la copia codificada en el megaplásmido, que posee un *cluster* Box completo capaz de metabolizar el benzoil CoA. De esta manera, las copias en el megaplásmido podrían complementar la ausencia de BoxA en el cromosoma. El *cluster* Box completo cerca de la BCL codificada en el megaplásmido podría explicar por qué esta es más eficiente pero menos abundante que la BCL en el cromosoma, considerando que los pasos siguientes en la reacción pueden ser co-transcritos, incluso para esta vía en la que *boxABC* y *boxR* están organizados de manera divergente con respecto a *boxD* y al gen de la BCL, tal como observamos en el contexto genómico. Se ha sugerido que un aumento en la unión de la ARN polimerasa en una orientación reduciría la transcripción en la dirección opuesta en secuencias promotoras con una organización divergente (Warman *et al.,* 2021). Esta organización genómica podría indicar que el *cluster boxABC* será transcrito de manera más eficiente en condiciones específicas que el *boxD*, dejando más enzimas libres para complementar la vía codificada en el cromosoma.

Nuestros datos también mostraron que tres especies de *Cupriavidus* exhibían una BCL en el cromosoma donde la vía inferior estaba notoriamente ausente. Este hallazgo sugiere que la degradación efectiva del sustrato a través de esta vía requiere la expresión de productos Box codificados por el crómido putativo. Al igual que la de *P. xenovorans* LB400, la organización cromosómica de las BCLs en unidades separadas también indica que la eficiencia de la degradación del sustrato está significativamente influenciada por la afinidad de cada enzima y su concentración.

Un análisis del potencial de biodegradación de compuestos aromáticos en Burkholderiales, centrado en la degradación aeróbica que involucra la activación del anillo aromático mediante oxigenasas (Pérez-Pantoja *et al.,* 2012), sugiere que, en este orden, el hábitat podría desempeñar un papel más relevante que el origen filogenético en la configuración de la versatilidad catabólica aromática. No obstante, se debe abordar esta observación con precaución. Nuestro estudio, reveló que el 81% de las proteínas Box predichas, es decir, las proteínas implicadas en la degradación aeróbica identificadas están codificadas en genomas pertenecientes a Burkholderiales aislados de diversos nichos, que abarcan desde hospederos humanos, animales, rizosfera, nódulos radiculares del suelo, pasando por aguas residuales, lodos y sedimentos (Tabla A4, anexo).

Algunos organismos de la familia Burkholderiales, se han aislado de una variedad de entornos, que incluyen suelos, cuerpos de agua y áreas contaminadas con hidrocarburos derivados del petróleo, metales o aguas residuales industriales y agrícolas. Los resultados obtenidos en este estudio indican que las cepas de Burkholderiales que poseen un PO de la BCL fueron aisladas no solo de lugares con contaminación por hidrocarburos de petróleo u otros contaminantes, sino también de sitios aparentemente no contaminados, así como de organismos como plantas, animales/humanos o sus secreciones. Además, investigaciones

previas realizadas por Pérez-Pantoja (*et al.*, 2012) han revelado que las cepas de Burkholderiales aisladas de patógenos humanos y zoonóticos presentan el potencial para catalizar la biodegradación de compuestos aromáticos mediante oxigenasas. En ese mismo estudio, se identificaron múltiples oxigenasas que degradan HAs en los crómidos putativos. El análisis realizado en el presente estudio revela que las vías de degradación aeróbica relacionadas con el *cluster* Box, encontrado en géneros como *Cupriavidus, Ralstonia* y *Paraburkholderia,* también se encuentran en los crómidos putativos. Este hallazgo sugiere que la acumulación de genes asociados con la degradación de HAs responde a la adaptación a diferentes ambientes.

La distribución observada de bacterias con genomas segmentados en diversos entornos parece estar relacionada con el tamaño del genoma, lo que permite la expansión de este a través de la adquisición de nuevas funciones alojadas en el crómido putativo (Pérez-Pantoja *et al.*, 2012; Dicenzo & Finan, 2017; Dicenzo *et al.*, 2019; Riccardi *et al.*, 2023). Estudios recientes comparando el tamaño del cromosoma y del crómido putativo en genomas segmentados de Alfa, Beta y Gammaproteobacteria sugieren que los crómidos putativos están implicados en el proceso de ganancia/pérdida de genes, probablemente influenciado por la selección evolutiva en respuesta a la adaptación a diferentes hospedadores y nichos (Riccardi *et al.*, 2023; Dranenko *et al.*, 2023). En su conjunto, estos estudios, junto con las conclusiones de este análisis, sugieren que las especies de Burkholderiales muestran múltiples estrategias de degradación de hidrocarburos y una notable versatilidad para colonizar diversos nichos ambientales.

