PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

MAESTRÍA EN BIOLOGÍA COMPUTACIONAL

Ensamblaje y anotación del genoma de una proteobacteria perteneciente a la familia Orbaceae a partir de datos metagenómicos del tracto digestivo de la mosca soldado negra *Hermetia illucens*

Trabajo previo a la obtención del título de

Magíster en Biología Computacional

YADIRA ESTEFANÍA SALGUERO SALAS

Quito, 2023

DERECHOS DE AUTOR

Expreso que soy la autora del presente trabajo de titulación y consulté las referencias indicadas en el mismo. Este trabajo no fue presentado de forma previa para la obtención de ningún grado académico.

La Pontificia Universidad Católica del Ecuador puede utilizar los derechos del trabajo de titulación de acuerdo con la Ley de Propiedad Intelectual y su normativa institucional.

_____

Mg. Yadira Salguero Salas

APROBACIÓN DEL DIRECTOR DEL TRABAJO DE TITULACIÓN

Certifico que el trabajo de la Mg. Yadira Salguero Salas para la obtención del título de Magíster en Biología Computacional se llevó a cabo bajo la normativa y reglamentación institucional y puede ser presentada para su calificación.

_____

Ing. Francisco Flores, Ph.D

HOJA DE EVIDENCIA ANTIPLAGIO

DEDICATORIA

A Dios y Amelia

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

RESUMEN

El objetivo de este trabajo fue ensamblar y anotar el genoma de una proteobacteria perteneciente a la familia Orbaceae, a partir de datos metagenómicos del tracto digestivo *de Hermetia illucens* en estado larvario y conocer qué genes y rutas metabólicas de esta bacteria están involucrados en el proceso de biotransformación de residuos.

El trabajo se dividió en tres etapas: la primera fue la eliminación de las lecturas del hospedero, la segunda consistió en el ensamblaje y anotación del metagenoma y la tercera en la búsqueda de genes y rutas metabólicas en el genoma de interés.

La primera etapa inició con un control de calidad con la herramienta *FASTQC*, para continuar con la limpieza de las secuencias con *trimmommatic* y la re-evaluación de su calidad. Los resultados indicaron que, el 5,07 % de las lecturas fueron eliminadas por su baja calidad. Luego se eliminaron las lecturas duplicadas (11,76 %) con el paquete *BBMap*. De los 17 677 745 de lecturas *paired end* restantes, el 59,69 % se alinearon y mapearon sobre el genoma de referencia del hospedero con la herramienta *Bowtie2*. S*amtools* permitió obtener un archivo .bam con las lecturas no mapeadas al genoma de referencia. Posteriormente, las lecturas se dividieron en dos archivos (*forward* y *reverse*), en formato FASTQ, con ayuda de la herramienta *bedtools*.

La segunda etapa inició con una predicción de la composición microbiana con *Kaiju*, para para su posterior comparación con árboles de especie generados luego del ensamblaje y anotación del metagenoma. Con esta herramienta se encontró que los microorganismos más abundantes pertenecieron a los filos Firmicutes, Proteobacteria, Bacteroidetes y Actinobacteria. Así mismo las especies con abundancia relativa alta fueron, *Hungatella hathewayi*, *Frischella perrara*, *Gilliamella apícola*, *Enterococcus* sp., *Providencia rettgeri*, entre otras. Para el ensamblaje del metagenoma se utilizaron tres herramientas: *metaSPAdes*, *IDBA-UD* y *MEGAHIT*, cuyos resultados fueron comparados con *Compare Assembled Contig Distributions*. La herramienta que permitió un mejor ensamblaje fue *metaSPAdes* cuyo *contig* más largo fue de 706 116 pb y un N50 de 22 421 pb. Posteriormente, los *contigs* metagenómicos ensamblados fueron agrupados en *bins* con las herramientas *MaxBin2*, *MetaBAT2* y *CONCOCT*, luego de lo cual se procedió a su optimización con *DASTool*, obteniéndose un porcentaje de agrupamiento del 38,52 % en 10 *bins* de alta calidad (BinScore > 0.7). La evaluación de la calidad de los *bins* se llevó a cabo con la herramienta CheckM, donde se encontró que el *bin* 007 pertenecía a una gammaproteobacteria y con base en conjuntos de genes marcadores que están presentes en una sola copia se determinó una

integridad del 99,44 % (número de conjuntos de marcadores presentes con respecto a los conjuntos de marcadores de referencia) y una contaminación del 0,05 % (número de genes marcadores que están presentes en varias copias en cada conjunto de marcadores). La anotación se llevó a cabo con *Annotate Multiple Microbial Assemblies with RASTtk (AMMA with RASTtk)*, donde el *bin* 007 fue asignado al género Orbus y tuvo una longitud total de 3 084 216 pb, con 2 881 genes, de los cuales el 97,29 % fueron codificantes.

La clasificación taxonómica con *Genome Taxonomic Database (GTDB)-Tk* reportó los mismos resultados que con *AMMA with RASTtk*. Posteriormente, se construyó un árbol de especies con *Species tree* que indicó que los microorganismos más cercanos encontrados para el *bin* 007 fueron de los géneros *Frischella* y *Gilliamella*, lo cual concuerda con los resultados obtenidos tanto en *Kaiju*, como en *AMMA with RASTtk* y *(GTDB)-Tk*, indicando que el *bin* 007 corresponde a Orbus sp. IPMB12 y al género Orbus, respectivamente; motivo por el cual se pudo asegurar que la bacteria pertenece a la familia Orbaceae. Adicionalmente, se realizó otro árbol filogenético con la herramienta *Bacterial Genome Tree* donde se estableció que el genoma posiblemente corresponde a una especie aún no caracterizada del género Orbus.

La tercera etapa inició con la herramienta *DRAM* que proporcionó un resumen del perfil funcional o metabólico del genoma anotado. La presencia de rutas metabólicas y funciones vinculadas con la utilización de carbono, nitrógeno, celulosa amorfa y producción de alcohol se relacionó con la eficacia de degradación de residuos orgánicos por parte de esta bacteria en los intestinos de *H. illucens*, contribuyendo en la bioconversión de residuos con un metabolismo anaeróbico facultativo. Posteriormente, se utilizó *Build FeatureSet from Genome* y *Merge FeatureSets* para construir un conjunto de características con genes pertenecientes a PET hidrolasas. Con MUSCLE y *Gblocks* se generó una alineación de secuencias múltiple de secuencias de proteínas recortada para eliminar regiones variantes. Por último, se usó *HMMERsearch from MSA* cuyo informe indicó que, no se encontraron genes pertenecientes a las PETasas pero no se descartó la posibilidad de que existan otras enzimas relacionadas con la degradación de PET.

ABSTRACT

This work aimed was to assemble and annotate a proteobacteria genome belonging to the Orbaceae family, based on metagenomic data from the digestive tract of *H. illucens* in the larval stage, and to know which genes and metabolic pathways of these bacterium are involved in the waste biotransformation process.

The work was divided into three stages: the first stage was the elimination of the host reads, the second stage consisted of the assembly and annotation of the metagenome, and the third one consisted of the search for genes and metabolic pathways in the genome of interest.

The first stage began with a quality control by using the *FASTQC* tool, to continue with the cleaning of the sequences with *trimmommatic* and the re-evaluation of their quality. The results indicated that 5.07 % of the reads were eliminated due to their low quality. Afterwards, duplicate reads (11.76%) were removed with the *BBMap* package. Of the remaining 17 677 745 paired end reads, 59.69 % were aligned and mapped to the host reference genome using *Bowtie2*. *Samtools* allowed to obtain a .bam file with the unmapped reads to the reference genome. Subsequently, the readings were then divided into two files (forward and reverse), in FASTQ format, with the help of the *bedtools* tool.

The second stage began with a prediction of the microbial composition with *Kaiju*, for its subsequent comparison with species trees generated after assembly and annotation of the metagenome. With this tool it was found that the most abundant microorganisms belonged to the Firmicutes, Proteobacteria, Bacteroidetes and Actinobacteria phyla. Likewise, the species with high relative abundance were *Hungatella hathewayi, Frischella perrara, Gilliamella apícola, Enterococcus sp., Providencia rettgeri*, among others. Three tools were used to assemble the metagenome: *metaSPAdes*, *IDBA-UD* and *MEGAHIT*, which results were compared with *Compare Assembled Contig Distributions*. The tool that allowed a better assembly was *metaSPAdes* whose longest contig was 706 116 bp and an N50 of 22 421 bp. Subsequently, the assembled metagenomic contigs were grouped into bins with the *MaxBin2, MetaBAT2* and *CONCOCT* tools, after that, they were optimized with *DASTool*, obtaining a grouping percentage of 38.52 % in 10 high-quality bins (BinScore > 0.7). The bins quality assessment was performed using the *CheckM* tool, where it was found that bin 007 belonged to a gammaproteobacteria and based on marker genes set that are present in a single copy, an 99.44 % completeness was determined (number of marker sets present relative to reference marker sets) and 0.05 % contamination (number of marker genes that are present in multiple

copies in each marker set). The annotation was performed with *Annotate Multiple Microbial Assemblies with RASTtk (AMMA with RASTtk)*, where bin 007 was assigned to the Orbus genus and it obtained a total length of 3 084 216 bp, with 2 881 genes, where 97,29 % of them turned out to be coding genes.

Taxonomic classification with *Genome Taxonomic Database (GTDB)-Tk* reported the same results as *AMMA with RASTtk.* Subsequently, a species tree was built with *Species tree* tool, which pointed out that the closest microorganisms found for bin 007 were from the genera *Frischella* and *Gilliamella*, which was consistent with the results obtained both in *Kaiju*, and in *AMMA with RASTtk* and *(GTDB)-Tk*, indicating that bin 007 corresponds to Orbus sp. IPMB12 and the Orbus genus, respectively; reason which was made possible to ensure that the bacterium belongs to the Orbaceae family. Additionally, another phylogenetic tree was made with the *Bacterial Genome Tree* tool where it was established that the genome possibly corresponds to a species of the Orbus genus that has not yet been characterized.

The third stage started with the *DRAM* tool that provided a summary of the functional or metabolic profile of the annotated genome. The presence of metabolic pathways and functions related to the use of carbon, nitrogen, amorphous cellulose, and alcohol production was related to the efficiency of organic waste degradation by this bacterium in of *H. illucens* intestin, contributing to the bioconversion of residues with a facultative anaerobic metabolism. Subsequently, *Build FeatureSet from Genome* and *Merge FeatureSets* were used to build a feature set with genes belonging to PET hydrolases. A trimmed multiple sequence alignment of protein sequences to remove variant regions was generated with *MUSCLE* and *Gblocks*. Finally, *HMMERsearch from MSA* was used, whose report indicated that no genes belonging to PETases were found, but the possibility that there are other enzymes related to PET degradation was not ruled out.

1. INTRODUCCIÓN

1.1. FORMULACIÓN DEL PROBLEMA

La generación de residuos sólidos es una de las mayores preocupaciones e inconvenientes a nivel mundial. Se considera como residuo a todo material sin valor económico para el usuario pero cuya recuperación e incorporación en el ciclo de vida de la materia representa un valor comercial (INEC, 2020). Los residuos pueden ser de dos tipos: orgánicos o inorgánicos. Los residuos orgánicos son derivados de la elaboración de alimentos y es indispensable diferenciar las pérdidas del desperdicio. Las pérdidas se generan al producir, cosechar, almacenar y transportar los alimentos; mientras que, los desperdicios ocurren al distribuirlos y consumirlos, y por lo tanto, tienen una relación directa con la conducta de vendedores y consumidores, que la mayor parte de las veces, deciden desechar los productos que todavía pueden aprovecharse (Salguero & Guevara, 2019).

Si los residuos orgánicos e inorgánicos fueran separados y clasificados de manera adecuada, la mayor parte podría reciclarse, no obstante, por lo general la población está acostumbrada a la producción, consumo y descarte. La materia orgánica separada podría utilizarse en la elaboración de abonos u otros productos, sin embargo, al estar combinada con residuos inorgánicos, tales como el plástico, no puede llevarse a cabo esta transformación.

Los plásticos presentes en los rellenos sanitarios se degradan por fraccionamiento hasta que alcanzan tamaños microscópicos, lo que hace inevitable su consumo por parte de insectos que predominan en lugares con materia en proceso de descomposición (Pazmiño, 2021). Se ha encontrado evidencia de que las larvas de insectos de *Plodia interpunctella* pueden degradar polietileno (PE) (J. Yang, Yang, Wu, Zhao, & Jiang, 2014; Y. Yang, Chen, Wu, Zhao, & Yang, 2015), *Tenebrio molitor* PE y poliestireno (PS) (Brandon et al., 2018), *Galleria mellonella* PS (Shan, Su, Zhao, & Wang, 2021), *Zophobas asatratus* PS y poliuretano (PU) (Luo et al., 2021), *Hermetia illucens* sustratos con macroplásticos de PVC (Lievens et al., 2022), además el peso larvario incrementa en presencia de PS (Cho, Kim, Kim, & Chung, 2020). La degradación de polímeros por parte de las larvas se da, en gran medida, gracias a la comunidad microbiana presente en sus intestinos

(Mitra & Das, 2023; Shan et al., 2021; J. Yang et al., 2014). De acuerdo con Sun, Prabhu, Aroney y Rinke (2022), la degradación de poliestireno (PS) se ve facilitada inicialmente por factores ambientales (radiación UV, fuerzas mecánicas, viento, entre otras) que permiten que el polímero reaccione con pequeñas moléculas de agua y oxígeno y se formen grupos carbonilo que son más accesibles a la degradación enzimática; así mismo, los microorganismos podrían utilizar hidroxilasas y deshidrogenasas para la formación de los grupos carbonilo. Posteriormente, enzimas con actividad catalítica como las lipasas, serina hidrolasas o PET hidrolasas (PETasas) podrían facilitar la descomposición de PS al dirigirse al grupo carbonilo (Sun, Prabhu, Aroney, & Rinke, 2022). No obstante, el conocimiento del proceso de degradación de los plásticos a través de insectos aún es limitado (Mitra & Das, 2023).

Así, existe un interés creciente en investigaciones que involucren a la degradación de materia orgánica y plásticos por parte de insectos en estado larvario (Kuan, Chan, & Gan, 2022; Pazmiño, 2021). Un insecto que ha llamado mucho la atención es la mosca soldado negra (*H. illucens),* por su gran voracidad de sustratos orgánicos (Bruno et al., 2019; ESR International, 2016; Kuan et al., 2022; Oliver, 2004; Zhineng, Ying, Bingjie, Rouxian, & Qiang, 2021) y en cuyo tracto digestivo pueden existir varias especies de microorganismos no descritas que necesitan ser caracterizadas para lograr encontrar genes importantes en la biotransformación de la materia orgánica o degradación de plástico. De acuerdo con Danso, *et al*. (2018), los genes pertenecientes a las PETasas actualmente conocidos son los siguientes: Cut190 (W0TJ64), cut1 (E9LVI0), cut-2 (E5BBQ3), Tcur_1278 (D1A9G5), cut1 (E9LVH7), cut (H6WX58), cut2 (E9LVH9), ISF6_4831 (A0A0K8P6T7), Deinma_1209, liplAF5-2_UB, J057_15340_Mna, pbsA, est_Psyc OLEAN_C07960_Oant, y se producen principalmente en los filos Actinobacteria, Proteobacteria y Bacteroidetes (Danso et al., 2018).

Los microorganismos del tracto digestivo de *H. illucens* juegan un papel protagónico en la degradación de residuos (Shelomi, Wu, Chen, Huang, & Burke, 2020). Investigaciones previas (Shelomi et al., 2020; Zhineng et al., 2021) y un ensayo preliminar con datos metagenómicos han indicado que de entre las bacterias, las proteobacterias de la familia Orbaceae son de las más abundantes en el tracto digestivo de la larva de este díptero. Por tal motivo, en este trabajo se busca ensamblar y anotar el genoma de una proteobacteria perteneciente a la familia Orbaceae, a partir de datos metagenómicos del tracto digestivo

de *H. illucens* en estado larvario para conocer qué genes y rutas metabólicas de estas proteobacterias están involucrados en el proceso de biotransformación de residuos.

1.2. JUSTIFICACIÓN

A nivel mundial, la generación de residuos orgánicos, solo debido a desperdicio y pérdidas de alimentos, está entre la tercera y cuarta parte de la producción destinada al consumo humano, lo que representa cerca de 1,3 billones de toneladas descartadas al año (FAO, 2022a; Salguero & Guevara, 2019). De acuerdo con la ONU, en el 2019, el 17 % de los alimentos disponibles para el consumidor (931 millones de toneladas) terminaron en basureros (ONU, 2021). Esto involucra un derroche económico que, para el 2017, fue calculado en 990 mil millones de dólares, mientras que en el 2019 se estimó que las pérdidas postcosecha, antes de la venta en tiendas, ascendían a los 400 billones de dólares al año (FAO, 2017, 2019, 2022b).

Además del dinero y los alimentos, se debe tener en cuenta que también se pierden los recursos previamente usados para producir, procesar, embalar, transportar y vender dichos alimentos, como por ejemplo mano de obra, tierra, agua, energía e insumos (FAO, 2012; ONU, 2021). La huella de agua azul por la generación de los alimentos que terminaron como residuos es de aproximadamente 250 km^3 de agua, mientras que, el uso de tierra se aproximó a 1400 000 000 ha, equivalente al 30 % de la superficie agrícola global (FAO, 2015; FAO, 2016; HLPE, 2014).

Otros inconvenientes de obtener alimentos que no se consumen y se convierten en residuos es la pérdida del valor agregado y la emisión innecesaria de gases de efecto invernadero que constituyen entre el 8 y 10 % de las emisiones globales (ONU, 2021; PNUMA, 2021). La huella de carbono se calculó en 4,4 billones de toneladas de CO_2, similar al 87 % del calentamiento global que produce el transporte terrestre (FAO, 2015; FAO, 2016; HLPE, 2014). Así mismo, los desperdicios y pérdidas de alimentos provocan una expansión e intensificación de la agricultura y, por lo tanto, afectan negativamente en la biodiversidad, cuya repercusión mundial no es fácilmente estimable (HLPE, 2014; ONU, 2021).

En el Ecuador, los residuos sólidos tienen tres destinos: los botaderos a cielo abierto, las celdas emergentes y los rellenos sanitarios (Figura 1.1). No obstante, la mala

administración de estos espacios puede ocasionar, como ha sucedido en otras ocasiones, un colapso temprano de estos sitios (Morán, 2020). Debido al grave impacto medioambiental, las Naciones Unidas plantearon dentro de los objetivos de desarrollo sostenible (ODS), reducir la generación de residuos hasta el año 2030 (PNUD, 2015), no obstante, en el Ecuador no se ha registrado ningún avance y, por el contrario, tiene una tendencia creciente de producción per cápita de residuos sólidos, pasando de 0,570 kg en el 2014 a 0,83 kg en el 2020 (INEC, 2017, 2021).

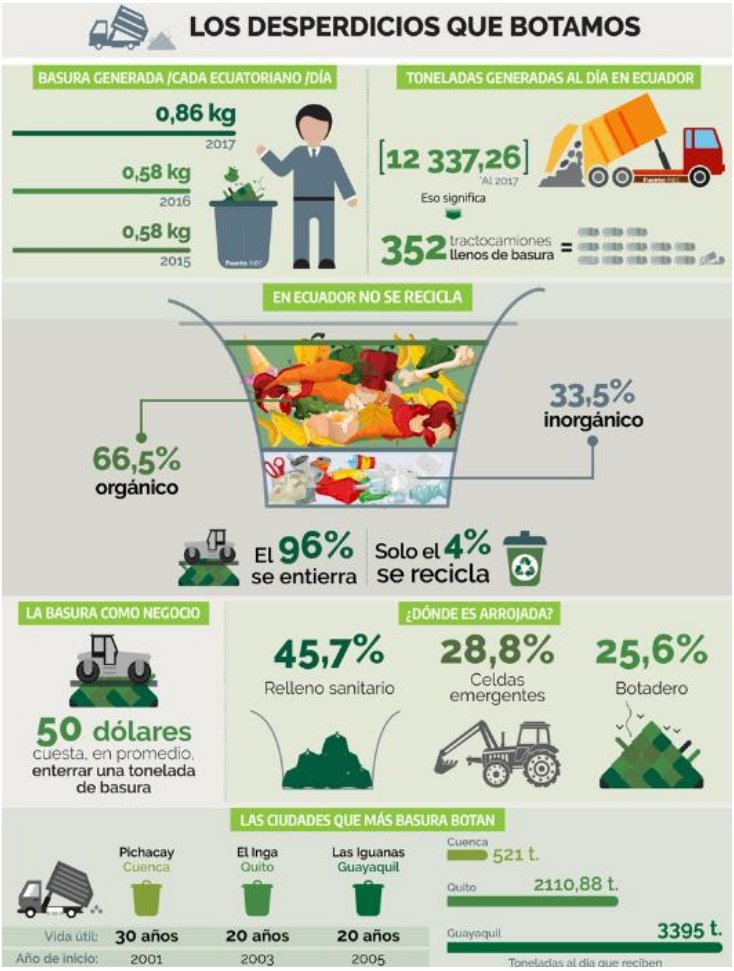


Figura 1.1. Residuos sólidos en el Ecuador (Morán, 2020)

Así, resulta importante encontrar alternativas que permitan una gestión de estos residuos que sea amigable con el ambiente.