Las especies que carecen de genomas segmentados también presentan parálogos de la BCL en bacterias pertenecientes a los géneros *Azoarcus, Aromatoleum* y *Thauera*. Estas copias se encuentran organizadas en diversos subclados dentro del Clado III, lo que sugiere su adquisición en eventos evolutivos independientes (Figura 8). Aquellas copias que se hallan más cercanas a la raíz se agrupan en el clado IIIC, exhibiendo un perfil filogenético anaeróbico. Por otro lado, las copias reunidas en los clados IIID1 y IIID2 fueron adquiridas en un evento posterior, dado que algunos de sus parálogos comparten un contexto genético enriquecido con proteínas pertenecientes al *cluster* Box (como *Azoarcus* sp. DN11, *Azoarcus* sp. CIB y *T. aromática* K172), las cuales se ubican específicamente en el clado IIID2. Esta característica les otorga a estas bacterias la capacidad de adaptarse a una amplia diversidad de condiciones ambientales. Estas observaciones concuerdan con el análisis genómico previo realizado en *Azoarcus* sp. CIB y *Aromatoleum aromaticum*, donde se evidenció que estas bacterias presentan un número significativamente elevado de elementos móviles y una distribución genética aleatoria para la degradación de compuestos aromáticos tanto en condiciones anaeróbicas como aeróbicas (Fernández *et al.*, 2014). Esta propiedad muestra la alta plasticidad adaptativa de los organismos pertenecientes a los géneros *Azoarcus* y *Aromatoleum* en lo que respecta a la degradación de hidrocarburos aromáticos.

Nuestra investigación reveló que el 87% de las BCLs analizadas y sus vías inferiores estaban codificados como copias individuales en sus respectivos genomas. Este fenómeno provoca que ciertas bacterias mantengan solo una copia para degradar el benzoato y los compuestos relacionados en condiciones de bajos niveles de oxígeno o agotamiento. Nuestros hallazgos destacan una mayor prevalencia del *cluster* Box, lo que lleva a considerar la redundancia observada en la degradación aeróbica de hidrocarburos monoaromáticos dentro de las

bacterias estudiadas. Esta consideración se deriva de la distribución observada de la clásica vía de degradación aeróbica del benzoato en las bacterias, que se basa en la hidroxilación del anillo aromático para producir catecol, que posteriormente se escinde por una dioxigenasa, como se documenta en varios estudios (Lykidis *et al.*, 2010; Pérez-Pantoja *et al*., 2003, 2012). Se confirmó la presencia de genes que codifican tanto vías aeróbicas como anaeróbicas en siete cepas pertenecientes a Burkholderiales, entre ellas *P. xenovorans* LB400. Cabe destacar que las vías de benzoato/catecol y box de estas cepas exhiben una expresión diferencial en diversas condiciones fisiológicas, como durante la fase de crecimiento (Denef *et al.*, 2006).

En cuanto a los sitios de aislamiento encontrados en genomas secuenciados se encontró únicamente un organismo aislado de ambientes marinos. Un análisis de los sitios de aislamiento de los genomas depositados en la base de datos KEGG mostró que 258 de los 6536 genomas se aislaron de ambientes marinos. Para evitar este sesgo, se siguió la metodología propuesta para encontrar ortólogos en 1526 genomas ensamblados a partir de metagenomas tomados del catálogo OceanDNA MAG (Nishimura *et al.*, 2022) y en 11 binns ensamblados del Golfo de México, para los que no se cuenta con medidas del contenido de hidrocarburos (Loza *et al.*, 2022). Los MAGs pertenecientes al catálogo OceanDNA MAG presentaron diferentes niveles de completitud y contaminación, algunos de ellos de baja calidad (Tabla A6, anexo). En contraste, los del Golfo de México presentaron altas calidades (Tabla A7, anexo), (Loza *et al.*, 2022). La aplicación del método en los MAGs permitió identificar 53 POs de la BCL, los cuales tienen en sus vecindades al menos una de las enzimas aeróbicas o anaeróbicas del contexto genómico propuesto. La distribución de las enzimas contextuales se vio afectada por la calidad del genoma fragmentado. Lo que significa que los MAGs que tienen baja calidad y poca completitud presentan un contexto menos conservado. El 17.5 % de los MAGs tomados del catálogo OceanDNA se ensamblaron de agrupamientos de muestras recogidas en diferentes profundidades en la misma región geográfica. Esta aproximación ayudó a los autores a obtener genomas fragmentados a expensas de un resultado más limpio, lo que explica la poca conservación del perfil filético. Los MAGs provenientes del Golfo de México tienen alta calidad e incluyen bacterias con capacidades degradantes de hidrocarburos, como *Parvibaculum lavamentivorans*, *Pseudomonas aestusnigri*, bacterias del género *Kineosporia*, *Sulfitobacter* sp. y una *Actinobacteria*_SGB36491 no clasificada (Loza *et al.*, 2022). No obstante, la inspección de estos MAGs utilizando el método propuesta mostró que las BCLs conservaba los dominios Pfam, AMP_binding y AMP-binding_C, pero carecía de las proteínas del contexto genómico propuesto y de un conjunto completo de los motivos MEME, lo que las clasifica como homólogos de las aril CoA ligasa, pero no como ortólogos de la BCL. Este resultado, es importante, ya que indica que los métodos de anotación actuales usados masivamente pueden ser imprecisos y que es necesario mejorar los métodos de anotación usando aproximaciones como las presentadas en este trabajo.

Un resultado llamativo mostró que el 80% de los MAGs predichos corresponden a la clase Alphaproteobacteria. Sin embargo, el 82.2% de estos presentó bajas calidades. En cambio, las enzimas Box se identificaron en binns de alta calidad, clasificados como *Cupriavidus gilardii* CR3, *Pusillimonas maritima*, *Antarctobacter heliothermus*, *Sulfitobacter* sp. HI0040, *Planktotalea frisia* y *Alcanivorax xenomutans*, distribuidos en las clases Beta, Alpha y Gammaproteobacteria. Cabe destacar la predicción realizada en *Cupriavidus gilardii* CR3, que ya se ha aislado y cultivado y que apareció entre genomas completamente secuenciados,

se aisló de un depósito natural de asfalto (Wang *et al.,* 2015). El identificador de este MAG aparece dentro de los MAGs marinos, sin embargo, esto es un error en los metadatos ya que su localzación es terrestre. Finalmente, podemos decir, que la calidad de los MAGs fue un factor crítico en la detección de la BCL y las vías río abajo. Esta investigación muestra que los MAGs de alta calidad son más propensos a conservar un contexto genómico conservado. Este hallazgo enfatiza la importancia de obtener genomas completos y de alta calidad.

11. Conclusiones.

En este estudio, se emplearon varias estrategias para predecir a los ortólogos de la BCL. A partir de la metodología empleada se logró distinguir especialmente a las enzimas BCL de otras aril CoA ligasas debido a la exploración de los motivos conservados, de los cuales, dos estaban localizados en el extremo N y eran compartidos por las benzoato CoA ligasas, pero no por otras aril CoA ligasas, y uno de los motivos BCL propuestos jugó un papel en el reconocimiento del sustrato.

El contexto genómico fue una herramienta útil para la predicción de ortólogos, así como para definir la vía de degradación del benzoil. En este paso, se consideró que la presencia del cluster box y las subunidades benzoil-CoA reductasa definían una vía aeróbica, anaeróbica o híbrida. La conservación del contexto generalmente proporciona una fuerte evidencia para predecir un gen como un ortólogo. Sin embargo, en el conjunto de datos utilizado, algunos casos mostraron la presencia de dos copias de la BCL en el mismo proteoma, tales copias conservaban los contextos genómicos propuestos, mostrando que la predicción de ortólogos sigue siendo complicada. También se demostró que el uso de múltiples pasos mejora la predicción de ortología. Tal aporte ayudó en la exploración de genomas ensamblados a partir de metagenomas, aunque en este caso resulta más complejo ya que algunos MAGs no eran de alta calidad y se pudo observar que el contexto no estaba muy conservado en algunos casos. De manera que es importante la obtención de MAGs de alta calidad para observar una mayor conservación del contexto genómico y de esta manera mejorar las predicciones.

El análisis de las distancias filogenéticas entre las BCLs predichas y su distribución en más de un cromosoma mostró que las copias localizadas en el cromosoma fueron adquiridas en el primer evento evolutivo, mientras que las copias localizadas en los crómidos putativos fueron adquiridas posteriormente en otro evento evolutivo. En este estudio se observó que algunas especies que presentaron a la BCL estaban asociadas con otros organismos parásitos o simbiontes, algunos fueron aislados de sitios contaminados con hidrocarburos aromáticos o algún otro contaminante. Además, el análisis de los MAGs permitió explorar el ecosistema marino, y la identificación de enzimas Box en diferentes clases, como Beta, Alpha y Gammaproteobacteria, muestra la diversidad de los organismos que poseen estas enzimas y su potencial para la biodegradación de hidrocarburos. Estos resultados sugieren que la capacidad de degradación de hidrocarburos se encuentra distribuida en una variedad de linajes microbianos y ambientes. Lo cual indica que los hidrocarburos aromáticos están ampliamente distribuidos en distintos ecosistemas y las bacterias han desarrollado diferentes estrategias para degradarlos. El progreso en la identificación bioinformática de estas vías ha incrementado nuestro conocimiento de la distribución de enzimas que degradan los