La mosca soldado negra, *Hermetia illucens*, es una mosca verdadera, de la clase Insecta, orden Diptera y familia Stratiomyidae (Shelomi et al., 2020). Aunque originalmente es de América, actualmente es posible encontrarla en regiones tropicales y templadas

alrededor del mundo (Čičková, Newton, Lacy, & Kozánek, 2015; Spranghers, Noyez, Schildermans, & De Clercq, 2017). Los adultos solo toman agua, no pican ni se acercan al humano y tampoco transmiten enfermedades (Wang & Shelomi, 2017). Sus larvas se han utilizado en la gestión de residuos a pequeña escala debido a que consumen materia orgánica, como paja de arroz, estiércol, desechos de alimentos, etc (De Smet, Wynants, Cos, & Van Campenhout, 2018; Del Hierro, Anrango, Ortiz, & Sánchez, 2021).

De entre todos los dípteros, *H. illucens* tiene la mayor eficiencia y diversidad de sustratos que pueden procesar (Khamis et al., 2020; Kim et al., 2011; Shelomi et al., 2020; Zhineng et al., 2021). Las larvas también se han estudiado como productos comestibles (Cedeño, 2021; Osimani et al., 2021) y en la producción de biodiesel (Q. Li et al., 2011) pues acumulan lípidos de su dieta para que el adulto, que no se alimenta, los utilice como energía. La materia orgánica que no es consumida, combinada con el excremento de las larvas, rico en nitrógeno, puede usarse como fertilizante (Del Hierro et al., 2021). Su tiempo de desarrollo larvario, que puede superar las tres semanas, es más largo que el de las moscas domésticas y carroñeras (menor a cinco días), lo que implica que una sola larva consumirá una mayor cantidad de sustrato y producirá pupas más grandes (Čičková et al., 2015). Además, en la etapa de prepupa, instintivamente dejan el sustrato y se trasladan a un lugar alto y limpio, un comportamiento llamado "autocosecha" que elimina la necesidad de trasladar a las larvas, lo que requeriría mucha mano de obra para su reproducción. Todos estos beneficios hacen que su crianza sea práctica y se han convertido en una herramienta adecuada para valorizar desechos por su forma de alimentación (Wang & Shelomi, 2017).

Los microorganismos intestinales de los insectos desempeñan un papel crucial en su crecimiento y desarrollo debido a que están involucrados en varias funciones, como la coordinación nutricional, la defensa contra las toxinas de las plantas, respuesta fisiológica, incremento de la esperanza de vida, desintoxicante de alimentos específicos, influencia en el desarrollo y potencial reproductivo, entre otros (Zhineng et al., 2021). Cabe mencionar que, en los intestinos de estos insectos se encuentran bacterias que poseen capacidades metabólicas y juegan un importante rol en la fermentación de carbohidratos complejos. Dado que los huéspedes animales carecen de enzimas para degradar la mayoría de los tipos de carbohidratos, estos nutrientes se dejan digerir por la microbiota intestinal, que, a su vez, libera ácidos grasos de cadena corta (productos de

fermentación) que pueden ser altamente beneficiosos para el metabolismo energético del huésped (Zheng et al., 2016).

Ensamblar genomas microbianos a partir de datos metagenómicos tiene un valor fundamental para comprender la ecología y el metabolismo microbiano pues permite dilucidar el potencial funcional de microorganismos sin necesidad de un cultivo previo (Sangwan, Xia, & Gilbert, 2016). La recuperación metagenómica de genomas bacterianos y arqueales completos o borradores proporciona una ruta para el análisis del potencial de un taxón dentro del contexto de comunidad y ecosistema. Esto permite comprender mejor la adaptación ecológica, las interacciones tróficas y versatilidad metabólica de organismos no cultivados y eco-genéticamente adaptados (Sangwan et al., 2016).

Según Zhineng et al. (2021) existen pocas investigaciones centradas en el análisis de la función biológica de la microbiota intestinal de las larvas de mosca soldado negra y la investigación enfocada en buscar microbiota que mejore la eficiencia en la utilización del alimento es casi nula. Entonces, es necesario desarrollar investigaciones para comprender mejor las características y funciones de la microbiota de las larvas de *H. illucens*, caracterizar especies no descritas del tracto digestivo de la mosca soldado negra y encontrar genes importantes para la biotransformación de la materia orgánica o degradación de plástico. Esto permitiría obtener información relevante para mejorar u optimizar la eficiencia en la biotransformación de los residuos y consecuentemente, se podría tener un impacto positivo para el medio ambiente.

1.3. OBJETIVOS

1.3.1. OBJETIVO GENERAL

Ensamblar y anotar el genoma de una proteobacteria perteneciente a la familia Orbaceae a partir de datos metagenómicos del tracto digestivo de larvas de mosca soldado negra (*Hermetia illucens*)

1.3.2. OBJETIVOS ESPECÍFICOS

- Eliminar lecturas del hospedero (*Hermetia illucens*) de los datos metagenómicos
- Ensamblar el metagenoma a partir de datos metagenómicos libres de lecturas del hospedero
- Seleccionar el genoma ensamblado de una proteobacteria perteneciente a la familia Orbaceae
- Anotar el genoma ensamblado de una proteobacteria perteneciente a la familia Orbaceae
- Buscar genes y rutas metabólicas de interés en el genoma anotado de la proteobacteria

2. REVISIÓN DE LA LITERATURA

2.1 GENÓMICA

En sus inicios, la Biología consistía únicamente en observar la naturaleza y en experimentar sobre partes aisladas de ella. La biología genómica proporcionó un nuevo tipo de información derivada de la naturaleza cuyos datos son igual de complejos que la naturaleza misma. A finales de los 80´s, la genómica hacía referencia a la obtención y análisis de información sobre genes y genomas, siempre y cuando dicha información haya sido generada de manera sistemática, por medio de recursos especializados en proyectos de secuenciación del ADN. A mediados de los 90´s, apareció el término genómica funcional que hacía enfoque en las funciones de los genes. Desde entonces, han aparecido varios términos como proteómica, transcriptómica, metagenómica, etc., motivo por el cual se ha sugerido que todo este conjunto se denomine "ómicas" (Brent, 2000).

2.1.1 TECNOLOGÍAS DE SECUENCIACIÓN

Las complejas preguntas planteadas en torno a la genómica exigen una profundidad de información que supera la capacidad de las tecnologías de secuenciación de ADN tradicionales (Illumina, 2022).

La metodología de secuenciación de ADN de alto rendimiento (secuenciación de próxima generación, NGS por sus siglas en inglés) ha evolucionado velozmente en la última década y continuamente se comercializan nuevos métodos (Slatko, Gardner, & Ausubel, 2018). Las tecnologías de secuenciación de ADN de lectura larga de tercera generación se utilizan cada vez más, proporcionando extensos conjuntos de herramientas genómicas que alguna vez estuvieron reservados para unos pocos organismos modelo seleccionados (Jung et al., 2020). A medida que la tecnología se desarrolla, también incrementan las aplicaciones correspondientes para la ciencia básica y aplicada (Slatko et al., 2018). En la **Figura 2.1** se resumen cuatro diferentes métodos de secuenciación.

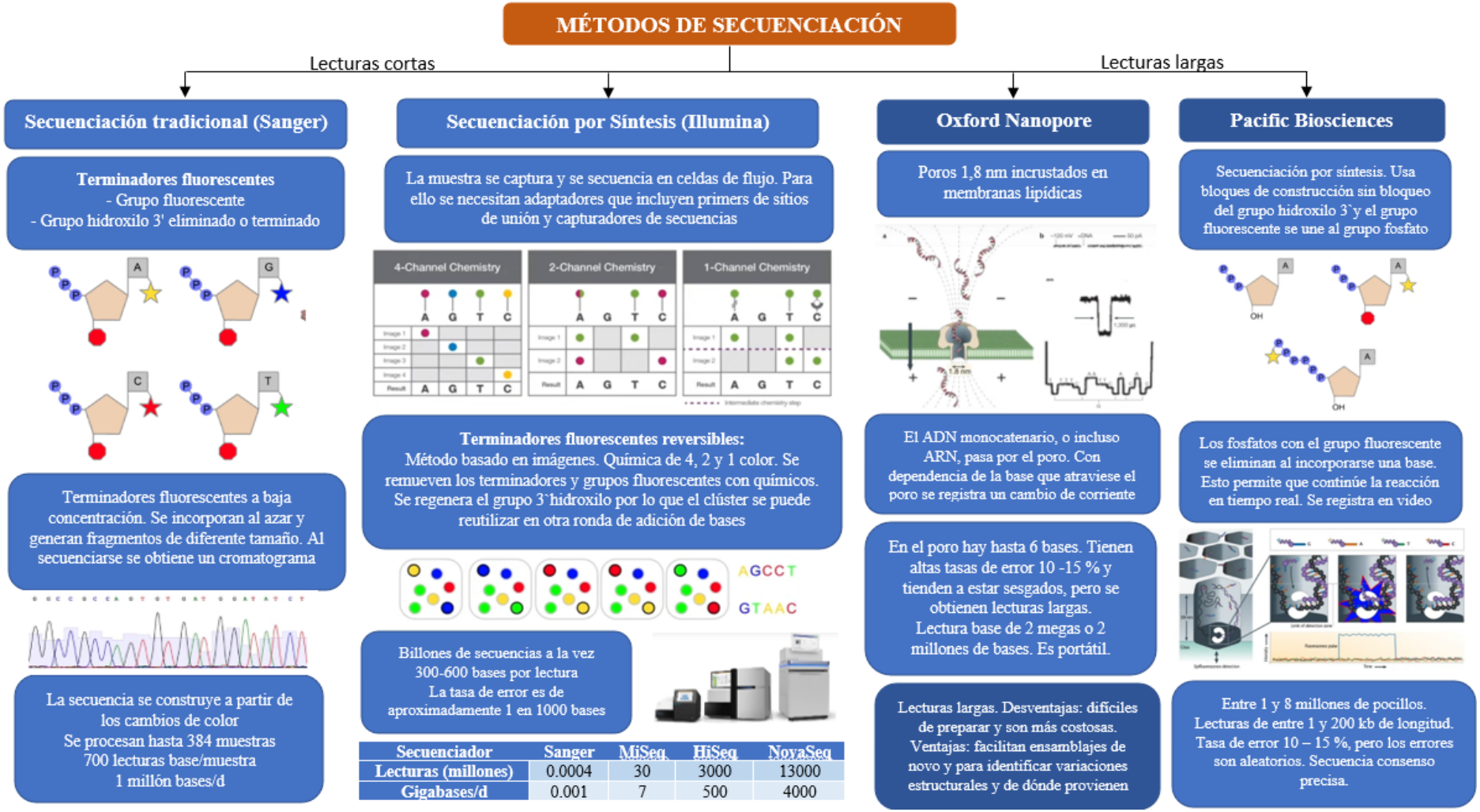


Figura 2.1. Métodos de secuenciación
Imágenes obtenidas de Chow E. (USFC), 2019

2.1.2 ENSAMBLAJE Y ANOTACIÓN DE GENOMAS

La secuenciación del genoma y el ensamblaje de novo, que alguna vez fueron dominio exclusivo de consorcios internacionales bien financiados, se han vuelto cada vez más asequibles, lo que se ajusta a los presupuestos de grupos de investigación individuales (Jung et al., 2020).

El seleccionar las plataformas de software y secuenciación más apropiadas y los pipelines de anotación para un nuevo proyecto de genoma puede ser desalentador porque las herramientas a menudo solo funcionan en contextos limitados. En genómica, la generación de un ensamblaje/anotación de alta calidad es indispensable para comprender mejor la biología de cualquier especie (Jung et al., 2020).

En la **Figura 2.2** se presenta un diagrama del proceso lógico para el ensamblaje de un genoma.

2.2 METAGENÓMICA

Como consecuencia de las mejoras en la eficiencia de la secuenciación del genoma, la investigación del microbioma se ha expandido rápidamente (Breitwieser, Lu, & Salzberg, 2019). El término microbioma se refiere a un microhábitat completo, incluidos sus microorganismos, sus genomas y el entorno que lo rodea (Liu et al., 2021). Para la investigación del microbioma, los enfoques basados en la secuenciación más utilizados son la metataxonómica y la metagenómica. Tanto la metataxonómica como la metagenómica pueden proporcionar información sobre la composición de especies de un microbioma. La metagenómica se refiere a la secuenciación aleatoria de ADN microbiano, sin seleccionar ningún gen en particular (Breitwieser et al., 2019).

La metagenómica "de escopeta", o *shotgun*, es una secuenciación no dirigida de todos los ("meta-") genomas microbianos de una muestra. Esta secuenciación puede utilizarse para establecer un perfil de la composición taxonómica, conocer el potencial funcional de las

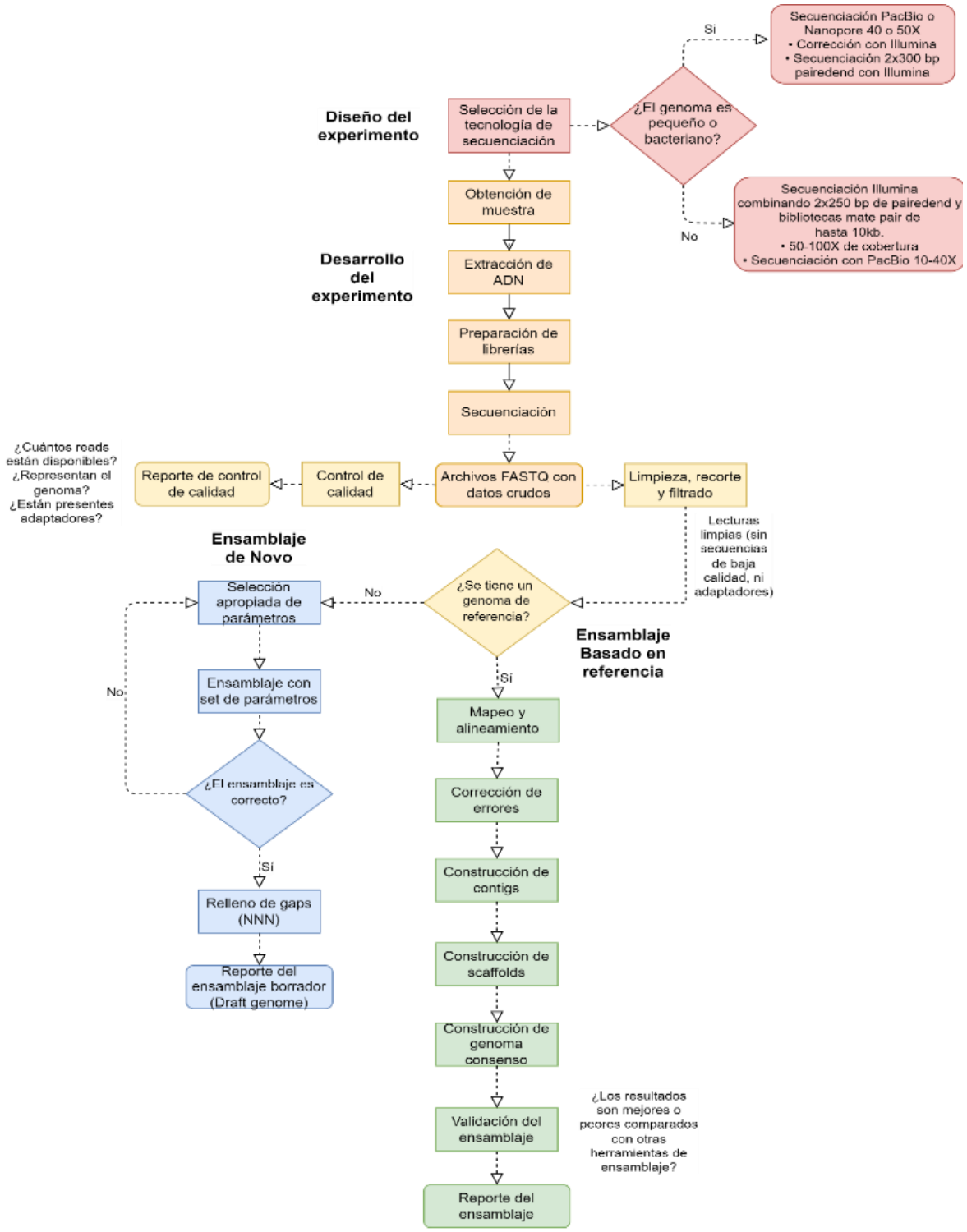


Figura 2.2. Diagrama de flujo para el ensamblaje de un genoma

comunidades de microorganismos, así como para obtener secuencias genómicas completas (Quince, Walker, Simpson, Loman, & Segata, 2017).

De acuerdo con Quince et al. (2017), luego del diseño experimental, el estudio típico de metagenómica *shotgun* consta de cinco pasos, representados en la **Figura 2.3.**

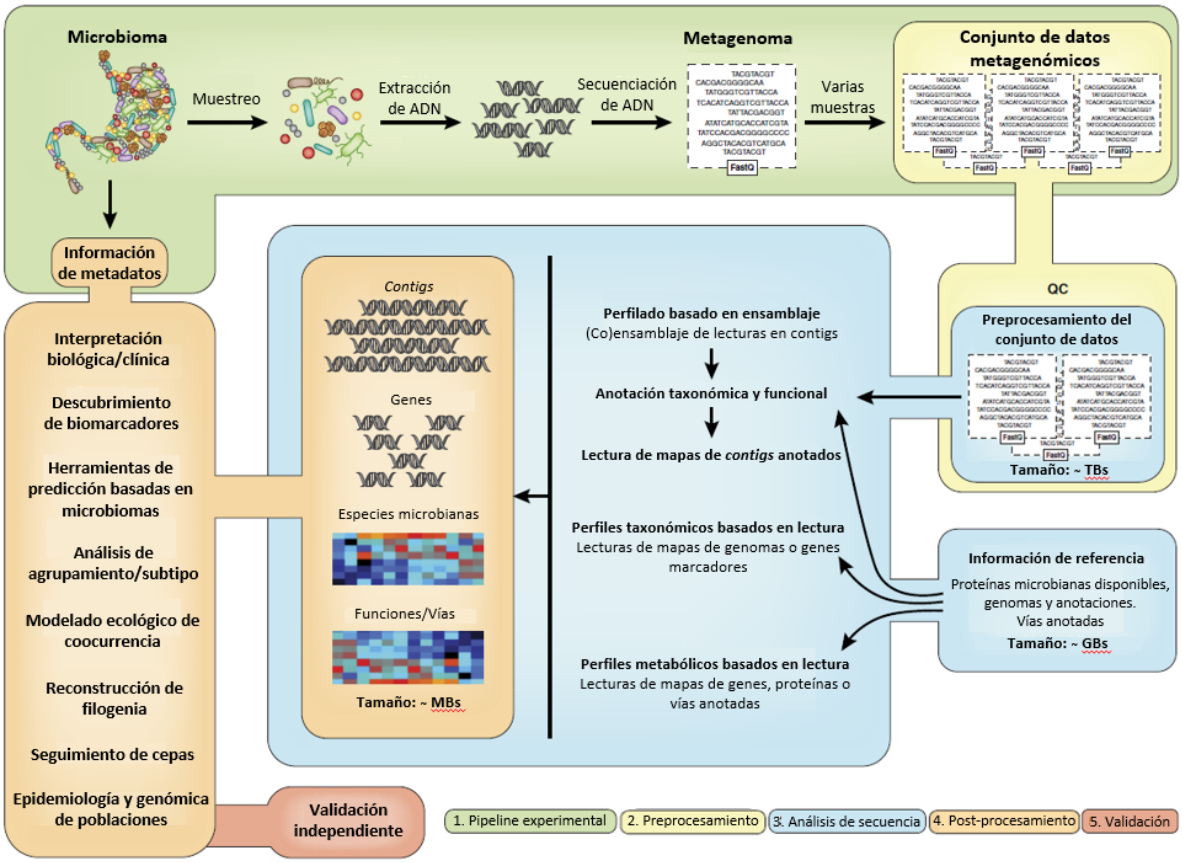


Figura 2.3. Resumen de un flujo de trabajo metagenómico
Modificada de Quince et al. (2017)

2.3 MOSCA SOLDADO NEGRA (*Hermetia illucens*)

La MSN pertenece al orden Diptera, familia Stratiomyidae, género *Hermetia*. Es carroñera y prospera en varios tipos de material orgánico en descomposición. Aunque suele ser una molestia, no tiene piezas bucales funcionales, por lo que no pica ni transmite enfermedades.