hidrocarburos en condiciones aeróbicas o anaeróbicas. Sin embargo, actualmente los estudios experimentales son limitados por lo que se deben realizar esfuerzos futuros para confirmar las predicciones obtenidas en especies distintas de las estudiadas experimentalmente. Tal como lo demuestra la presencia de *Cupriavidus gilardii* CR3 en los MAGs analizados, un organismo previamente aislado y cultivado (Wang *et al.*, 2015), lo que permitió validar dicha predicción de esta especie como degradadora de hidrocarburos aromáticos. Por lo que los estudios de organismos aislados ayudan a validar y expandir la comprensión de la biodegradación de hidrocarburos.

De manera general, los hallazgos presentados proporcionan una visión más completa de la distribución de enzimas BCL en la naturaleza, se destaca la importancia de la calidad de los MAGs, la filogenia de los organismos y la anotación funcional precisa. Estos resultados pueden ser fundamentales para futuras investigaciones sobre la biodegradación de hidrocarburos y la ecología microbiana en ambientes marinos y terrestres.

12. Perspectivas

- La validación experimental es esencial para confirmar las predicciones obtenidas en especies distintas de las estudiadas experimentalmente. Se pueden llevar a cabo experimentos para verificar la actividad de las enzimas BCL en diferentes organismos y condiciones, lo que contribuiría a una comprensión más sólida de la biodegradación de hidrocarburos.

- Profundizar en la caracterización funcional de las enzimas BCL y su interacción con otros genes y proteínas en el contexto genómico, especialmente el regulador de la transcripción BdzR/BoxR. Esto podría incluir estudios de expresión génica y regulación para comprender mejor cómo estas enzimas contribuyen a la degradación de hidrocarburos.

- Desarrollar métodos y técnicas mejoradas para obtener MAGs de alta calidad, lo que permitiría obtener una mayor conservación del contexto genómico y, por lo tanto, mejorar las predicciones de ortología.

- Investigar la coevolución de genes relacionados con la degradación de hidrocarburos y las enzimas BCL podría proporcionar información valiosa sobre las adaptaciones microbianas a diferentes entornos contaminados.

- Utilizar los conocimientos derivados de esta investigación para desarrollar estrategias de biorremediación de ambientes contaminados con hidrocarburos. Esto podría

incluir el desarrollo de biotecnologías que aprovechen la enzima BCL y las vías río abajo para la biodegradación de contaminantes.

13. Referencias

Aguilar-Bultet, L., and Falquet, L. (2015). Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos. Revista de salud animal, 37(2), 125-132.

Altenhoff, A. M., Glover, N. M., and Dessimoz, C. (2019). Inferring orthology and paralogy. Methods Mol Biol. 1910:149-175.

Anderson, R., Mawji, E., Cutter, G., Measures, C. and Jeandel, C. (2014). GEOTRACES: Changing the way we explore ocean chemistry. Oceanography, 27, 50–61.

Arnold, M. E., Kaplieva-Dudek, I., Heker, I., and Meckenstock, R. U. (2021). Aryl Coenzyme A Ligases, a subfamily of the adenylate-forming enzyme superfamily. Applied and environmental microbiology, 87(18), e00690-21.

Bailey, T.L., and Gribskov, M. (1998). Combining evidence using p-values: Application to sequence homology searches. Bioinformatics 14, 48–54.

Bains, J., and Boulanger, M. J. (2007). Biochemical and structural characterization of the paralogous benzoate CoA ligases from *Burkholderia xenovorans* LB400: defining the entry point into the novel benzoate oxidation (*box*) pathway. Journal of molecular biology, 373(4), 965-977

Biller, S., Berube, P., Dooley, K. *et al*. (2018). Marine microbial metagenomes sampled across space and time. Sci Data 5, 180176.

Brzeszcz, J., and Kaszycki, P. (2018). Aerobic bacteria degrading both n-alkanes and aromatic hydrocarbons: an undervalued strategy for metabolic diversity and flexibility. Biodegradation, 29, 359-407.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). Trimal: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.

Carmona, M. Zamarro, M. T. Blazquez, B. Durante-Rodriguez, G. Juarez, J. F. Valderrama, J. A. Barragan, M. J. L. Garcia, J. L. Diaz, E. (2009). Anaerobic catabolism of aromatic compounds: a genetic and genomic view. Microbiology and Molecular Biology Reviews, 73(1), 71–133.

Chang A., Jeske L., Ulbrich S., Hofmann J., Koblitz J., Schomburg I., Neumann-Schaal M., Jahn D, Schomburg D. (2021). BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Research, 49, D498-D508.