Por el contrario, tiene como beneficio el hacer menos adecuados los medios de reproducción de moscas domésticas (Oliveira, Doelle, List, & Reilly, 2015).

Debido a su utilidad y características únicas, esta especie se cría industrialmente y se usa como organismo modelo para la investigación básica. La MSN se distribuye naturalmente en los trópicos y subtrópicos, pero también se puede criar en interiores bajo condiciones controladas; en consecuencia, ahora se distribuye por todo el mundo. Actualmente, optimizar a la MSN para reciclar tipos particulares de desechos es particularmente desafiante porque no se sabe nada sobre su genética, lo que impide el uso de la tecnología molecular para mejorar sus características para el reciclaje de desechos (Zhan et al., 2020).

2.3.1 DESCRIPCIÓN MORFOLÓGICA Y CICLO DE VIDA

La mosca adulta no se alimenta y vive entre cinco y ocho días, período en el cual se aparea. Su ciclo de vida dura aproximadamente entre 40 a 45 días (De Smet et al., 2018) y se representa en la **Figura 2.4**, donde se aprecian cinco etapas: huevo, larva, pre-pupa, pupa y adulto.

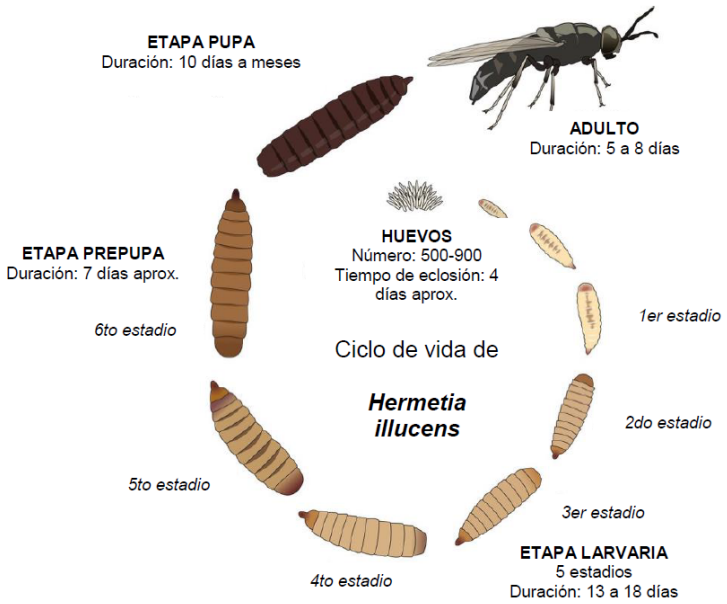


Figura 2.4. Ciclo de vida de *H. illucens* (Pazmiño, 2021)

Para desarrollar el presente trabajo de titulación se utilizarán datos obtenidos a partir de MSN capturadas en Puerto Quito – Ecuador. La identificación del ciclo de vida y la descripción morfológica de esta mosca fueron realizadas previamente por Pazmiño (2021) y se resume en la **Tabla 2.1**.

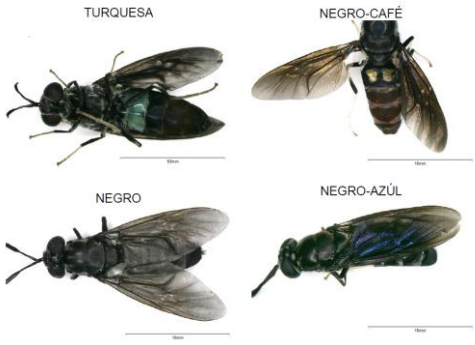
2.3.2 DEGRADACIÓN DE RESIDUOS

La MSN tiene la capacidad de reciclar muchos tipos de desechos orgánicos de manera eficiente y efectiva. A través de este proceso de reciclaje, los flujos de desechos se convierten en productos valiosos, como proteínas para alimentación animal, grasa para bioenergía y compost que se puede utilizar como fertilizante. El uso de MSN para reciclar alimentos y desechos animales tiene muchos otros beneficios. El reciclaje de estos nutrientes da como resultado la reducción de olores nocivos, emisiones de dióxido de carbono, bacterias patógenas y antibióticos (Beskin et al., 2018; De Smet et al., 2018).

A nivel mundial es bien conocido que estas larvas son eficientes en la bioconversión de residuos orgánicos. Es así que, la empresa ESR International desarrolló y patentó un proceso de bioconversión (US patent 6 780 637) que utiliza estas larvas y alcanza una reducción del 95 % en el peso y volumen de residuos alimenticios en pocas horas (Oliver, 2004). Lo interesante de este proceso es que no requiere agua, energía ni productos químicos. Además, es autónomo y no genera efluentes ni gases de efecto invernadero, a excepción de poca cantidad de CO_2 (ESR International, 2016). La unidad de bioconversión está desarrollada para su uso a nivel doméstico, pero puede fabricarse en cualquier tamaño. Por otra parte, continuamente se están desarrollando estudios que evalúan el potencial de estas larvas para su industrialización debido a que su uso se considera un medio rentable y amigable con el ambiente (Beskin et al., 2018; De Smet et al., 2018).

Entre las investigaciones desarrolladas con el tracto digestivo de la mosca soldado negra se encuentran las desarrolladas por Kim et al. (2011) que realizaron una caracterizaron bioquímica de las enzimas digestivas liberadas por la glándula salival y el intestino de la mosca soldado negra (Kim et al., 2011).

Tabla 2.1. Descripción morfológica y ciclo de vida de *H. illucens* (Pazmiño, 2021)

Etapa del ciclo de vida	Imagen	Duración (días)	Descripción
Huevos		5	Ovoidal con tamaño medio de 4 ± 0,5 mm y coloración beige que se torna amarillenta en la incubación.
Larvas	Primer estadío 	1-3	Alargada y aplanada dorsiventralmente, con tegumento endurecido. Su coloración es beige, pero con dependencia del estadio varía su tonalidad. Tiene tres regiones, con 11 segmentos: 8 secciones abdominales, con quetotaxia (distribución de sedas o pelos) característica, tres secciones torácicas y la cápsula cefálica, que se presenta estrecha con boca sin mandíbulas y ojos laterales. En el extremo redondeado se distingue el ano en la parte ventral. Al alcanzar 20 mm de longitud inician su transformación a pre-pupa, definido por un cambio de coloración a una tonalidad café oscura y en su comportamiento.
	Segundo estadío 	5	
	Tercer estadío 	13- 15	
	Cuarto estadío 	>20	
	Quinto estadío (pre-pupa) 	>20	
Pupa		-	Coloración café oscura. Requiere menos alimento y prefiere zonas calientes y secas.
Adulto		-	Longitud media: 21,1 mm. Cuatro coloraciones. Sin mandíbulas. Tiene manchas traslúcidas en el segundo segmento abdominal. Posee dimorfismo sexual.

En cuanto a microorganismos del tracto digestivo, está el estudio de Shelomi et al. (2020) en el que se analizó la microbiota bacteriana de larvas criadas en dos tipos de desechos alimentarios (pulpa de soja de posproducción y desechos de cafetería posconsumo), y se encontró que las larvas tienen una microbiota conservada cuyos componentes varían geográficamente (Shelomi et al., 2020).

Osimani et al. (2021) estudiaron la diversidad microbiana del sustrato de cría, excremento y larvas de *H. illucens* alimentadas con un subproducto del tostado del café (Cs) suplementado con diferentes cantidades de microalgas (*Schizochytrium limacinum* o *Isochrysis galbana*). La microbiota predominante de las larvas alimentadas con Cs fue *Paenibacillus*, en las alimentadas con dietas que contenían *I. galbana*, según el nivel de inclusión de algas, predominó *Enterococcus, Lysinibacillus, Morganella* y *Paenibacillus*, mientras que las alimentadas con dietas que contenían *S. limacinum* se caracterizaron por una abundancia relativa alta de *Brevundimonas, Enterococcus, Paracoccus* y *Paenibacillus*, con dependencia del nivel de inclusión de algas. Sus resultados indicaron que el tipo de alimentación influye en la abundancia relativa de los microorganismos presentes y sobre la base de estos resultados se planteó la posibilidad de tener un mayor control de las especies microbianas presentes en las larvas (Osimani et al., 2021).

Bruno et al. (2019) estudiaron la microbiota del intestino medio de las larvas, mediante secuenciación del gen 16S y demostraron que las comunidades bacterianas difieren en cantidad y composición de acuerdo con la dieta y región del intestino. Los principales grupos taxonómicos bacterianos encontrados a nivel de filo en este estudio fueron: *Bacteroidetes, Firmucutes, Proteobacteria, Fusobacteria, Tenericutes y Actinobacteria*, mientras que a nivel de género se presentan en la Figura 2.5. La microbiota de la parte anterior mostró la mayor diversidad microbiana, que disminuyó gradualmente a lo largo del intestino medio, mientras que la carga bacteriana tuvo una tendencia opuesta, siendo máxima en la región posterior. Además, mostraron que la influencia del contenido microbiano de la dieta se limitaba a la parte anterior del intestino medio y que la actividad de alimentación de las larvas no afectaba significativamente a la microbiota del sustrato (Bruno et al., 2019).

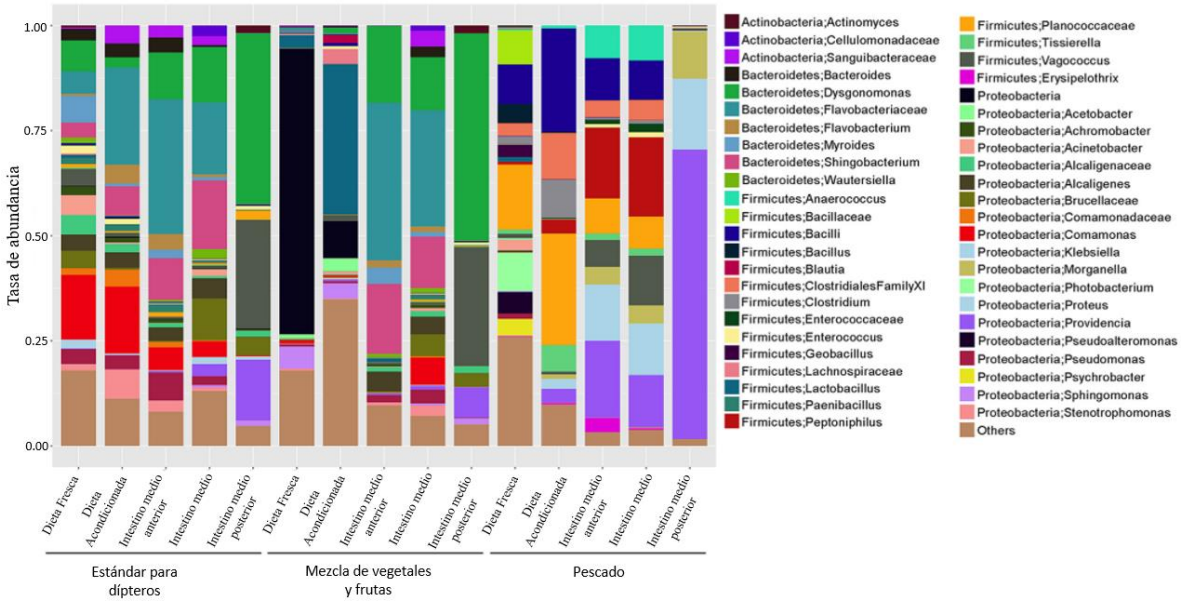


Figura 2.5. Abundancia relativa de géneros bacterianos identificados en las muestras de intestino medio de *H. illucens* y tres dietas (Bruno et al., 2019)

De Smet *et al*.(2018) realizaron una revisión acerca de la relación entre la comunidad de microorganismos con el desempeño y salud del insecto. En la revisión, se encuentra información sobre la caracterización de la microbiota y la dinámica dependiente de la composición del sustrato. Además, se enfoca en las interacciones entre los microorganismos y los insectos durante su desarrollo. Y, se discute cómo el conocimiento del microbioma puede desencadenar estrategias para: optimizar la obtención de biomasa de insectos por medio de la manipulación de la microbiota, el descubrimiento de nuevas enzimas, por medio de su interacción, y la obtención de antimicrobianos con base en la inmunidad, empleando organismos enteros o sus moléculas (De Smet et al., 2018).

Khamis *et al*. (2020) estudiaron la variabilidad genética y el complejo de microbiomas intestinales de poblaciones de MSN domesticadas y recolectadas en el medio silvestre de seis continentes utilizando el gen COI mitocondrial y la metagenómica 16S. Las secuencias generadas a partir del estudio se vincularon con las accesiones de *H. illucens* KC192965.1, KM967419.1, KY817115.1, FJ794367.1, FJ794361.1 y FJ794355.1 de GenBank. Las distancias genéticas entre las muestras de *H. illucens* del estudio y las de las accesiones de GenBank oscilaron entre 0,0091 y 0,0407, mientras que las muestras de *H. sexmaculata* y *H. albitarsis* se separaron claramente de todos los *H. illucens* por distancias de 0,1745 y 0,1903,

respectivamente. Los hallazgos sobre la diversidad genética revelaron una ligera variación filogeográfica entre las poblaciones de MSN en todo el mundo. Los datos de 16S representaron familias de bacterias intestinales larvales con géneros económicamente importantes que podrían presentar riesgos para la salud tanto para animales como para humanos por lo que recomiendan el tratamiento previo de las materias primas y medidas posteriores a la cosecha de las larvas para minimizar el riesgo de contaminación por patógenos a lo largo de la cadena de valor de alimentos basados en insectos (Khamis et al., 2020).

En cuanto a análisis de función biológica de la microbiota intestinal, Zhineng *et al.* (2021) estudiaron la microbiota intestinal presente en las larvas de mosca soldado negra para determinar su tipo y características funcionales. Para ello seleccionaron larvas en cuarto a quinto estadio y les ofrecieron diferente alimentación. Sus resultados indicaron que el intestino de las larvas contiene más de 2 300 géneros y 11 000 especies de bacterias y que la mayor abundancia relativa a nivel de filo correspondió a Proteobacteria, Bacteroidetes, Firmicutes, y Actinobacteria; mientras que, la mayor abundancia a nivel de especie incluyó Enterococcus, Acinetobacter, Providencia, Enterobacter y Myroides. Entre esas bacterias, se consideró que Firmicutes desempeñaba un rol importante en la digestión del estiércol animal y que Bacteroidetes degrada materia orgánica de alto peso molecular. Además, acotaron que la presencia de Proteobacterias es una posible fuente de enfermedad microbiana y, por lo tanto, se deben tener precauciones al utilizarlas con fines alimenticios. También encontraron que la alimentación de las larvas influye en el tipo de microorganismos presentes y, consecuentemente, en las funciones de la microbiota. Así, los microorganismos presentes en larvas alimentadas con 75 % de salvado de trigo y 25 % de soya en polvo pueden promover la eficiencia en la utilización del alimento; mientras que, si las larvas se alimentan con 75 % de salvado de trigo, 25 % de soya en polvo suplementado con 1 % de tetraciclina los microorganismos presentes pueden desempeñar un papel importante en la supervivencia de las larvas en ambientes hostiles (Zhineng et al., 2021). Estos autores no abordaron completamente la relación entre los microorganismos y su función, pero mencionaron que tienen investigaciones en marcha para poder hacerlo.

En Ecuador, las investigaciones realizadas hasta el momento con *H. illucens* están enfocadas a la simulación de su ciclo de vida (Romero, 2022), caracterización y reproducción del insecto con fines de biotransformación de materia orgánica (Del Hierro et al., 2021; Morales, 2021) y como fuente alimenticia (Cedeño, 2021; Holalla, 2021; Sumba, 2016). Recientemente, en el año 2021 se obtuvo por primera vez información genética de la mosca soldado negra. Pazmiño (2021), informó sobre la variabilidad genética existente entre poblaciones con marcadores moleculares: ADN de la subunidad ribosomal grande 28S (28S rADN), espaciador transcrito interno 2 (ITS2) y gen citocromo c oxidasa mitocondrial I (COI). Además, estudió el efecto que tiene la introducción de un 5 % de microplásticos en la dieta larvaria sobre la generación de biomasa y el crecimiento. En esta investigación se encontró que existe una disminución de la producción de biomasa en larvas alimentadas con polietileno; sin embargo, no se obtuvo una diferencia estadísticamente significativa en la producción de biomasa de larvas alimentadas con fundas de poliestireno, mientras que hubo un incremento en la biomasa de larvas alimentadas con material de fundas de ácido poliláctico. En cuanto a la tasa de crecimiento de las larvas, no se presentó una diferencia estadísticamente significativa para ninguno de los tratamientos empleados (Pazmiño, 2021).

2.4 PROTEOBACTERIAS DE LA FAMILIA ORBACEAE

Dentro del dominio de las bacterias se tiene el filo Proteobacteria cuyo nombre proviene de las palabras griegas *Proteus*, dios del océano con la capacidad de cambiar de forma, y *bakterion*, varilla pequeña. A este filo pertenecen las bacterias Gramnegativas relacionadas con las de los miembros del orden *Pseudomonadales* de acuerdo con secuencias del gen 16S ARNr (Garrity, Bell, & Lilburn, 2015b). De acuerdo con el manual de Bergey´s, sobre la base del análisis filogenético de las secuencias de dicho gen, este filo contiene cinco clases:

- Alphaproteobacteria
- Betaproteobacteria
- Gammaproteobacteria
- Deltaproteobacteria
- Epsilonproteobacteria

Gamma corresponde a la tercera letra del alfabeto griego y por tanto, Gammaproteobacteria representa la tercera clase de las proteobacterias, dentro de la cual se encuentra el orden Orbales con su familia Orbaceae (Garrity, Bell, & Lilburn, 2015a).

Hasta enero de 2023, el manual de Bergey indicaba que esta familia tiene tres géneros: *Frischella, Gilliamella* y *Orbus;* no obstante, según la Lista de nombres procarióticos con posición en la nomenclatura (cuyas siglas en inglés son LPSN), existe un cuarto género, el *Zophobihabitans* (LPSN, 2021).

La especie tipo de este último género es *Z. entericus*, también denominada IPMB12^T (=KACC 22323^T =KCTC 82347^T =LMG 32079^T =BCRC 80908^T) aislada de un supergusano (*Zophobas morio*). Las características de esta cepa son las siguientes: anaerobia facultativa, inmóvil, cocoide o en forma de bastón y de colonias translúcidas, con oxidasa positiva, catalasa y β-galactosidasa negativas, no puede hidrolizar urea, pero tiene capacidad para reducir el nitrato a nitrito, fermentar glucosa para producir ácido e hidrolizar esculina. El perfil lipídico polar consta de fosfatidiletanolamina (PE), fosfatidilglicerol (PG), difosfatidilglicerol, un fosfoaminoglicolípido no caracterizado y un aminofosfolípido no caracterizado (APL). El contenido de ADN genómico G+C es del 39,3 % en moles. La única quinona respiratoria identificada fue la ubiquinona-8 (Q-8), como en todos los miembros conocidos de esta familia (Kuo et al., 2021).

El género *Orbus* incluye dos especies: *O. sasakiae,* aislada del intestino de mariposa y *O. hercynius* obtenida a partir de heces de jabalí. Estas bacterias pueden presentarse como cocos con un diámetro de 0,5 – 1,0 μm o como bastoncillos cortos de 1,0 - 1,5 μm de longitud y 0,8 μm de ancho; además, son mesófilas, sin motilidad dependiente de flagelos, quimioheterótrofas con metabolismo aeróbico y anaeróbico facultativo (Wilharm, 2019). En la Tabla 2.2 se presentan diferencias entre ambas especies. Es importante señalar que su descripción se basa en un único aislado en cada caso, por lo que actualmente se desconoce la variabilidad de rasgos en ambas especies (Wilharm, 2019).