Chen, S., Zhou, Y., Chen, Y., Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884–i890.

Clark L, Leatherby D, Krilich E, Ropelewski AJ, Perozich J. (2018). In silico analysis of class I adenylate-forming enzymes reveals family and group specific conservations. Plos One, 13, e0203218.

Davidson, P., Eutsey, R., Redler, B., Hiller, N.L., Laub, M.T., and Durand, D. (2018). Flexibility and constraint: Evolutionary remodeling of the sporulation initiation pathway in Firmicutes. Plos Genet. 14, 1–33.

Dasgupta, D., Ghosh, R., and Sengupta, T. K. (2013). Biofilm-mediated enhanced crude oil degradation by newly isolated Pseudomonas species. International Scholarly Research Notices, 2013.

de Melo, A. P. Z., Hoff, R. B., Molognoni, L., de Oliveira, T., Daguer, H., and Barreto, P. L. M. (2022). Disasters with oil spills in the oceans: Impacts on food safety and analytical control methods. Food Research International, 111366.

Denef, V.J., Klappenbach, J.A., Patrauchan, M.A., Florizone, C., Rodrigues, J.L.M., Tsoi, T. V., Verstraete, W., Eltis, L.D., and Tiedje, J.M. (2006). Genetic and genomic insights into the role of benzoate-catabolic pathway redundancy in *Burkholderia xenovorans* LB400. Applied and Environmental Microbiology. 72, 585–595

Denef, V. J., Patrauchan, M. A., Florizone, C., Park, J., Tsoi, T. V., Verstraete, W., and Eltis, L. D. (2005). Growth substrate-and phase-specific expression of biphenyl, benzoate, and C1 metabolic pathways in *Burkholderia xenovorans* LB400. Journal of Bacteriology .187(23), 7996-8005.

Díaz, E., Jiménez, J. I., and Nogales, J. (2013). Aerobic degradation of aromatic compounds. Current Opinion in Biotechnology, 24(3), 431-442.

Dicenzo, G.C., Mengoni, A., and Perrin, E. (2019). Chromids aid genome expansion and functional diversification in the family Burkholderiaceae. Molecular, Biology and Evolution. 36, 562–574.

Dicenzo, G.C., and Finan, T.M. (2017). The Divided Bacterial Genome. Molecular Microbiology. 81, 1–37.

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 (2015). Bioinformatics, 31 (10), 1674–1676, https://doi.org/10.1093/bioinformatics/btv033

Dranenko, N. O., Rodina, A. D., Demenchuk, Y. V., Gelfand, M. S., and Bochkareva, O. O. (2023). Evolutionary trajectories of secondary replicons in multipartite genomes. bioRxiv, 2023-04.

Edgar, R.C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nature communications. 13.

England, P. G., Pelletier, D. A., Dispensa, M., Gibson, J., and Harwood, C. S. (1997). A cluster of bacterial genes for anaerobic benzene ring biodegradation. Proceedings of the National Academy of Sciences, 94(12), 6484-6489.

Fernández Juárez, J. (2011). Caracterización de los genes *mbd* de *Azoarcus* sp. CIB: Descripción de una nueva ruta para la degradación anaeróbica de compuestos aromáticos

Fernandez, H., Prandoni, N., Fernandez-Pascual, M., Fajardo, S., Morcillo, C., Diaz, E., and Carmona, M. (2014). *Azoarcus* sp. CIB, an anaerobic biodegrader of aromatic compounds shows an endophytic lifestyle. PLoS One, 9(10), e110771.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). Sequence analysis CD-HIT: accelerated for clustering the next-generation sequencing data. 28, 3150–3152.

Fuchs, G. (2008). Anaerobic Metabolism of Aromatic Compounds. Annals of the New York Academy of Sciences, 1125, 82–99.

Georg Fuchs. (2002). Genes coding for a new pathway of aerobic benzoate metabolism in Azoarcus evansii. Journal of Bacteriology, 184(22), 6301-6315.

Gescher, J., Zaar, A., Mohamed, M., Schägger, H., and Fuchs, G. (2002). Genes coding for a new pathway of aerobic benzoate metabolism in *Azoarcus evansii*. Journal of Bacteriology, 184(22), 6301-6315.

Geissler, J. F., Harwood, C. S., and Gibson, J. (1988). Purification and properties of benzoate-coenzyme A ligase, a *Rhodopseudomonas palustris* enzyme involved in the anaerobic degradation of benzoate. Journal of bacteriology, 170(4), 1709-1714.

Golby, S., Ceri, H., Gieg, L.M., Chatterjee, I., Marques, L.L.R., and Turner, R.J. (2012). Evaluation of microbial biofilm communities from an Alberta oil sands tailings pond. FEMS Microbiology Ecology 79, 240–250.