Tabla 2.2. Diferencias entre Orbus hercynius y Orbus sasakiae

(Wilharm, 2019)

Características	*O. hercynius*	*O. sasakiae*
Morfología celular	Bastoncillos cortos y cocos	Cocos
Oxidasa	+	-
Ureasa	+	-
B-Galactosiasa	-	+
Esterasa	-	+
N-acetil-β-glucosaminidasa	-	+
Asimilación de threalosa	+	-
Asimilación de celobiosa, L-fucosa, D-galactosa, α-lactosa, lactulosa, melibiosa, β-metil-D-glucósido, D-psicosa, rafinosa, ácido cítrico, ácido α-cetobutírico, ácido bromosuccínico, L-treonina, uridina, α-D-glucosa 1-fosfato, dextrina, Tweens 40 y 80 y N-acetil-D-galactosamina	-	+
Lípidos polares	PE, PG y 2 APL	PG, PE, 1 PL y 2 APL
Contenido de G+C en el ADN (% mol)	36,4* / 38,8	32,1

PL fosfolípido no identificado
*(Volkmann et al., 2010)

2.5 RUTAS METABÓLICAS ASOCIADAS CON LA BIOCONVERSIÓN DE RESIDUOS ORGÁNICOS

El tracto gastrointestinal de *H. illucens* se puede dividir en tres partes: el intestino anterior, medio y posterior; y, puede albergar diferentes comunidades bacterianas. El hecho de que las larvas de MSN (LMSN) puedan digerir una amplia gama de sustratos orgánicos, incluso en descomposición, plantea interrogantes sobre el papel que juega el microbioma del intestino en esta actividad y los mecanismos asociados (De Smet et al., 2018).

Así, Jiang *et* al. (2019), por medio de análisis de ARNr 16S, estudiaron cómo el microbioma intestinal de las LMSN puede influir en la comunidad microbiana del compostaje de desechos alimenticios y promover la degradación de la materia orgánica. Los resultados indicaron que, la comunidad microbiana de las muestras del sustrato que contenía las larvas se volvía más similar, con el tiempo, a la presente en el intestino de las LMSN; y a su vez, mayor cantidad de bacterias encontradas en los desechos alimenticios colonizaban el intestino de las LMSN (Jiang et al., 2019). Lo que permite explicar por qué en los estudios que alimentan a las LMSN con diferentes dietas se presentan cambios en la estructura de las bacterias intestinales (De Smet et al., 2018; Jiang et al., 2019).

A continuación, se resumen otros hallazgos importantes de ese estudio.

En la **Figura 2.6** se presenta la abundancia relativa de grupos de funciones metabólicas presentes en la microbiota de los intestinos de MSN (Jiang et al., 2019).

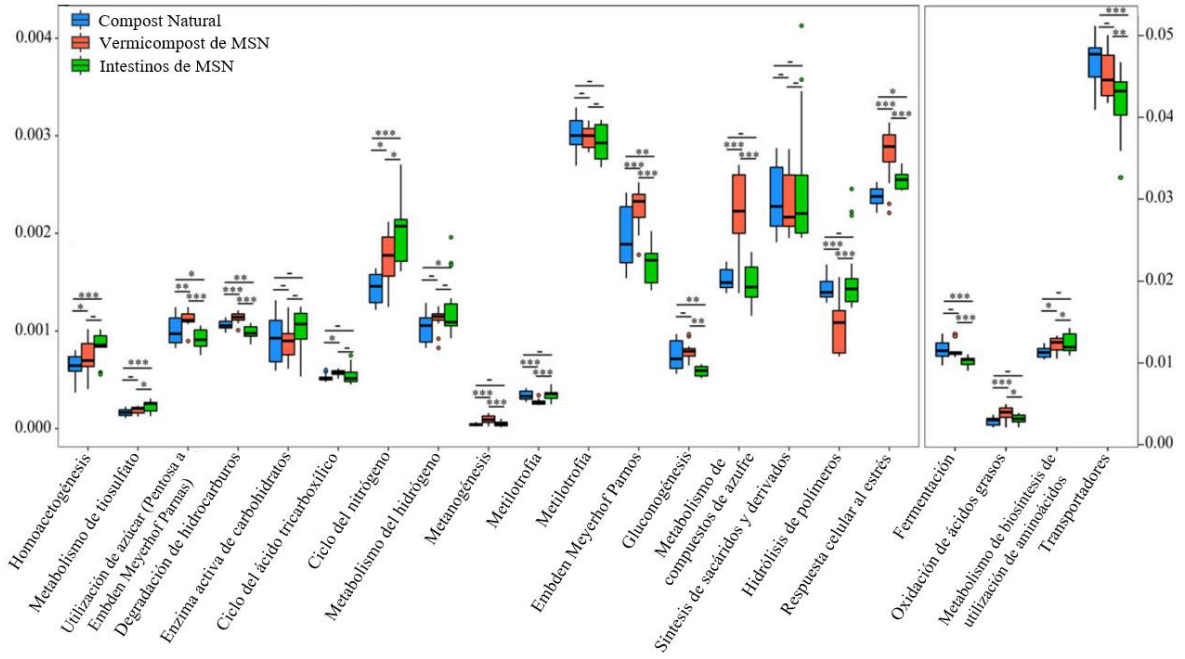


Figura 2.6. Diagramas de caja y bigotes de la abundancia relativa de grupos funcionales con base en la base de datos Asignaciones de Ontologías Funcionales para Metagenomas (FOAM por sus siglas en inglés).
-p>0.05; *p<0.05; **p<0.01; ***p<0.001
Modificada de Jiang et al. (2019)

Se determinó que, en presencia de las LMSN se pueden fortalecer 11 grupos de funciones metabólicas de la microbiota presente en el compost donde estaban presentes las larvas, incluidas las enzimas activas de carbohidratos, el ciclo del ácido tricarboxílico, el metabolismo del hidrógeno, Embden-Meyer-hof-Parnas, el metabolismo de los compuestos de azufre, la homoacetogénesis y la utilización de azúcar. Sin embargo, la metanogénesis y la hidrólisis de polímeros en el microbioma de esa muestra se debilitaron significativamente (P <0.05) en comparación con las del compost en ausencia de las LMSN.

Para mostrar las diferencias en las funciones metabólicas de elementos químicos, los investigadores identificaron genes funcionales asociados con los metabolismos de C, N y S,

así como tres tipos de enzimas, según la base de datos KEGG, y conservaron los genes más abundantes (**Figura 2.7**). Salvo excepciones, de manera general, encontraron que los genes seleccionados y las funciones metabólicas asociadas eran más altas en el sustrato que contenía LMSN y en el intestino de MSN que en el mismo sustrato compostado en ausencia de las larvas (Jiang et al., 2019).

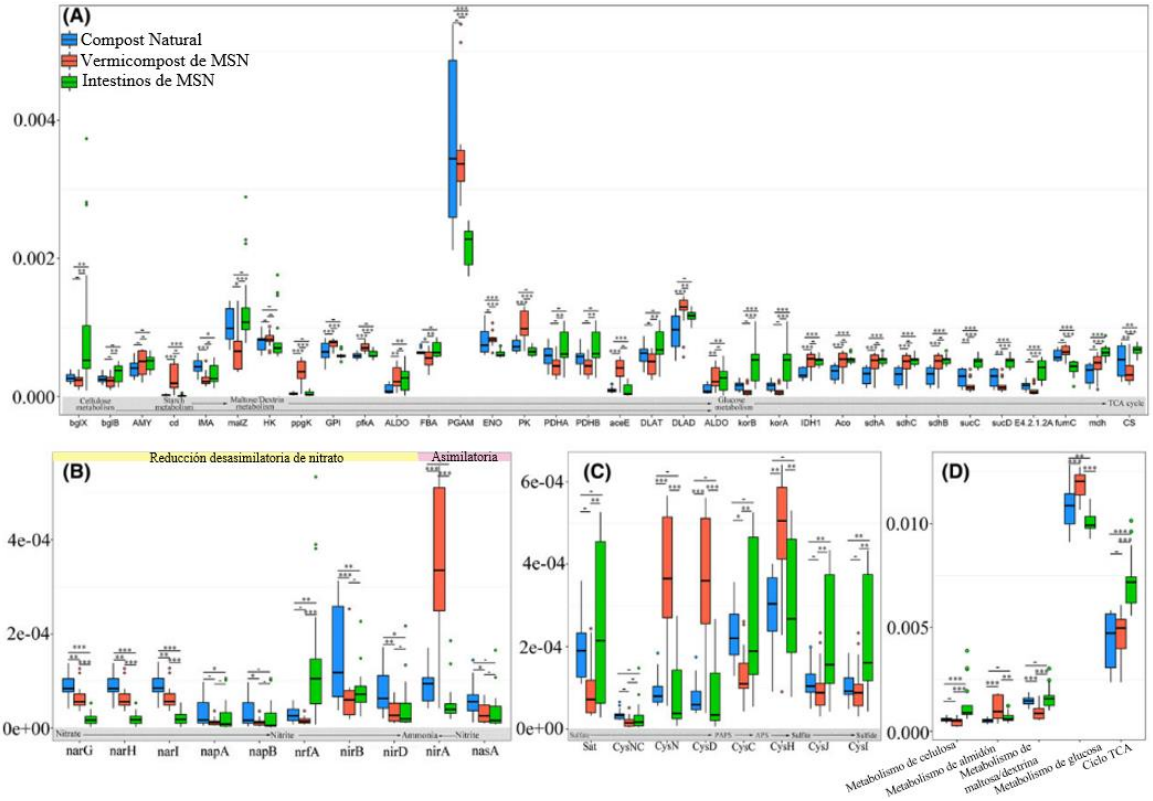


Figura 2.7. Diagramas de caja y bigotes de la abundancia relativa de genes con base en KEGG. Genes funcionales relacionados con el metabolismo de: (A) carbono; (B) nitrógeno; (C) azufre; (D) Genes relacionados con diferentes pasos del metabolismo de carbono. ¯p>0.05; *p<0.05; **p<0.01; ***p<0.001
Modificada de Jiang et al. (2019)

Para evaluar las posibles relaciones entre el género bacteriano más abundante y los genes de función metabólica, calcularon las correlaciones de rango de Spearman (**Figura 2.8**) y retuvieron los genes que estaban significativamente relacionados (P < 0,05) con el género. La mayor abundancia relativa de genes en el intestino de la MSN, incluidos Sat, CysC, malZ, korAB y sucCD, se correlacionó positivamente con Dysgonomonas, Ureibacillus, Enterococcus y RsaHF231. Dysgonomonas también se correlacionó positivamente con nrfA,

bglX, mdh y CS. Las bacterias pertenecientes a Bacillaceae se correlacionaron positivamente con CS, E4.2.1.2A, sucCD, korAB, IMA, cysC y Sat; mientras que, Bacillus fue la bacteria predominante en todas las muestras y tuvo la mayor abundancia en las muestras de intestino (Jiang et al., 2019).

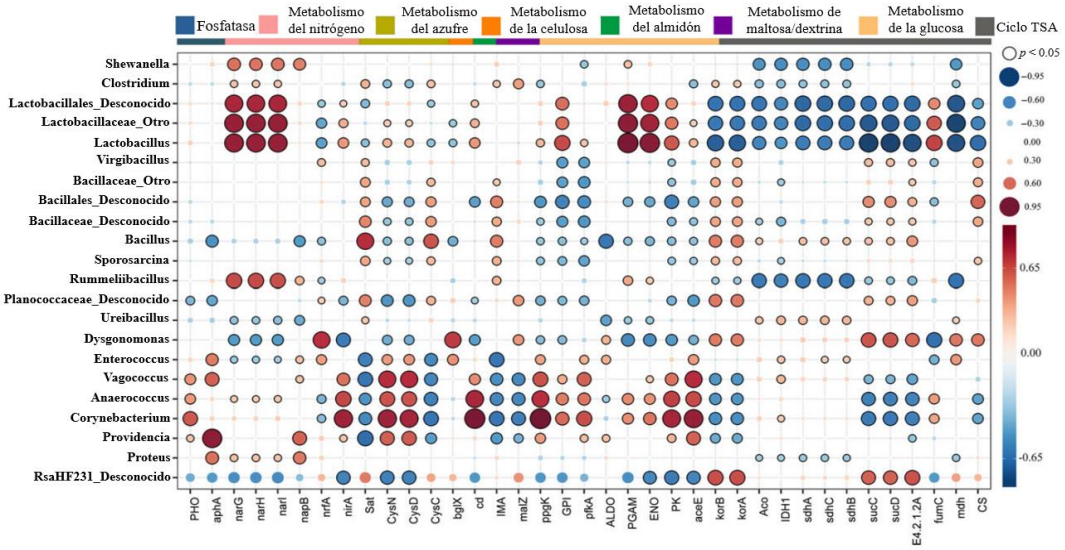


Figura 2.8. Correlaciones de rango de Spearman entre genes funcionales (filas) y géneros (columnas) en todos los grupos.

Los colores rojo y azul representan correlaciones positivas y negativas, respectivamente. El tamaño del círculo y el color de saturación son proporcionales a la magnitud de la correlación. Las correlaciones estadísticamente significativas ($p<0.05$) están indicadas por el perímetro en color negro. Modificada de Jiang et al. (2019)

En consecuencia, los resultados de Jiang et al. (2019) indicaron que los microorganismos más abundantes tuvieron correlaciones positivas con los genes asociados y los grupos funcionales metabólicos más abundantes, como enzimas activas de carbohidratos, metabolismo del hidrógeno, ciclo del nitrógeno y metabolismo de los compuestos de azufre. Además, que el cambio significativo de bacterias en el sustrato debido a la transformación por parte de las larvas provocó un aumento en la abundancia de varios genes metabólicos (Jiang et al., 2019).

Este estudio puede dar una idea de las rutas metabólicas importantes en la degradación del material orgánico. No obstante, se debe tener en cuenta que las funciones metabólicas están estrechamente relacionadas con las características de las bacterias presentes y esto a su vez con su dieta. En otras palabras, la composición del sustrato está asociado a la abundancia relativa de las vías metabólicas microbianas (De Smet et al., 2018; Jiang et al., 2019).

3. METODOLOGÍA

Para el desarrollo del proyecto se utilizaron archivos FASTQ fruto de la secuenciación previa, de larvas de mosca soldado negra alimentadas con una mezcla que contenía 5 % de poliestireno y 95 % de banano. Se partió de dos archivos FASTQ (uno *forward* y uno *reverse*) obtenidos con la tecnología de Illumina, en un equipo NovaSeq. La muestra utilizada tuvo un output de 12 Gb (Pazmiño, 2021).

3.1. ELIMINACIÓN DE LAS LECTURAS DEL HOSPEDERO

La eliminación de las lecturas del hospedero de los datos metagenómicos se llevó a cabo con el flujo de trabajo de la Figura 3.1. Para ello se utilizó Linux, por medio del clúster de CEDIA (PuTTY). El código de programación empleado para el efecto se presenta en el Anexo 1. Los parámetros se fijaron con base en el protocolo de análisis metagenómico del Laboratorio de diagnóstico molecular IDgen.

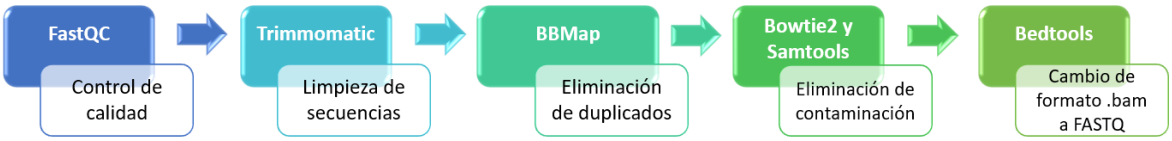


Figura 3.1. Flujo de trabajo para la eliminación de las secuencias del hospedero

3.1.1. CONTROL DE CALIDAD

El control de calidad de los datos de secuenciación se realizó con la herramienta *FastQC* v0.11.9 (Babraham Institute, 2020). Para esta herramienta no se especificó ningún parámetro pues el único requisito para su ejecución fue seleccionar el objeto de entrada que en este caso fueron los archivos tanto de la secuencia *forward* como de la secuencia *reverse*. *FastQC* v0.11.9 generó un informe, con una serie de tablas y gráficas por módulo, cuyo análisis se presenta en la **sección 4.1.1**.

3.1.2. LIMPIEZA DE SECUENCIAS Y RE-EVALUACIÓN DE LA CALIDAD

Para mejorar la calidad de las secuencias, se realizó un recorte con la herramienta *Trimmomatic* v0.39 (Bolger, Lohse, & Usadel, 2014) donde las lecturas se truncaron en función de su calidad promedio. Para ello, se trabajó con una codificación de calidad base *phred33*, con 8 subprocesos para mejorar el rendimiento del equipo, aunque si este parámetro no se especifica se elige automáticamente. Además, se especificó *trimlog logfile* para crear un registro de los recortes. En caso de que existan adaptadores, se utilizó *Illuminaclip* con una lista de adaptadores *TruSeq3*.

Los parámetros más importantes fueron los siguientes: Slidingwindow: 5:20, que realizó un recorte de ventana deslizante, cortando una vez que la calidad promedio de la ventana caía por debajo del umbral. Como se puede notar, el tamaño de la ventana se especificó en 5 pb que indica el número de bases para promediar, mientras que la calidad promedio requerida se fijó en 20. Leading: 5, eliminó las bases que tenían un valor de calidad inferior al umbral desde el inicio. Trailing:5, eliminó las bases con un valor inferior al umbral desde el final. Por último, Minlen:50 eliminó las lecturas con un valor menor a la longitud mínima especificada.

Una vez eliminadas las lecturas de baja calidad se procedió a re-evaluar la calidad de las secuencias con *FastQC* v0.11.9 para visualizar los resultados de la limpieza. Dichos resultados se presentan en la **sección 4.1.2.**

3.1.3. ELIMINACIÓN DE LECTURAS DUPLICADAS

La eliminación de lecturas duplicadas a partir de los archivos FASTQ se llevó a cabo con la herramienta *Dedupe*, del paquete *BBMap* v38.34. Esta herramienta facilita el ensamblaje posterior debido a que elimina las subsecuencias que comparten un porcentaje de identidad objetivo, consideradas como duplicadas (SourceForge, 2022). Los objetos de entrada para esta herramienta fueron los objetos de salida con lecturas emparejadas de la herramienta

Trimmomatic v0.39. Como salida se nombra un objeto sin las lecturas duplicadas (*out*) y otro objeto solo con las lecturas duplicadas eliminadas (*outd*).
Otra herramienta utilizada dentro del mismo paquete de *BBMap* v38.34 fue *Reformat* (SourceForge, 2022) para reformatear las lecturas en formato FASTQ y de esta manera obtener dos archivos sin intercalar con las lecturas sin duplicar como el 100 %.

3.1.4. ELIMINACIÓN DE LA CONTAMINACIÓN

Una vez eliminadas las lecturas duplicadas, se procedió a eliminar las lecturas del hospedero. Para ello se buscó y se descargó de NCBI el genoma de referencia de *H. illucens* en formato FASTA. A partir del genoma de referencia se creó una base de datos con la herramienta *Bowtie2* v2.4.2 (Johns Hopkins University, 2022; Langmead & Salzberg, 2012).

Posteriormente, se llevó a cabo la alineación y mapeo de las lecturas sobre el genoma de referencia con la misma herramienta. Los objetos de entrada fueron: la base de datos del genoma del hospedero, los archivos obtenidos con *Reformat* y se especificó como salida un archivo .sam (-S).

Luego, se cambió el formato del archivo generado a .bam y se filtraron y ordenaron las secuencias no alineadas con la herramienta *samtools* v1.10-3 (Danecek et al., 2021; SourceForge, 2012).

Se procedió a dividir las lecturas en archivos *paired end* para las secuencias *forward* y *reverse* con la herramienta *bedtools* v.2.25.0 (Quinlan, 2021; Quinlan & Kindlon, 2022). Estos archivos se utilizaron para crear el objeto *paired end library* para continuar su procesamiento en KBase como se indica en el apartado 3.2.

3.2. ENSAMBLAJE Y ANOTACIÓN DEL METAGENOMA

Una vez importados los datos en KBase, para el ensamblaje y anotación del metagenoma se continuó con el flujo de trabajo que se presenta en la Figura 3.2. Los parámetros utilizados

fueron los recomendados por Chivian, Clarck y Jungbluth (2020) en su tutorial: "*Metagenome-Assembled Genome Extraction from a Compost Microbiome Enrichment*" (Chivian, Clark, & Jungbluth, 2020).

3.2.1. CLASIFICACIÓN TAXONÓMICA (*KAIJU*)

El flujo de trabajo empezó con la herramienta *Kaiju* v1.7.3 (Menzel, Ng, & Krogh, 2016) para predecir la composición microbiana de la muestra, en todos los niveles taxonómicos (filo, clase, orden, familia, género y especie), en función de las similitudes de proteínas. Se utilizo la base de datos de NCBI RefSeq (no Euks), con un filtro de baja abundancia de 0,1 para mostrar solo los géneros que comprendan al menos este porcentaje del total de lecturas y un porcentaje de submuestra de 10 de cada conjunto de datos para una rápida ejecución.