Gulick, A.M. (2009). Conformational dynamics in the acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. ACS Chemical Biology. 4, 811–827.

Harwood, C.S., Burchhardt, G., Herrmann, H., and Fuchs, G. (1998). Anaerobic metabolism of aromatic compounds via the benzoyl-CoA pathway. FEMS Microbiology Ecology. 22, 439–458

Hirakawa, H., Hirakawa, Y., Greenberg, E. P., and Harwood, C. S. (2015). BadR and BadM proteins transcriptionally regulate two operons needed for anaerobic benzoate degradation by *Rhodopseudomonas palustris*. Applied and environmental microbiology, 81(13), 4253-4262.

Huang IK, Pei J, Grishin NV. (2013). Defining and predicting structurally conserved regions in protein superfamilies. Bioinformatics, 29(2), 175-181.

Hon, T., Mars, K., Young, G., Tsai, Y. C., Karalius, J. W., Landolin, J. M., ... and Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. Scientific data, 7(1), 399.

Heider J., Alfred M. Spormann, Harry R. Beller, Friedrich Widdel. (1998). Anaerobic bacterial metabolism of hydrocarbons. FEMS Microbiology Reviews, 22(5), 459–473.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases, and drugs. Nucleic Acids Research. 45, D353–D361.

Karl, D. M., and Church, M. J. (2014). Microbial oceanography and the Hawaii Ocean Time-series programme. Nature Reviews Microbiology, 12(9), 699–713.

Karigar, C. S., and Rao, S. S. (2011). Role of microbial enzymes in the bioremediation of pollutants: a review. Enzyme research.

Kawaguchi, K., Shinoda, Y., Yurimoto, H., Sakai, Y., & Kato, N. (2006). Purification and characterization of benzoate-CoA ligase from *Magnetospirillum* sp. strain TS-6 capable of aerobic and anaerobic degradation of aromatic compounds. FEMS microbiology letters, 257(2), 208-213.

Khot, V., Zorz, J., Gittins, D. A., Chakraborty, A., Bell, E., Bautista, M. A., ... and Bhatnagar, S. (2022). CANT-HYD: A curated database of phylogeny-derived Hidden Markov Models for annotation of marker genes involved in hydrocarbon degradation. Frontiers in Microbiology, 12, 764058.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A bioinformatician guide to metagenomics. Microbiology and Molecular Biology Reviews, 72(4), 557-578.

Kelly, A., and Salgado, P.S. (2019). The engulfasome in C. difficile: Variations on protein machineries. Anaerobe *60*, 102091.

Kumar, V., Shahi, S. K., and Singh, S. (2018). Bioremediation: an eco-sustainable approach for restoration of contaminated sites. Microbial bioprospecting for sustainable development, 115-136.

Lapidus, A. L., and Korobeynikov, A. I. (2021). Metagenomic data assembly–the way of decoding unknown microorganisms. Frontiers in Microbiology, 12, 613791.

Laczi, K., Erdeiné Kis, Á., Szilágyi, Á., Bounedjoum, N., Bodor, A., Vincze, G. E., ... and Perei, K. (2020). New frontiers of anaerobic hydrocarbon biodegradation in the multi-omics era. Frontiers in Microbiology, 11, 590049.

Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, Do L, Land ML, Pelletier DA, Beatty JT, Lang AS, Tabita FR, Gibson JL, Hanson TE, Bobst C, Torres JL, Peres C, Harrison FH, Gibson J, Harwood CS (2004). Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. Nature Biotechnology. (1):55-61. doi: 10.1038/nbt923. Epub 2003 Dec 14. PMID: 14704707.

López Barragán, M. J., Díaz, D., García, J. L., and Carmona, M. (2004). Genetic clues on the evolution of anaerobic catabolism of aromatic compounds. Society for General Microbiology.

Letunic, I., and Bork, P. (2021). Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research. *49*.

Lin, H. H., and Liao, Y. C. (2016). Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Scientific reports, 6(1), 24175.

Lykidis, A., Pérez-Pantoja, D., Ledger, T., Mavromatis, K., Anderson, I.J., Ivanova, N.N., Hooper, S.D., Lapidus, A., Lucas, S., González, B., *et al*. (2010). The complete multipartite genome sequence of *Cupriavidus necator* JMP134, a versatile pollutant degrader. PLoS One.

Loza, A., García-Guevara, F., Segovia, L., Escobar-Zepeda, A., Sanchez-Olmos, M. D. C., Merino, E., Gutierrez-Rios, R. M. (2022). Definition of the metagenomic profile of ocean water samples from the gulf of mexico based on comparison with reference samples from sites worldwide. Frontiers in Microbiology, 12, 781497.