3.2.2. ENSAMBLAJE DEL METAGENOMA Y COMPARACIÓN DE *CONTIGS*

El ensamblaje del metagenoma se llevó a cabo con tres diferentes herramientas, a partir del objeto *paired end library*, en formato Fastq, creado previamente. Las herramientas empleadas fueron: *metaSPAdes* v3.15.3 (Nurk, Meleshko, Korobeynikov, & Pevzner, 2017), *IDBA-UD* v1.1.3 (Peng, Leung, Yiu, & Chin, 2012) y *MEGAHIT* v1.2.9 (D. Li, Liu, Ruibang, Sadakane, & Lam, 2015). Para las tres herramientas, la longitud mínima de *contig* seleccionada fue de 2000.

Posteriormente, se utilizó *Compare Assembled Contig Distributions* v1.1.2 (KBase, 2019) con el fin de obtener un resumen comparativo de los ensamblajes llevados a cabo con las tres herramientas antes mencionadas y seleccionar la que permita un mejor ensamblaje. Para ello se utilizaron los tres objetos resultantes del ensamblaje de cada herramienta.

3.2.3. AGRUPAMIENTO DE CONTIGS Y OPTIMIZACIÓN DE *BINS*

El agrupamiento de los *contigs* metagenómicos ensamblados en *bins* (linajes) utilizando la profundidad de cobertura, la composición de nucleótidos y los genes marcadores se llevó a

cabo con las herramientas: *MaxBin2* v2.2.4 (Wu, Simmons, & Singer, 2016), *MetaBAT2* v1.7 (Kang, Froula, Egan, & Wang, 2015) y *CONCOCT* v1.1. (Alneberg et al., 2014). Los objetos de entrada para las tres herramientas fueron el objeto resultante del ensamblaje seleccionado y el objeto *paired end library*.

Para las tres herramientas se trabajó con los valores preestablecidos de los diferentes parámetros:

En *MaxBin2* v2.2.4 se empleó con un conjunto de 107 marcadores que corresponden principalmente a linajes bacterianos y un umbral de probabilidad de 0.8 que indica la confianza que debe cumplir el algoritmo de maximización de expectativas para que un *contig* se agrupe con un *bin*. Los *contigs* con valores por debajo del umbral se consideraron "sin clasificar". Para *MetaBAT2* v1.7 se consideró una longitud mínima de *contig* de 2500 pb, al igual que para *CONCOCT* v1.1.

Posteriormente, los *bins* fueron optimizados mediante desreplicación, agregación y puntuación con *DASTool* v1.1.2 (Sieber et al., 2018). Los objetos de entrada fueron: el archivo de ensamblaje seleccionado y los tres archivos resultantes del agrupamiento de *contigs* de cada una de las tres herramientas utilizadas. Se trabajó bajo los parámetros preestablecidos, utilizando *diamond* para la identificación de genes de una sola copia.

3.2.4. EVALUACIÓN DE LA CALIDAD DE *BINS*

Una vez optimizados los *bins* se procedió a evaluar su calidad por medio de estimaciones de integridad y contaminación utilizando la herramienta *CheckM* v1.0.18 (Parks, Imelfort, Skennerton, Hugenholtz, & Tyson, 2015). Esta herramienta genera conjuntos de genes marcadores específicos de clado para cada *bin* e informa la resolución taxonómica posible.

Para evaluar la calidad, identifica y cuenta los conjuntos de marcadores o genes universales de copia única (GCU) de cada *bin*, que deben existir dentro de un linaje filogenético. Los GCU consisten principalmente en genes de mantenimiento o que codifican proteínas ribosómicas y hay varias listas de ellos que pueden utilizarse como referencia (Olm,

2017). Así, la integridad se estima como el número de conjuntos de marcadores presentes en un genoma teniendo en cuenta que sólo una porción de un conjunto de marcadores puede ser identificado. Mientras que, la contaminación se estima a partir del número de genes marcadores que están presentes en varias copias en cada conjunto de marcadores, ya que solo debe existir una copia (Parks et al., 2015).

El objeto de entrada fue el obtenido con *DASTool* v1.1.2, mientras que los parámetros empleados fueron los preestablecidos, es decir, árbol reducido, que seleccionó el conjunto de marcadores específicos del linaje para los genomas, y guardar las gráficas generadas.

Dado que se pretende evaluar el potencial funcional de un determinado genoma, se buscó la máxima integridad y mínima contaminación. Para continuar con el flujo de trabajo se extrajeron los ensamblajes de cada *bin* con la herramienta *BinUtil* v1.0.2 (Arkin et al., 2018). El objeto de entrada para esta herramienta fue el generado por *DASTool* v1.1.2.

3.2.5. ANOTACIÓN DEL GENOMA

La anotación de los genomas pertenecientes a cada *bin* se llevó a cabo con la herramienta *Annotate Multiple Microbial Assemblies with RASTtk (AMMA with RASTtk) - v1.073* (Brettin et al., 2015). El objeto de entrada fue el resultado de la herramienta *BinUtil* v1.0.2. Se anotó los objetos de tipo ensamblaje (*conjunto de contigs*) para obtener objetos con características de codificación y no codificación, es decir objetos de tipo genoma.

Los parámetros utilizados en esta herramienta fueron los predeterminados, es decir seleccionar todas las opciones a excepción de "*Call features prophage phispy*" debido a que vuelve al procesamiento lento. De los genomas anotados únicamente se seleccionó el de interés para mostrar los resultados, aquel con máxima integridad y mínima contaminación.

3.2.6. CLASIFICACIÓN TAXONÓMICA

Se procedió a obtener la clasificación taxonómica con la herramienta *Genome Taxonomic Database* (GTDB)-Tk v1.7.0 (Chaumeil, Mussig, Hugenholtz, & Parks, 2020; Jain, Rodriguez-R, Phillippy, Konstantinidis, & Aluru, 2018; Matsen, Kodner, & Armbrust, 2010), para los diez *bins*, y se compararon los resultados con los obtenidos con *AMMA with RASTtk*. El objeto de entrada fue el obtenido en la anotación con *AMMA with RASTtk* donde se fijó el porcentaje mínimo de alineación con el valor predeterminado de 10.

Posteriormente se construyó un árbol de especies con la herramienta *Species tree* v2.2.0 (Price, Dehal, & Arkin, 2010). La relación está determinada por la similitud de alineación con 49 dominios COG (*Clusters of Orthologous Groups*), que son genes universales centrales definidos por familias. Para la construcción del árbol filogenético, la aplicación utiliza *FastTree2* v2.1.10, una herramienta para estimar de manera rápida la filogenia de máxima verosimilitud aproximada. Al combinar el subconjunto de genomas públicos de KBase estrechamente relacionados con los genomas proporcionados, se crea el árbol filogenético (Price et al., 2010). El objeto de entrada fue el set de genomas proporcionado por la herramienta *AMMA with RASTtk* y el parámetro número de genomas públicos más cercanos que contendrá el árbol se fijó en 36.

De manera adicional, se realizó un segundo árbol filogenético con el método *Codon Tree* (Cock et al., 2009; Davis et al., 2016; Edgar, 2004; Stamatakis, 2014; Stamatakis, Hoover, & Rougemont, 2008) en la página del Centro de recursos bioinformáticos bacterianos y virales (BV-BRC) (BRC, n.d.; Wattam et al., 2014). Para ello se ingresó a la herramienta *Bacterial Genome Tree*, se escogieron los genomas bacterianos relacionados con el genoma del *bin* seleccionado de acuerdo con los resultados previos obtenidos y según su disponibilidad en la base de datos de la herramienta. Se seleccionó la carpeta de salida de los resultados y se escribió el nombre del archivo. El número de genes seleccionado fue de 100, las deleciones y duplicaciones máximas permitidas se fijó en 5.

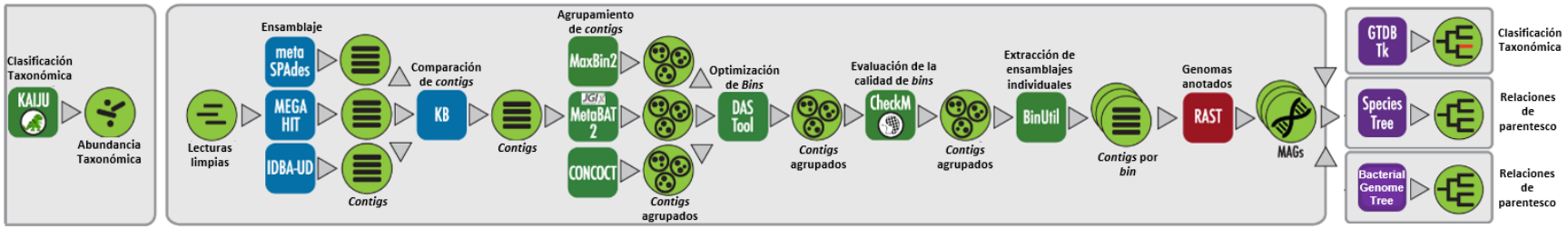


Figura 3.2. Flujo de trabajo para el ensamblado del metagenoma

Modificado de Chivian et al. (2020)

3.3. BÚSQUEDA DE GENES Y RUTAS METABÓLICAS DE INTERÉS

Para finalizar, se descargó el genoma del *bin* seleccionado y se procedió con el flujo de trabajo de la Figura 3.3, para la búsqueda de genes y rutas metabólicas de interés en el genoma anotado de una proteobacteria perteneciente a la familia Orbaceae. Los parámetros se fijaron de acuerdo con las recomendaciones del tutorial "*Microbial Genomics in KBase: Gene Feature Analysis"* (Allen, 2021)

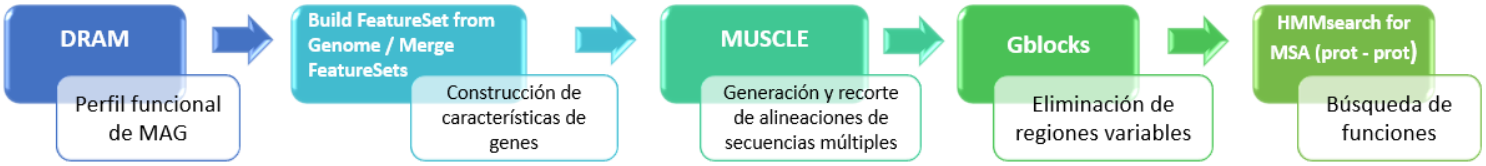


Figura 3.3. Flujo de trabajo para la búsqueda de genes y rutas de interés

Modificado de Allen (2021)

3.3.1. PERFIL FUNCIONAL O METABÓLICO DEL ORGANISMO

El flujo de trabajo en KBase continuó con la herramienta *DRAM v0.1.2* (Shaffer et al., 2020) para obtener un resumen del perfil funcional o metabólico del organismo. *DRAM v0.1.2* predice secuencias de codificación y anota los genes de un objeto de ensamblaje de KBase con bases de datos por lo que el objeto de entrada fue el generado por *DASTool* v1.1.2. Como parámetros se fijó una longitud mínima de *contig* a anotar de 2000, tabla de traducción 11 pues representa el código genético para traducir los genes predichos a aminoácidos, un umbral de puntuación de 60 *bits* para asignar un resultado en búsquedas tipo *blast* y un umbral de puntuación de *bits* de búsqueda inversa de 350.

3.3.2. GENES RELACIONADOS CON LA HIDRÓLISIS DE PET

El primer paso para la identificación de genes relacionados con la hidrólisis de PET fue la creación de un conjunto de características con la herramienta *Build FeatureSet from Genome* v1.2.6 (Arkin et al., 2018) de cada uno de los genes descritos por Danso et al. (2018) y previamente enlistados en la sección 1.1. Posteriormente, se unieron en un solo conjunto de funciones con la herramienta *Merge FeatureSets* v1.7.4 (Arkin et al., 2018) para su posterior búsqueda en el genoma ensamblado, cuyo único parámetro requerido es una descripción.

A partir del conjunto de genes de PETasas generado como objeto, se obtuvo una alineación de secuencias múltiple (MSA) de secuencias de proteínas con *MUSCLE* v3.8.425 (Edgar, 2004). El objeto de entrada fue el generado con *Merge FeatureSets* v1.7.4. En caso de que no exista convergencia, como parámetros se fijaron los valores máximos de 16 iteraciones y 0,5 horas.

Para obtener una MSA recortada que mantenga únicamente los bloques conservados se utilizó *Gblocks* v0.91b (Castresana, 2000). Como entrada se utilizó el objeto generado con *MUSCLE* v3.8.425. Se estableció un nivel de recorte de 1 y los valores predeterminados para: número mínimo de secuencias requeridas con un residuo conservado (0), número mínimo de secuencias requeridas con un residuo en una posición que flanquea una posición conservada (0), número máximo de posiciones no conservadas (columnas) que se pueden

incluir en un bloque (8) y el número mínimo de posiciones (columnas) necesarias para incluir un bloque en el MSA de salida (10).

Para finalizar, se buscaron las coincidencias de los genes con la herramienta *HMMER Search from MSA (prot-prot)* v3.3.2. La herramienta convierte la MSA de proteínas en un modelo oculto de Markov (HMM) que se utiliza para buscar en una base de datos de secuencias de proteínas. Como entradas se utilizó el objeto generado con *Gblocks v0.91b* y el genoma ensamblado del *bin* seleccionado y anotado con *DRAM v0.1.2*. El único parámetro que se fijó fue el valor e en 0,001 y, por tanto, los valores por debajo de este no se informaron.

4. RESULTADOS Y ANÁLISIS DE RESULTADOS

4.1. ELIMINACIÓN DE LAS LECTURAS DEL HOSPEDERO

4.1.1. CONTROL DE CALIDAD

En la Tabla 4.1 se presentan las estadísticas básicas para las secuencias *forward* y *reverse*. La longitud de secuencia proporciona el valor de la secuencia más corta y larga, no obstante, en la tabla se informa un único valor por lo que se puede concluir que las secuencias obtenidas tienen la misma longitud de 150 pb.

Tabla 4.1. Estadísticas básicas para las secuencias *forward* y *reverse*

Medida	Valor
Nombre del archivo	forward1.txt
Tipo de archivo	Llamada de bases convencionales
Codificación	Sanger / Illumina 1.9
Secuencias totales	21 103 105
Secuencias marcadas como de baja calidad	0
Longitud de la secuencia	150
% GC	41

Medida	Valor
Nombre del archivo	reverse1.txt
Tipo de archivo	Llamadas de bases convencionales
Codificación	Sanger / Illumina 1.9
Secuencias totales	21 103 105
Secuencias marcadas como de baja calidad	0
Longitud de la secuencia	150
% GC	41

En la Figura 4.1 se muestra la calidad de la secuencia por base para las secuencias *forward* y *reverse*. Se puede apreciar que los puntajes de calidad caen desde el principio hacia el final de las lecturas, lo cual era esperado dado que las lecturas fueron generadas por secuenciación por síntesis de Illumina. En este caso, no se presentan errores inesperados por avería de la instrumentación puesto que no existen caídas repentinas de la calidad o un gran porcentaje de lecturas de baja calidad que indiquen un problema de la instalación de secuenciación, por lo que no sería necesario contactar con el centro de secuenciación. Entonces, la caída en la calidad hacia el final de las lecturas podría explicarse por el decaimiento de la señal o fase (Babraham Institute, 2020).

En las gráficas se puede notar que las barras de error alcanzan puntajes de calidad altos o de buena calidad (superiores a 30) hasta el nucleótido 124 para la secuencia *forward* y hasta el nucleótido 74 para la secuencia *reverse*. Para las siguientes posiciones en las lecturas se obtuvieron barras de error que alcanzan puntajes de calidad medios o de calidad razonable (superiores a 24). No obstante, el puntaje de calidad medio (línea azul) indica que todos los

nucleótidos presentaron una calidad superior a 32, tanto para la secuencia *forward* como para la secuencia *reverse*.

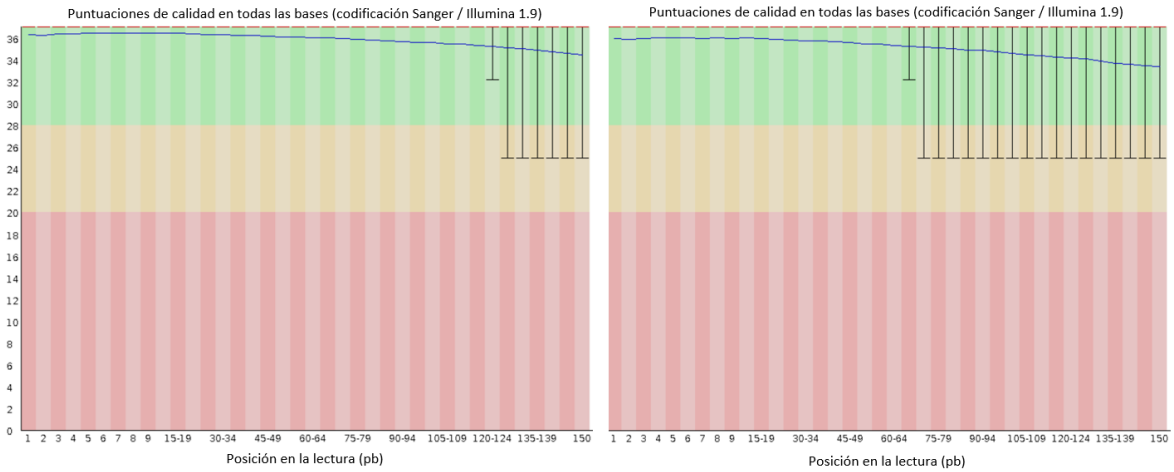


Figura 4.1. Calidad de la secuencia por base para las secuencias *forward* (izquierda) y *reverse* (derecha)

De acuerdo con el Babraham Institute, se espera que las gráficas de contenido de GC por secuencia tengan una forma más o menos normal, mientras que una distribución inusual podría indicar contaminación o sesgos. Si se presentan picos agudos la muestra está contaminada por compuestos específicos (por ejemplo, dímeros), mientras que los picos anchos pueden representar contaminación con otra especie (Babraham Institute, 2020). En la Figura 4.2 se presentan las gráficas obtenidas del contenido de GC. Las distribuciones se muestran uniformes, donde los picos centrales prácticamente coinciden con la distribución teórica y no se presentan picos adicionales, lo que indica ausencia de contaminación por lo que no se generaron advertencias de ningún tipo en el informe (Anexo II).

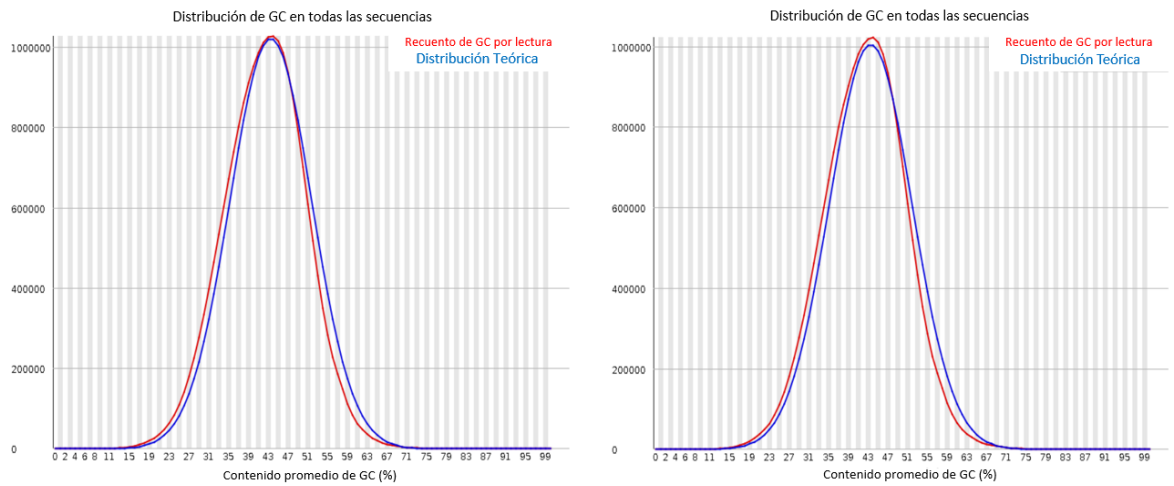


Figura 4.2. Contenido de GC por secuencia, forward (izquierda) y reverse (derecha)

La gráfica de calidad de secuencia por mosaico presentó una advertencia (Anexo II). Esta gráfica permite asociar el puntaje de calidad de cada mosaico en todas las bases con la celda de flujo para conocer si la pérdida de calidad está asociada con alguna parte específica de la celda. Las advertencias pueden generarse por problemas como presencia de burbujas, desechos o manchas en la celda (Babraham Institute, 2020) pero esto puede solucionarse con la eliminación de fragmentos de baja calidad.

4.1.2. LIMPIEZA DE SECUENCIAS Y RE-EVALUACIÓN DE LA CALIDAD

Los resultados de FastQC v0.11.9 luego de eliminar las lecturas de baja calidad (no se presentaron adaptadores) con herramienta *Trimmomatic* v0.39, se muestran en la Tabla 4.2. Se puede notar que el 94,93 % de las lecturas (20 033 887) pasaron el filtro de calidad, mientras que el restante 5,07 % de las lecturas fueron eliminadas por su baja calidad. La longitud de secuencia más corta fue de 50 pb, tanto para la secuencia forward como para la secuencia reverse.

Tabla 4.2. Estadísticas básicas para las secuencias forward (izquierda) y reverse (derecha) luego de la limpieza

Medida	Valor
Nombre del archivo	PS_leftP.fastq.gz
Tipo de archivo	Llamada de bases convencionales
Codificación	Sanger / Illumina 1.9
Secuencias totales	20 033 887
Secuencias marcadas como de baja calidad	0
Longitud de la secuencia	50 - 150
% GC	41

Medida	Valor
Nombre del archivo	PS_rightP.fastq.gz
Tipo de archivo	Llamada de bases convencionales
Codificación	Sanger / Illumina 1.9
Secuencias totales	20 033 887
Secuencias marcadas como de baja calidad	0
Longitud de la secuencia	50 - 150
% GC	41

En la Figura 4.3 se puede apreciar que, aunque para la secuencia reverse las barras de error a partir del nucleótido 104 alcanzan una calidad media o razonable (superior a 24), para los 150 nucleótidos de ambas secuencias los puntajes de calidad fueron altos o de buena calidad (superior a 30), lo cual es adecuado para el análisis bioinformático.

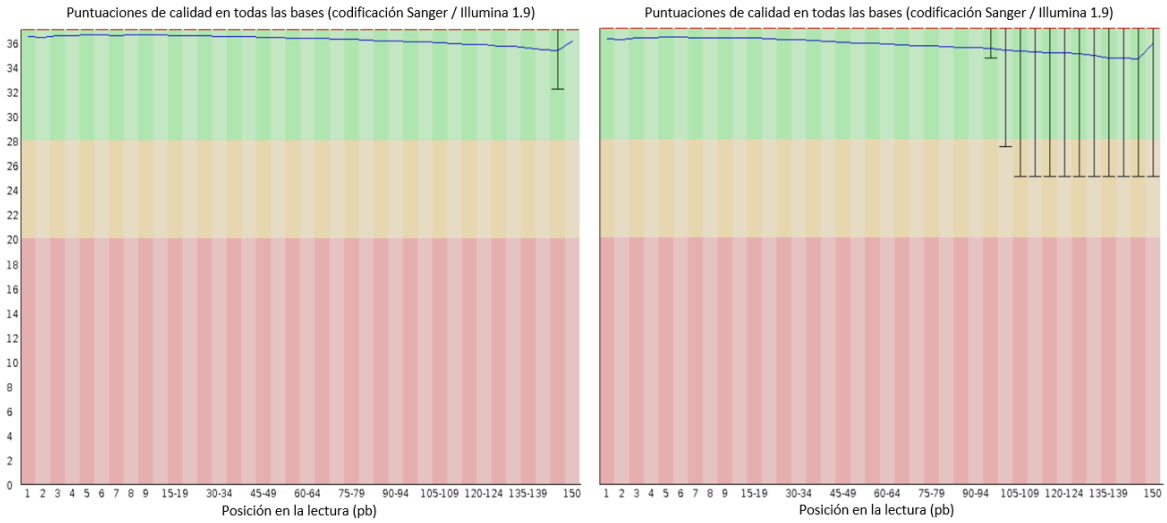


Figura 4.3. Calidad de la secuencia por base para las secuencias forward (izquierda) y reverse (derecha) luego de la limpieza de las secuencias

Si bien hubo una mejora en los puntajes de calidad luego de la limpieza, no hubo un cambio en la mayor frecuencia observada del promedio de las puntuaciones de calidad por lectura, pues se mantuvo en 36 para ambas secuencias.

Se debe mencionar que antes y después de la limpieza se tienen alertas en las gráficas de contenido de secuencia por base. Las advertencias en este tipo de gráfica se generan cuando se presenta un valor superior al 10 % en la diferencia entre A y T, o G y C, lo cual es normal en torno a los 12 pb iniciales pues es un sesgo técnico que aparece de cualquier librería generada por tagmentación y no puede corregirse mediante recorte pero como no representa ninguna secuencia sesgada individualmente no afecta en el análisis posterior (Babraham Institute, 2020). Dado que luego de los 12 pb la gráfica mostró líneas paralelas de contenido similar entre A y T, y entre G y C, se hizo caso omiso a estas advertencias.

Luego del recorte también se generó una advertencia en la gráfica de distribución de longitud de secuencia. Esto se debe a que las secuencias ya no poseen la misma longitud, pero es algo normal por lo que las advertencias de este apartado también pueden ignorarse.

4.1.3. ELIMINACIÓN DE LECTURAS DUPLICADAS

En la Figura 4.4 se puede observar que, de las 40 067 774 lecturas ingresadas, el 11,76 % estuvieron duplicadas, lo cual representó el 12 % de los nucleótidos iniciales.

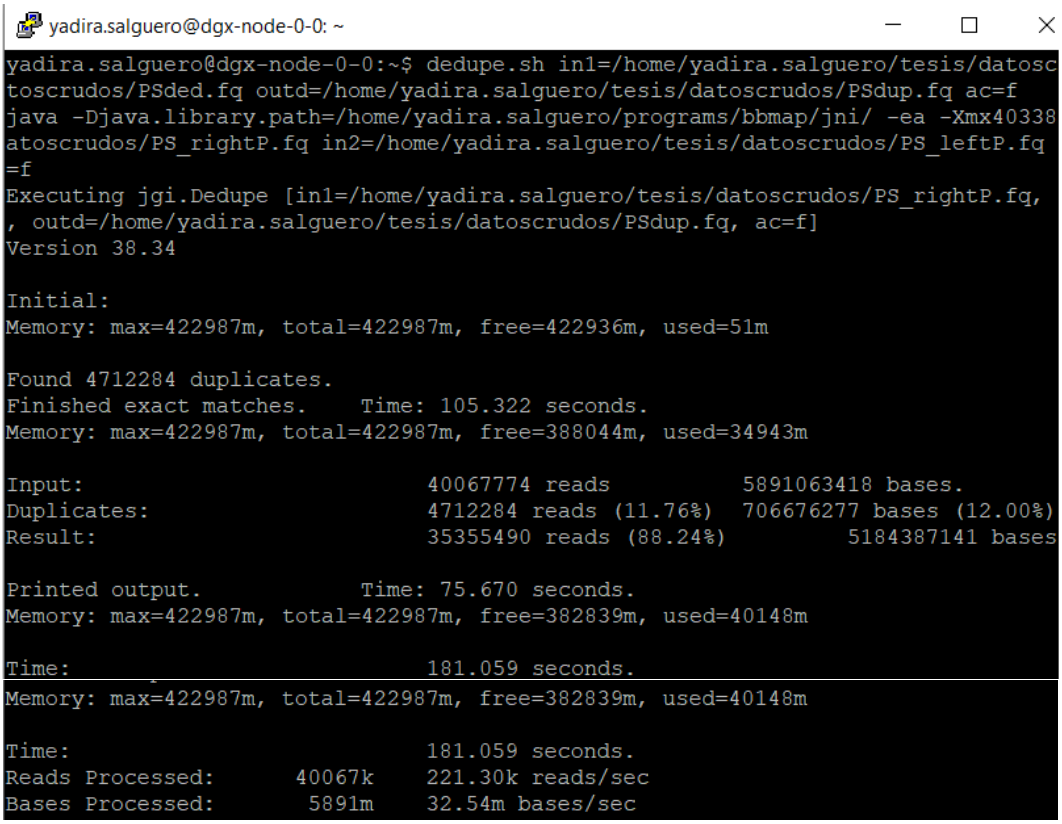


Figura 4.4. Resultados obtenidos con *Dedupe*

En la Figura 4.5 se muestra el reformateó de las lecturas en formato FASTQ, obteniéndose archivos que consideran las 35 355 490 lecturas sin duplicar como el 100 %.

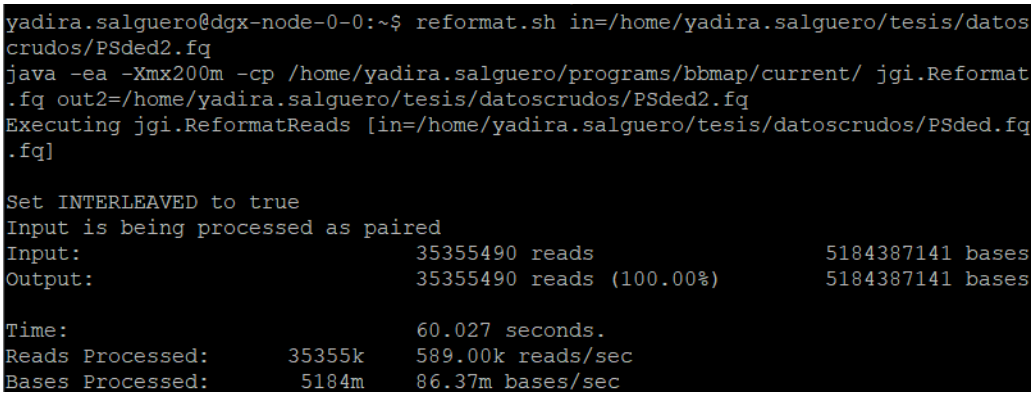


Figura 4.5. Reformateo de las lecturas en formato FASTQ con *Reformat*

4.1.4. ELIMINACIÓN DE CONTAMINACIÓN

Como se observa en las Figuras 4.6 y 4.7, el genoma de referencia de *H. illucens*, encontrado en NCBI, tiene el número de acceso GCF_905115235.1 (NCBI, 2020) y formato FASTA.

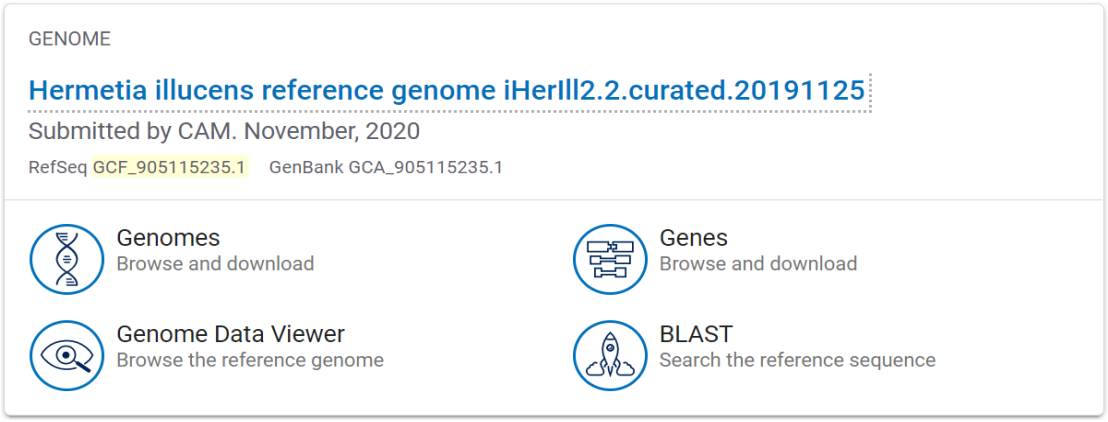


Figura 4.6. Resultado de la búsqueda del genoma de referencia de *H. illucens* en NCBI

```
yadira.salguero@dgx-node-0-0:~/tesis/datoscrudos$ head GCF_Hermetia.fna
>NC_051849.1 Hermetia illucens chromosome 1, iHerIll2.2.curated.20191125, whole
genome shotgun sequence
atttcgtcatcatcaaatcacaggctatacacttgctctcagacgctgataccctcagatctcaagatcacaaatcaggt
tccttttcattattttgcaagtacttaacgagaaattaacatttaaatcctggaaaagtgatttttcagcttttcctcca
ttttcgtcatcatcaaatcacaggctatacacttgctctcagacgctgataccctcagatctcaagatcacaaatcaagt
tccttttcaatattttgcaagtacttaacgagaaattaacatttaaatcctggaaaagtgatttccagcttttcctcca
ttttcgtcatcatcaaatcacaggctatacacttgctctgagacgctgataccctcagatctcaagatcacaaatcaggt
tccttttcattattttgcaagtacttaacgagaaattaacatttaaatcctggaaaagtgatttttcagcttttcctcca
ttttcgtcatcatcaaatcacaggctatacacttgctctcagacgctgataccctcagatctcaagatcacaaatcaggt
tccttttcattattttgcaagtacttaacgagaaattaacatttaaatcctggaaaagtgatttttcagattttcctcta
ttttcgtcatcatcaaatcacaggctatacacttgctctcagacgctgatatcctcagatctcaagatcacaaatcaggt
```

Figura 4.7. Genoma de referencia de *H. illucens* en formato FASTA

En la Figura 4.8 se muestra la base de datos creada a partir del genoma de referencia con la herramienta *Bowtie2* v2.4.2.

```
yadira.salguero@dgx-node-0-0:~/tesis/datoscrudos$ ls
anteriores              mapped_and_unmapped.sam  PS_leftP.fq
GCF_Hermetia.fna        PS                       PS_leftU.fastq.gz
Hermetia_DB.1.bt2       PSded1.fq                PS_rightP_fastqc2.html
Hermetia_DB.2.bt2       PSded2.fq                PS_rightP_fastqc.zip
Hermetia_DB.3.bt2       PSded.fq                 PS_rightP.fq
Hermetia_DB.4.bt2       PSdup.fq                 PS_rightU.fastq.gz
Hermetia_DB.rev.1.bt2   PS_leftP_fastqc2.html
Hermetia_DB.rev.2.bt2   PS_leftP_fastqc.zip
```

Figura 4.8. Base de datos obtenida con el genoma de *H. illucens*

Los resultados indicaron que se alinearon y mapearon sobre el genoma de referencia del hospedero 8 759 603 lecturas PE concordantemente y 296 709 lecturas PE discordantemente, lo que representó un 51,23 % de las 17 677 745 lecturas PE de entrada (Figura 4.9). Además, del 48,77 % restante, se alinearon y mapearon 2 989 770 lecturas SE, es decir un 8,46 %

adicional, lo que indicó un 59,69 % de alineamiento general. Al final, se generó un archivo .sam que mantuvo tanto las lecturas mapeadas como las lecturas no mapeadas.

```
yadira.salguero@dgx-node-0-0:~$ export PATH=$PATH:/home/yadira.salguero/programs
/bowtie2-2.4.2-sra-linux-x86_64:$PATH
yadira.salguero@dgx-node-0-0:~$ bowtie2 -x /home/yadira.salguero/tesis/datoscrud
os/Hermetia_DB -1 /home/yadira.salguero/tesis/datoscrudos/PSded1.fq -2 /home/yad
ira.salguero/tesis/datoscrudos/PSded2.fq -S /home/yadira.salguero/tesis/datoscru
dos/mapped_and_unmapped.sam
17677745 reads; of these:
  17677745 (100.00%) were paired; of these:
    8918142 (50.45%) aligned concordantly 0 times
    6187132 (35.00%) aligned concordantly exactly 1 time
    2572471 (14.55%) aligned concordantly >1 times
    ----
    8918142 pairs aligned concordantly 0 times; of these:
      296709 (3.33%) aligned discordantly 1 time
    ----
    8621433 pairs aligned 0 times concordantly or discordantly; of these:
      17242866 mates make up the pairs; of these:
        14253096 (82.66%) aligned 0 times
        1716026 (9.95%) aligned exactly 1 time
        1273744 (7.39%) aligned >1 times
59.69% overall alignment rate
```

Figura 4.9. Lecturas mapeadas y no mapeadas frente al genoma de *H. illucens*

Luego, como resultado del cambio de formato del archivo .sam y el filtrado de las secuencias no alineadas con la herramienta *samtools* v1.10-3, se obtuvo un archivo .bam que contenía únicamente las secuencias con lecturas pareadas que no fueron mapeadas al genoma de referencia del hospedero, es decir las 14 253 096 lecturas que no pertenecen a *H. illucens*.

Así mismo, la herramienta *samtools* v1.10-3 permitió organizar las lecturas, generando un archivo .bam con las lecturas ordenadas.

Como resultado de procesar este archivo con la herramienta *bedtools* v.2.25.0 se obtuvieron los archivos FASTQ de las secuencias *forward* (Figura 4.10) y *reverse* (Figura 4.11) para su procesamiento en KBase.

```
yadira.salguero@dgx-node-0-0:~/tesis/datoscrudos$ head r1.fastq
@A00253:464:HGFW7DSX2:1:1101:1045:30984/1
GCATTTTTATTATGAAAGAAATATAAAGGCAATGGTTTAAAGTACCTTTCAATTTCTATGTTAATAATAACGGCAAAATGAAGGATAACTTTAAGTTTTTTTATAGATT
ATAAGCTTAGCTG
+
:FFFF,FFFF,FF:FFFFFFFFFFFFFFFFFF:FFFFFF:FFF,FF,F:FFF:FFFFFFFFFFFFFF,F,,FFFF,F,FFFFFFFF:FFFFF,FF,F:FFFFF,F,F
FFFFFF::FF,,F
@A00253:464:HGFW7DSX2:1:1101:1072:12305/1
GGCCGAAATTCCAGTTGCAGCCCTGTCCGCTACGGCTTTGACTTGCTCCAAGAAG
+
F:FFFFFFF:F,FFFF:,FF,F:FFF::F,:F,FF:FFF:FFFFFFF:F:FF,F:
@A00253:464:HGFW7DSX2:1:1101:1090:20102/1
GGACGCTTTTGCCGTTCATGTGGCTCATCTGAAGTGGCTCTTACCACGAAGTCTTATCGGTTATCTGATACCGTGGGAACCGGCATGATCCTCTAACATATGCTCTTCG
AGGATTCCTGAAGTAGGTGCTCTAAGGTCGGTTACTTTCGG
yadira.salguero@dgx-node-0-0:~/tesis/datoscrudos$
```

Figura 4.10. Secuencia *forward* en formato FASTQ

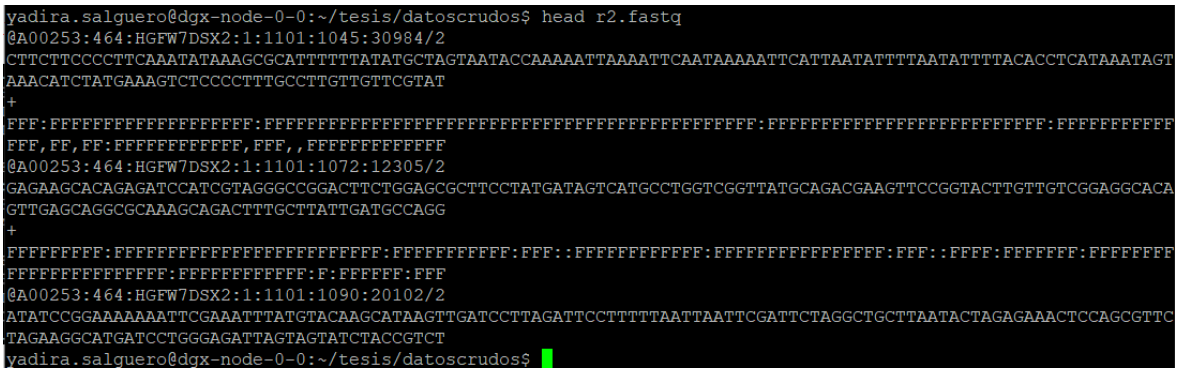


Figura 4.11. Secuencia reverse en formato FASTQ

4.2. ENSAMBLAJE Y ANOTACIÓN DEL METAGENOMA

4.2.1. CLASIFICACIÓN TAXONÓMICA (*KAIJU*)

Con la base de datos de NCBI RefSeq, se predijo la composición microbiana de la muestra en función de las similitudes de proteínas para una comparación posterior con los árboles de especies generados a partir del ensamblaje y la anotación del metagenoma.

En las siguientes figuras, generadas con la herramienta *Kaiju* v1.7.3 se presenta la clasificación taxonómica a nivel de filo, clase (Figura 4.12), orden (Figura 4.13), familia (Figura 4.14), género (Figura 4.15) y especies (Figura 4.16) encontradas en el tracto intestinal de la MSN.