Henkin, T. M., and Peters, J. E. (2020). Snyder and Champness molecular genetics of bacteria. John Wiley & Sons.

Marahiel MA, Stachelhaus T, Mootz HD. (1997). Modular peptide synthetases involved in nonribosomal peptide synthesis. Chemical Reviews 97:2651–2674.

Martínez-Amador, P., Castaneda, N., Loza, A., Soto, L., Merino, E., ... & Gutierrez-Rios, R. M. (2019). Prediction of protein architectures involved in the signaling-path signaling-pathway initiating sporulation in Firmicutes. BMC Research Notes, 12(1), 686.

Martín-Moldes, Z., Zamarro, M.T., del Cerro, C., Valencia, A., Gómez, M.J., Arcas, A., Udaondo, Z., García, J.L., Nogales, J., Carmona, M., *et al*. (2015). Whole-genome analysis of *Azoarcus* sp.strain CIB provides genetic insights to its different lifestyles and predicts novel metabolic features. Systematic and Applied Microbiology. 38, 462–471

Mistry, J., Chuguransky, S., Paladin, L., Qureshi, M., Raj, S., Richardson, L.J., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Williams, L., *et al*. (2021). Pfam: The protein families database in 2021. Nucleic Acids Research. 49, D412–D419.

Mikheenko, A., Gurevich, A., Saveliev, V. (2016). MetaQUAST: evaluation of metagenome assemblies. Bioinformatics, 32(7), 1088-1090.

Meckenstock, R. U., and Mouttaki, H. (2011). Anaerobic degradation of non-substituted aromatic hydrocarbons. Current Opinion in Biotechnology, 22(3), 406-414.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. Bioinformatics, 32(7), 1088-1090.

Mouttaki, H., James, K. L., Loo, R. R. O., *et al.* (2022). The Acyl-Proteome of *Syntrophus aciditrophicus* reveals metabolic relationships in benzoate degradation. Molecular & Cellular Proteomics, 21(4).

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution. 32, 268–274.

National Research Council. (2007). The new science of metagenomics: revealing the secrets of our microbial planet.

Nishimura, Y., and Yoshizawa, S. (2022). The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originating from various marine environments. Scientific Data, 9(1), 1-11.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Research. 27(5):824-834.

Parab, Vivek; Phadke, Manju (2020). Co-biodegradation studies of naphthalene and phenanthrene using bacterial consortium. Journal of Environmental Science and Health, Part A, (), 1–13. doi:10.1080/10934529.2020.1754054

Pérez-Pantoja, D., Donoso, R., Agulló, L., Córdova, M., Seeger, M., Pieper, D.H., and González, B. (2012). Genomic analysis of the potential for aromatic compounds biodegradation in Burkholderiales. Environmental Microbiology. 14, 1091–1117.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., ... and Searson, S. (2015). Open science resources for the discovery and analysis of Tara Oceans data. Scientific data, 2(1), 1-16.

Porter, A.W., and Young, L.Y. (2014). Benzoyl-CoA, a Universal Biomarker for Anaerobic. Degradation of Aromatic Compounds. Advances in Applied Microbiology. 88:167-203.

Prakash, T., and Taylor, T. D. (2012). Functional assignment of metagenomic data: challenges and applications. Briefings in bioinformatics, 13(6), 711-727.

Riccardi, C., Koper, P., Innocenti, G., Dicenzo, G.C., Fondi, M., Mengoni, A., and Perrin, E. (2023). Independent origins and evolution of the secondary replicons of the class Gammaproteobacteria. Microbiology Genomics 9, 1–12.

Robinson, S.L., Terlouw, B.R., Smith, M.D., Pidot, S.J., Stinear, T.P., Medema, M.H., and Wackett, L.P. (2020). Global analysis of adenylate-forming enzymes reveals b-lactone biosynthesis pathway in pathogenic nocardia. J. Biol. Chem. 295, 14826–14839.

Schühle, K., Gescher, J., Feil, U., Paul, M., Jahn, M., Schägger, H., and Fuchs, G. (2003). Benzoate-coenzyme a ligase from *Thauera aromatica*: An enzyme acting in anaerobic and aerobic pathways. Joutnal of Bacteriology. 185, 4920–4929.

Simon C Potter, Aurélien Luciani, Sean R Eddy, Youngmi Park, Rodrigo Lopez, Robert D Finn. (2018). HMMER web server: 2018 update. Nucleic Acids Research, 46(W1), W200–W204.

Soto-Avila, L., Ciria-Moreno, C., Merino, E., Segovia, L., Soberón, X., and Gutiérrez-Ríos, R. M. (2012). A general profile for P2 promoter sequences in bacteria of the Gammaproteobacteria and Betaproteobacteria classes. Molecular Biosystems, 8(12), 3212-3219.