Los resultados indican que, en los intestinos de larvas de MSN los microorganismos más abundantes a nivel de filo son Firmicutes, Proteobacteria, Bacteroidetes y Actinobacteria lo cual concuerda con lo reportado por otros autores (Zhineng et al., 2021); mientras que, los 15 géneros más importantes son: *Providencia, Enterococcus, Frischella, Dysgonomonas, Hungatella, Gilliamella, Pelosinus, Lacrimispora, Clostridium, Orbus, Vagococcus, Lachnoclostridium, Klebsiella, Raoultella y Lactobacillus*. A nivel de género y especie, la abundancia relativa bacteriana difiere de la indicada por otros investigadores, lo cual puede deberse a que el tipo de alimentación influye en la abundancia relativa de microorganismos presentes (Osimani et al., 2021). Sin embargo, las especies con abundancia relativa alta que coincidieron con el presente trabajo son: *Hungatella hathewayi, Frischella perrara, Gilliamella apicola, Enterococcus* sp. *y Providencia rettgeri* (Zhineng et al., 2021).

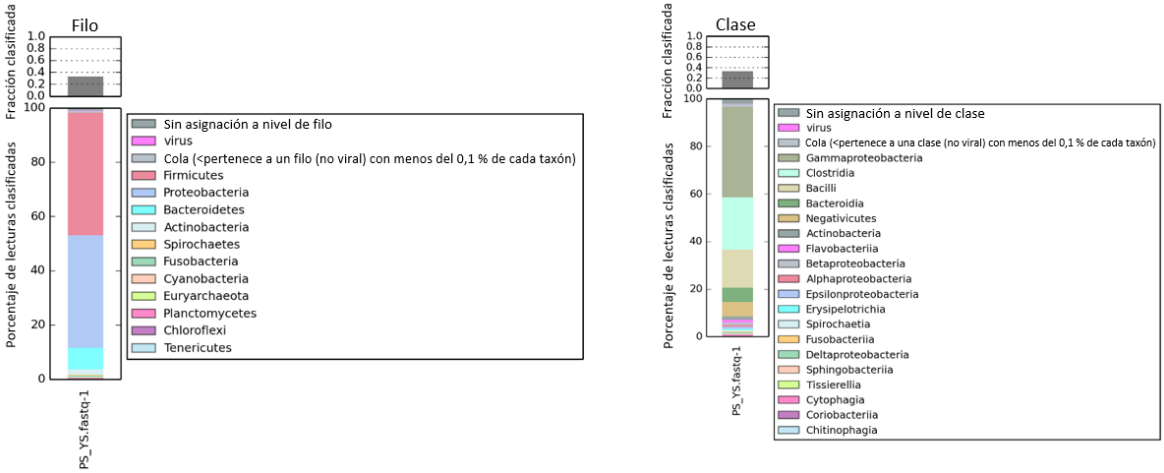


Figura 4.12. Clasificación taxonómica a nivel de filo y clase

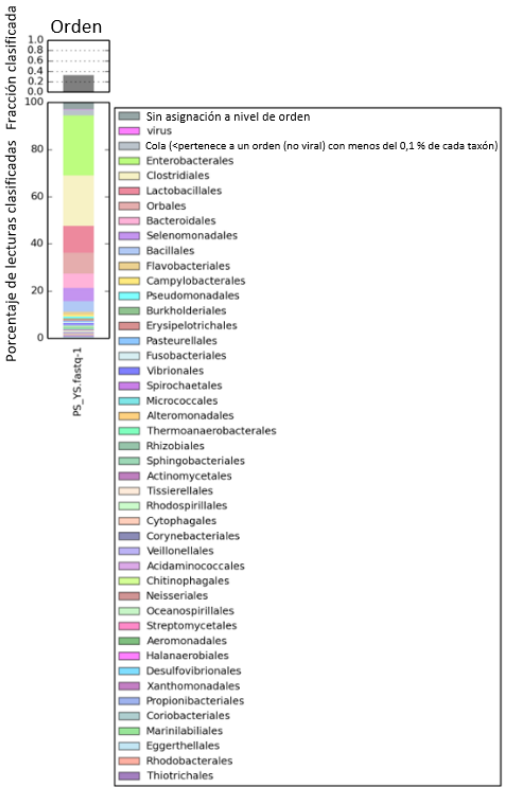


Figura 4.13. Clasificación taxonómica a nivel de orden

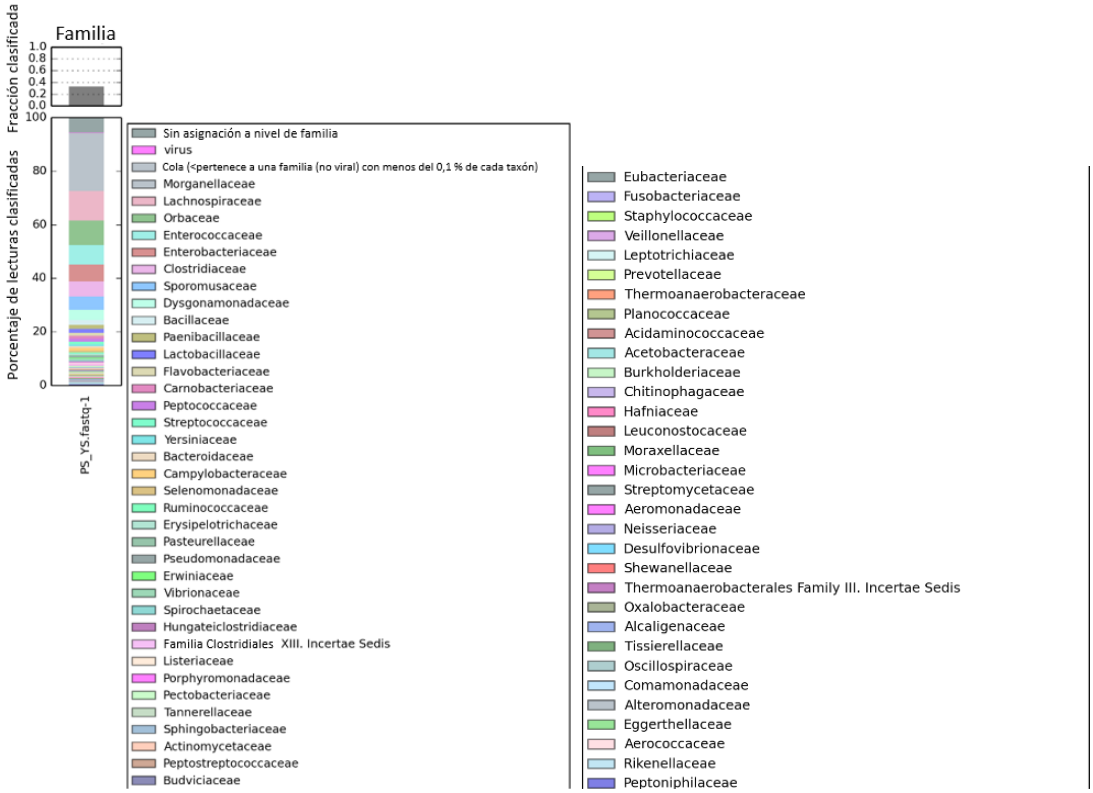


Figura 4.14. Clasificación taxonómica a nivel de familia

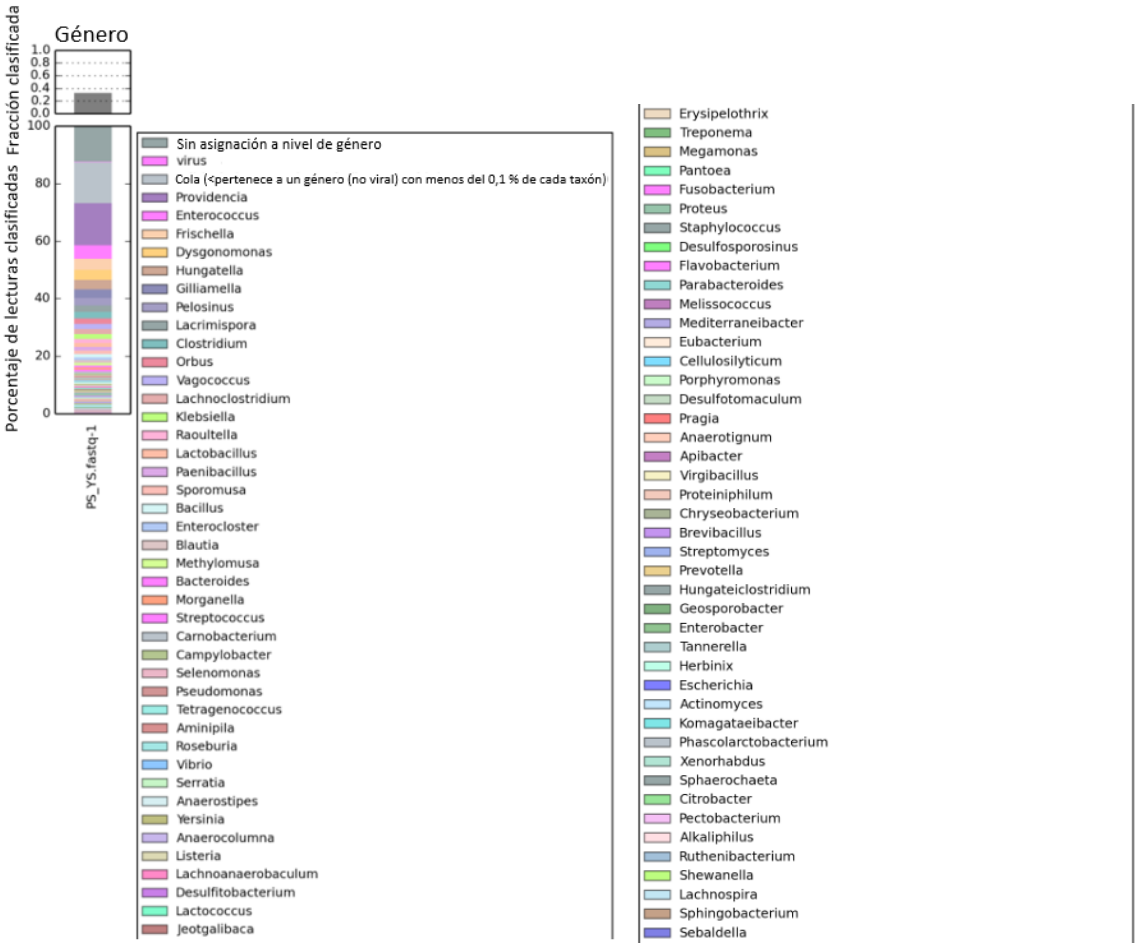


Figura 4.15. Clasificación taxonómica a nivel de género

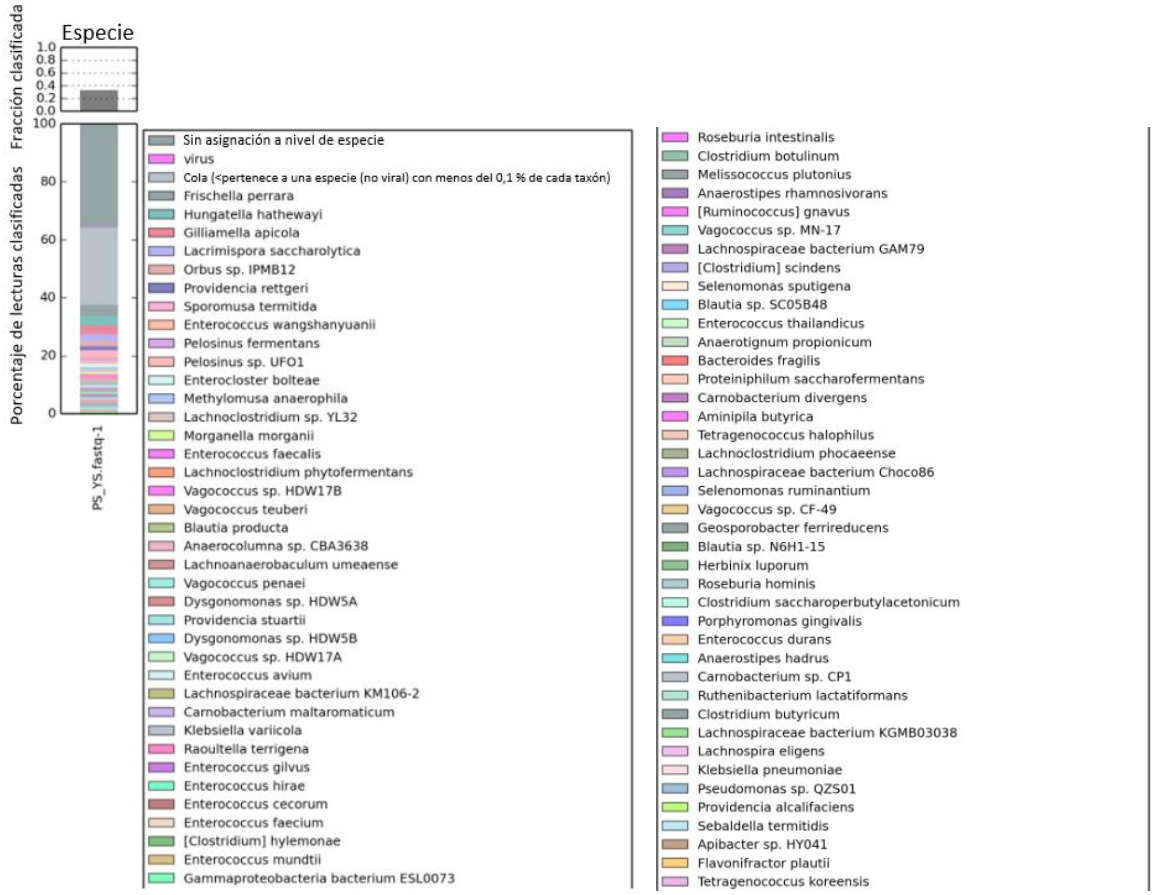


Figura 4.16. Clasificación taxonómica a nivel de especie

En la Figura 4.17 se presentan las especies y abundancia relativa del metagenoma en un diagrama de KRONA. De acuerdo con este, se identificó un 0,2 % de virus y 0,5 % de otros organismos diferentes a bacterias y virus.

Se puede notar que, uno de los filos con mayor representación en el metagenoma bacteriano es el de las proteobacterias (42 %) con las gammaproteobacterias como las más abundantes (38 % de las bacterias y 92 % de las proteobacterias). Dentro de esta clase destacan los órdenes enterobacterales (26 % de las bacterias, 62 % de las proteobacterias y 67 % de las gammaproteobacterias) y orbales (9 % de las bacterias, 22 % de las proteobacterias y 24 % de las gammaproteobacterias). En cuanto a familia, el 60 % de las lecturas clasificadas se asignaron a la Orbaceae (Figura 4.14), dentro de la cual se puede encontrar a *Orbus* sp. IPMB12 (2 % de las bacterias, 5 % de las proteobacterias, 5 % de las gammaproteobacterias y 22 % de la familia Orbaceae), *Gilliamella apicola* (3 % de las bacterias, 7 % de las proteobacterias, 8 % de las Gammaproteobacterias y 32 % de la familia Orbaceae) *y Frischella perrara* (4 % de las bacterias y 43 % de la familia Orbaceae).

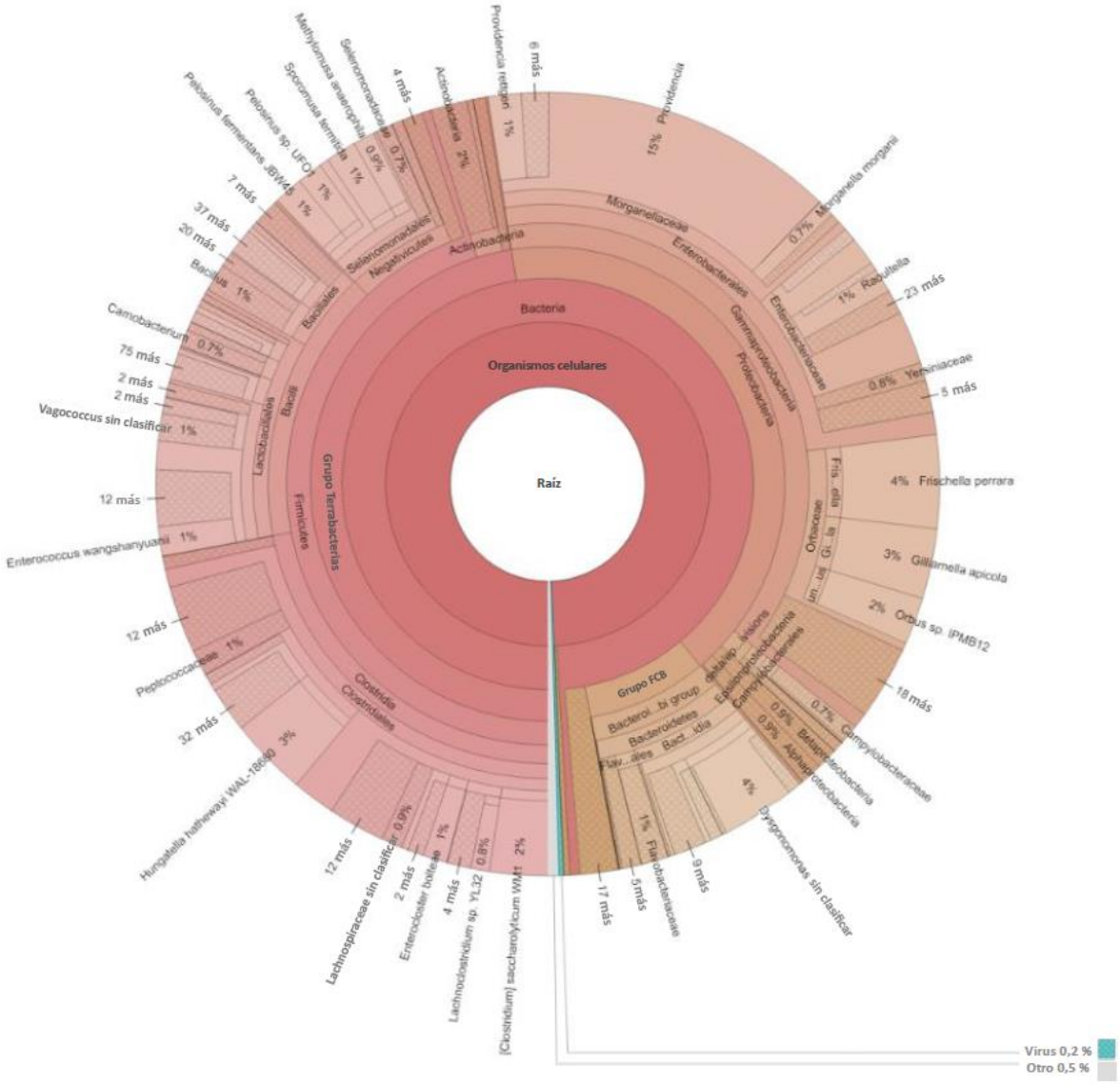


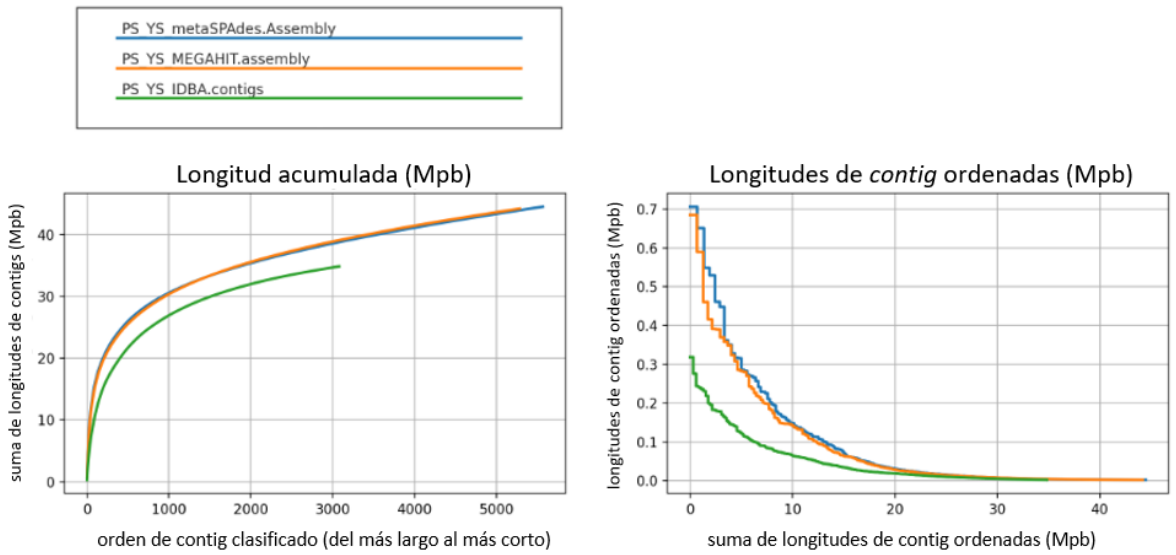
Figura 4.17. Diagrama circular de KRONA

4.2.2. ENSAMBLAJE DEL METAGENOMA Y COMPARACIÓN DE *CONTIGS*

En la Figura 4.18 se presenta el resultado de la herramienta *Compare Assembled Contig Distributions* v1.1.2, el cual es un resumen comparativo de los ensamblajes llevados a cabo con: *metaSPAdes* v3.15.3, *IDBA-UD* v1.1.3 y *MEGAHIT* v1.2.9.