Suvorova, I.A., and Gelfand, M.S. (2019). Comparative genomic analysis of the regulation of aromatic metabolism in betaproteobacteria. Frontiers Microbiology. 10, 1–18.

She, R., Chu, J. S. C., Wang, K., Pei, J., and Chen, N. (2009). GenBlastA: enabling BLAST to identify homologous gene sequences. Genome research, 19(1), 143-149.

Scoma, A., Yakimov, M. M., Daffonchio, D., and Boon, N. (2017). Self-healing capacity of deep-sea ecosystems affected by petroleum hydrocarbons: Understanding microbial oil degradation at hydrocarbon seeps is key to sustainable bioremediation protocols. EMBO reports, 18(6), 868-872.

Timothy, L. B., James, J., Charles, E. G., and William, S. N. (2015). The MEME suite. Nucleic Acids Research, 43(W1), W39-W49.

Thornburg, C.K., Wortas-Strom, S., Nosrati, M., Geiger, J.H., and Walker, K.D. (2015). Kinetically and crystallographically guided mutations of a benzoate CoA ligase (*bada*) elucidate mechanism and expand substrate permissivity. Biochemistry 54, 6230–6242.

Valderrama, J.A., Durante-Rodríguez, G., Blázquez, B., García, J.L., Carmona, M., and Díaz, E. (2012). Bacterial degradation of benzoate: Cross-regulation between aerobic and anaerobic pathways. Journal of Biological Chemistry. 287, 10494–10508.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A. and Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. science, 304(5667), 66-74.

Wang, X., Chen, M., Xiao, J., Hao, L., Crowley, D. E., Zhang, Z., ... & Wu, J. (2015). Genome sequence analysis of the naphthenic acid degrading and metal resistant bacterium *Cupriavidus gilardii* CR3. Plos One, 10(8), e0132881.

Warman, E.A., Forrest, D., Guest, T., Haycocks, J.J.R.J., Wade, J.T., and Grainger, D.C. (2021). Widespread divergent transcription from bacterial and archaeal promoters is a consequence of DNA-sequence symmetry. Nature Microbiology. 6, 746–756.

Whitman WB, Rainey F, Kämpfer P, Trujillo M, Chun J, DeVos P, Hedlund B, Dedysh S. (2018). Bergey's manual of systematics of archaea and bacteria. Bergey's Manual Trust.

Xin G, Cai Z, Cai S, Xu H, Zhou G. (2014). Cloning and characterization of a new benzoate CoA ligase from *Pseudomonas citronellolis.* International Journal of Molecular Sciences, 15(8), 14321–14336.

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., and Zhang, L. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. Computational and Structural Biotechnology Journal, 19, 6301-6314.

14. Anexo

TABLAS: Las tablas descritas a continuación se presentan en la siguiente liga **https://docs.google.com/spreadsheets/d/147J0C_IoIqgDBBjqtMyD527uQ6NPwR3X/edit?usp=sharing&ouid=100425136616765459611&rtpof=true&sd=true**

Tabla 1A. Arquitectura Pfam de las proteínas del contexto contenidas en organismos modelo. Se muestra la composición de los dominios Pfam presentes en cada proteína perteneciente al contexto genómico en los organismos con evidencia experimental de presentar la enzima BCL.

Tabla 2A. Perfil filético de la enzima Benzoato CoA ligasa y las proteínas de su contexto genómico. Se presentan los POs de la BCL, 132 pertenecientes a la categoría FC_FM, 13 pertenecientes a la categoría IC_FM y 3 pertenecientes a la categoría FC_DM. Se denota con un número 1 la presencia de la proteína y con un 0 la ausencia.

Tabla 3A. Proteínas parálogas de la BCL. Se presentan aquellos organismos que presenten dos copias de la BCL en sus genomas, así como el contexto genómico. Los organismos que pertenecen a distintas categorías son separados entre sí.

Tabla 4A. **Sitios de aislamiento de los POs obtenidos.** Se muestra la descripción del sitio de aislamiento u muestreo del cual las secuencias analizadas fueron obtenidas.

Tabla 5A. POs en genomas ensamblados de metagenomas. Se muestra el perfil filético de los 58 POs obtenidos a partir del análisis de los MAGs, los cuales están organizados de acuerdo con la presencia de las proteínas del contexto genético

Tabla 6A. Calidad de los MAGS tomados de OceanDNA MAGs. Se muestran las calidades de los 58 MAGs que presentan los POs de la BCL.

Tabla 7A. Calidad de los MAGS pertenecientes al Golfo de México. Se muestran las calidades de 11 bins del Golfo de México.