Estos resultados permitieron comparar los ensamblajes en términos de longitud y distribución de tamaño de los *contigs*. Dado que es más deseable tener *contigs* más largos, se seleccionó el ensamblaje realizado con *metaSPAdes* para continuar con el flujo de trabajo del proyecto.

Las herramientas de evaluación de calidad para ensamblajes *QUAST* y *Compare Assembled Contig Distributions*, indicaron que con metaSPAdes se obtuvieron 5 565 *contigs* mayores o iguales a 1 000 pb, 661 *contigs* mayores o iguales a 10 000 pb y 60 mayores o iguales a 100 000 pb. El contig más largo obtenido fue de 706 116 pb y un N50 de 22 421 pb.



Mejor — Peor

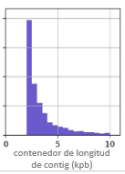
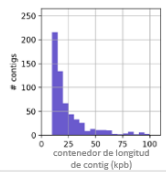
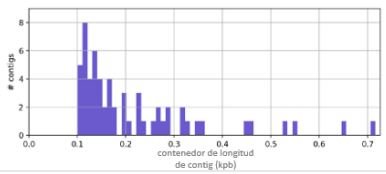
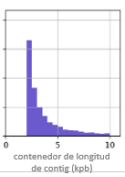
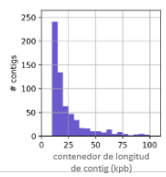
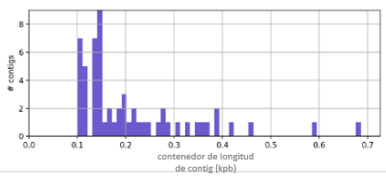
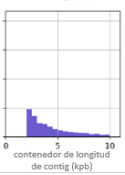
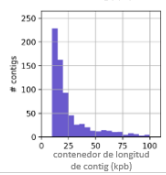
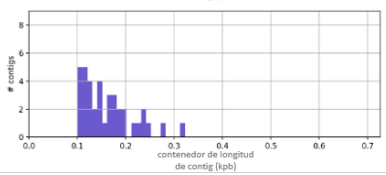
ENSAMBLAJE	Contig más largo (pb)	Nx (Lx)		Longitud (pb)	Núm. Contigs	Sum. Longitud (pb)	Histograma longitud de contig (1 pb <= lon < 10 kpb)	Histograma longitud de contig (10 pb <= lon < 100 kpb)	Histograma longitud de contig (lon >= 100 kpb)
PS_YS_metaSPAdes.Assembly	706116	N50:	22421	>= 10^6	0	0			
		L50:	(281)	>= 10^5	60	13056925			
		N75:	4637	>= 10^4	661	27820926			
		L75:	(1509)	>= 10^3	5565	44487191			
		N90:	2509	>= 500	5565	44487191			
		L90:	(3564)	>= 1	5565	44487191			
PS_YS_MEGAHIT.assembly	685061	N50:	21337	>= 10^6	0	0			
		L50:	(295)	>= 10^5	58	12311443			
		N75:	5178	>= 10^4	685	27667007			
		L75:	(1458)	>= 10^3	5282	44172503			
		N90:	2637	>= 500	5282	44172503			
		L90:	(3330)	>= 1	5282	44172503			
PS_YS_IDBA.contigs	318425	N50:	23747	>= 10^6	0	0			
		L50:	(278)	>= 10^5	38	6180993			
		N75:	8333	>= 10^4	744	24696065			
		L75:	(901)	>= 10^3	3078	34831859			
		N90:	3885	>= 500	3078	34831859			
		L90:	(1830)	>= 1	3078	34831859			

Figura 4.18. Comparación de distribuciones de *contigs* ensamblados con IDBA, MEGAHIT y metaSPAdes

4.2.3. AGRUPAMIENTO Y OPTIMIZACIÓN DE *BINS*

El agrupamiento de *contigs* utilizando *MaxBin2* v2.2.4 resultó en 11 *bins* con el 91,1 % de los *contigs* agrupados. Con *MetaBAT2* v1.7 se obtuvieron 18 *bins* con el 52,36 % de los *contigs*. Mientras que, *CONCOCT* v1.1. obtuvo 49 *bins* con el 64,51 % de los *contigs*.

Los resultados de la optimización de *bins* con *DASTool* v1.1.2 indicaron que la cantidad total de *contigs* en el ensamblaje fue de 5 565, la cantidad de *contigs* agrupados 2 144 (38,52 %) y finalmente, la cantidad de *bins* optimizados fue de 10 pues como se muestra en la Figura 4.19, esta herramienta solo selecciona los *bins* de más alta calidad (BinScore > 0.7).

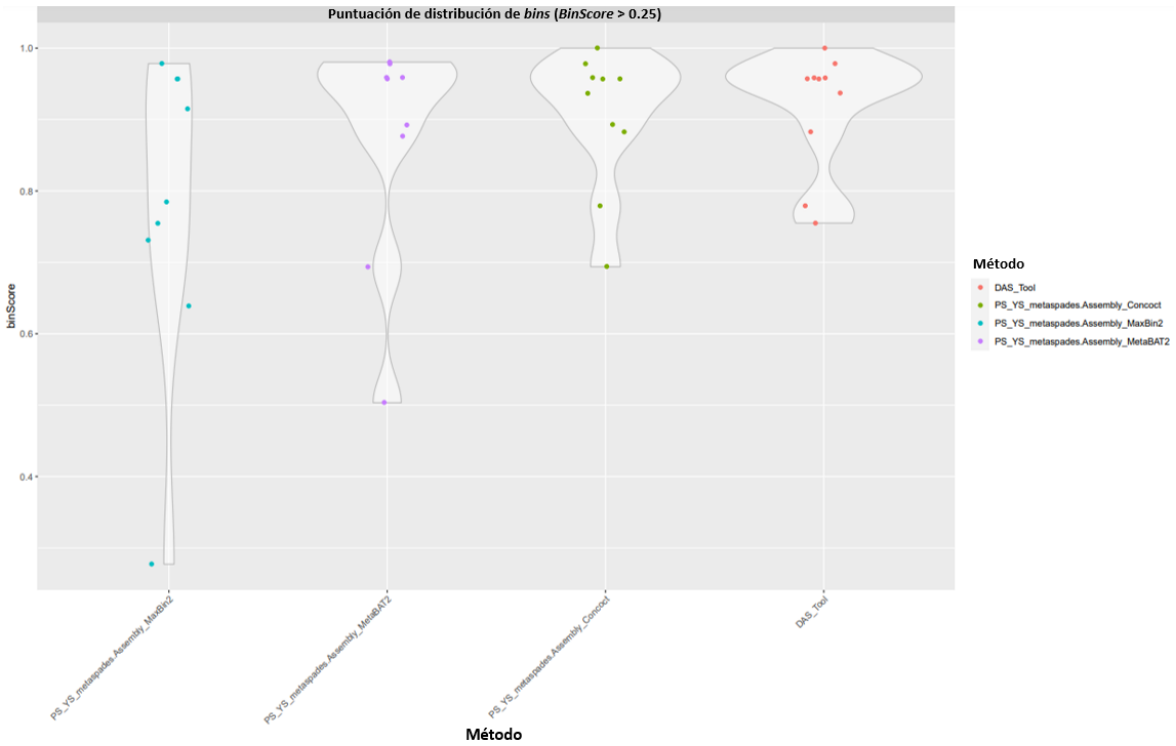


Figura 4.19. Comparación de métodos de agrupamiento en *bins*

4.2.4. EVALUACIÓN DE LA CALIDAD DE BINS

Las estimaciones de integridad y contaminación de los genomas recuperados del metagenoma, proporcionadas por la herramienta *CheckM* v1.0.18, se muestran en la Figura 4.20 y en la Tabla 4.3.

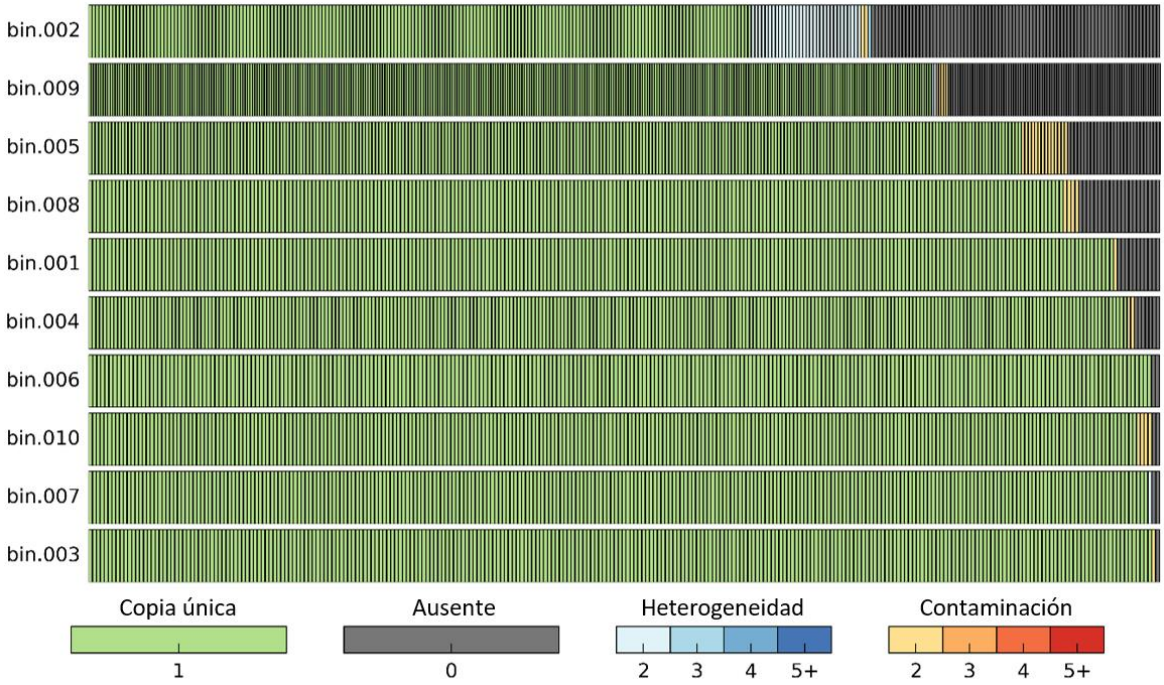


Figura 4.20. Integridad y contaminación de los genomas recuperados del metagenoma

En la Tabla 4.3 se proporciona el número de genomas utilizados para generar el conjunto de marcadores y el número de marcadores generados. Los genes marcadores son de copia única por lo que si aparece más de uno en un genoma o *bin* es indicativo de contaminación. Esto se indica en las columnas que van desde 2 a +5. Los marcadores filogenéticos específicos de clado ausentes se muestran en la columna 0. La integridad o completitud se obtiene por la proporción de marcadores que están presentes con respecto al número total de marcadores utilizados. La presencia de una solo copia del marcador se indica en la columna 1. Lo ideal es tener todos los genes marcadores en exactamente una copia.

Tabla 4.3. Evaluación de linaje de *CheckM*

Bin	Linaje marcador	# Genomas	# Marcadores	# Conjuntos de marcadores	0	1	2	3	4	5+	Completitud	Contaminación
bin.001	o__Clostridiales	155	278	158	11	266	1	0	0	0	95.57	0.63
bin.002	c__Gammaproteobacteria	899	312	185	84	193	34	1	0	0	70.3	15.05
bin.003	k__Bacteria	433	273	183	1	271	1	0	0	0	99.45	0.27
bin.004	p__Firmicutes	100	295	158	7	286	2	0	0	0	96.59	1.27
bin.005	c__Bacilli	586	325	181	28	283	14	0	0	0	88.26	4.28
bin.006	o__Clostridiales	172	263	149	2	261	0	0	0	0	98.66	0.0
bin.007	c__Gammaproteobacteria	965	277	177	2	274	1	0	0	0	99.44	0.05
bin.008	o__Clostridiales	155	278	158	21	253	4	0	0	0	90.17	1.74
bin.009	c__Epsilonproteobacteria	111	445	271	88	350	7	0	0	0	77.54	1.35
bin.010	o__Clostridiales	155	278	158	2	272	4	0	0	0	99.35	2.22

Se puede corroborar que DASTool seleccionó los *bins* de alta calidad pues únicamente seleccionó genomas considerados completos (Tabla 4.3; mayores al 90 %; Sieber et al., 2018), con una contaminación menor al 3 % y genomas considerados borrador (Tabla 4.3; mayores al 70 %; Sieber et al., 2018), con una contaminación entre 1,35 y 15,05 %.

Dado que *CheckM* genera conjuntos de genes marcadores específicos de clado para cada *bin*, en la Tabla 4.3 también se presenta la resolución taxonómica posible en la columna "*Marker Lineage*" para cada uno de los *bins*.

Se puede notar que los *bins* 002 y 007 pertenecen a gammaproteobacterias; no obstante, el *bin* 007 presenta mayor integridad y mínima contaminación por lo que fue seleccionado para futuros análisis. Para ello, con *BinUtil* v1.0.2 se extrajo cada *bin* como un objeto. Esta herramienta indicó que la secuencia genómica perteneciente al *bin* 007 estuvo compuesta por 17 *scaffolds*, con un porcentaje de guanina citocina del 35,65 % y una longitud total de 3 084 216 bp (nucleótidos).

4.2.5. ANOTACIÓN DEL GENOMA

Los tres campos de metadatos esenciales de los ensamblajes son: código genético, dominio y nombre científico. De manera predeterminada, el código genético de genomas de bacterias y arqueas es 11, el dominio es bacteriano y el nombre científico es "taxón desconocido". En la Tabla 4.4 se presenta la taxonomía asignada al *bin* 007 con la herramienta *Annotate Multiple Microbial Assemblies with RASTtk* v1.073 (*AMMA with RASTtk*).

Tabla 4.4. Descripción general del *bin* 007 obtenida con *AMMA with RASTtk*

Nombre del objeto	bin.007.fasta_assembly.RAST	**Taxonomía**
Nombre científico	Taxón desconocido	d__Bacteria p__Proteobacteria c__Gammaproteobacteria o__Enterobacterales f__Enterobacteriaceae g__Orbus s__
Dominio	B	
Código genético	11	
Fuente	KBase	
ID de la fuente	bin.007.fasta_assembly.RAST	
Tamaño	3 084 216	

Se puede apreciar que el género asignado fue *Orbus* sin embargo, el orden y la familia asignados fue Enterobacterales y Enterobacteriaceae, respectivamente y no Orbales y Orbaceae como se esperaba, lo cual puede deberse a la desactualización de la base de datos utilizada para la asignación pues el orden Orbales y la familia Orbaceae fueron propuestos en el 2013 (Kwong & Moran, 2013) y publicados en el manual de Bergey, que contiene la descripción más completa y autorizada de la diversidad bacteriana y de arqueas, en el 2019 (Wilharm, 2019).

El resumen de la herramienta indicó que, de los 2 881 genes encontrados, 78 (2,71 %) no eran codificantes. Así, el genoma de salida del *bin* 007 tuvo las siguientes características:

- Genes codificantes: 2 803 (97,29 %)
- Repetición sin codificación: 32 (1,11 %)
- ARN no codificante: 46 (1,60 %)

Estos resultados están dentro de lo esperado pues para la familia Orbaceae se encontraron los siguientes datos: *G. apicola* posee un único cromosoma de 3 139 412 pb que codifica 2 809 proteínas y 51 ARNt (Kwong, Engel, Koch, & Moran, 2014). En *Z. entericus* (*Orbus* sp. IPMB12) se identificaron 2 304 genes que codifican proteínas y 46 genes de ARNt. *O. hercynius* posee 2 086 genes codificantes de proteínas y 45 tRNA (Wilharm, 2019). Por otro lado, de acuerdo con la base de datos de KBase, el número de genes que codifican proteínas para *O. hercynius* es de 2 107, *Frischella perrara* 2 267, *Orbus* sp. IPMB12 tiene 2 356 y *G. apicola* 2 747.

4.2.6. CLASIFICACIÓN TAXONÓMICA

La clasificación taxonómica obtenida con *Genome Taxonomic Database* (GTDB)-Tk v1.7.0 para los diez *bins* fue la misma obtenida con *AMMA with RASTtk* y se presenta en la Tabla 4.5.

Para el *bin* 007 se estableció como taxonomía de ubicación más cercana a *O. hercynius* (GCF_003634275.1; Tabla 4.6). Lo cual concuerda con lo indicado por Zhineng et al. (2021) al reportar que *O. hercynius* es una especie de abundancia relativa alta en el intestino de

larvas de *H. illucens* (alimentadas con 75 % de salvado de trigo y un 25 % de polvo de soja). No obstante, para el *bin* 007 la identidad de nucleótidos promedio (ANI, por sus siglas en inglés) fue tan solo del 79,42 %. Esto también puede estar relacionado con la falta de aislamiento y caracterización de varias especies, así como la desactualización de la base de datos, mencionada anteriormente.

Tabla 4.5. Clasificación taxonómica obtenida con *AMMA with RASTtk* y GTDB-Tk

Bin	Filo	Clase	Orden	Familia	Género	Especie
001	Fimicutes_A	Clostridia	Lachnospirales	Lachnospiraceae	CHH4-2	
002	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Providencia	rettgeri
003	Bacteroidota	Bacteroidia	Bacteroidales	Dysgonomonadaceae	Dysgonomonas	
004	Firmicutes_C	Negativicutes	Selenomonadales_A	Dendrosporobacteraceae	Dendrosporobacter	
005	Firmicutes	Bacilli	Lactobacillales	Vagococcaceae		
006	Fimicutes_A	Clostridia	Lachnospirales	UBA5962	UBA5962	
007	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Orbus	
008	Fimicutes_A	Clostridia	Lachnospirales	Lachnospiraceae		
009	Campylobacterota	Campylobacteria	Campylobacterales	Campylobacteraceae	Campylobacter_B	
010	Firmicutes_A	Clostridia	Lachnospirales	Lachnospiraceae		

Tabla 4.6. Taxonomía de ubicación más cercana obtenida en GTDB

User Genome	Classification	FastANI Reference	FastANI Reference Radius	FastANI Taxonomy	FastANI ANI	FastANI Alignment Fraction	Closest Placement Reference	Closest Placement Taxonomy	Closest Placement ANI	Closest Placement Alignment Fraction	Classification Method	Note	Other Related References	MSA AA Percent	RED Value
bin.002.fasta_assembly	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacterales; f__Enterobacteriaceae; g__Providencia; s__Providencia rettgeri	GCF_900455085.1	95	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacterales; f__Enterobacteriaceae; g__Providencia; s__Providencia rettgeri	97.02	0.67	GCF_900455085.1	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacterales; f__Enterobacteriaceae; g__Providencia; s__Providencia rettgeri	97.02	0.67	taxonomic classification defined by topology and ANI	topological placement and ANI have congruent species assignments	GCF_013255915.1, s__Providencia rettge...	65.34	-
bin.007.fasta_assembly	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacterales; f__Enterobacteriaceae; g__Orbus; s__	-	-	-	-	-	GCF_003634275.1	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacterales; f__Enterobacteriaceae; g__Orbus; s__Orbus hercynius	79.42	0.49	taxonomic classification defined by topology and ANI	-	-	97.38	0.9684462388

De acuerdo con Kuo *et al.* (2021), IPMB12^T y *O. hercynius* tienen varias similitudes. Estos investigadores encontraron que, si bien la distribución de genes en diferentes categorías funcionales entre todas las especies de la familia *Orbaceae* es similar (utilizando la base de datos de eggNOG), IPMB12^T y *O. hercynius* CN3^T tienen menos genes asociados con el grupo funcional de motilidad. Además de que, al realizar un mapa de calor de la distribución relativa transformada por Z-score y el número de genes que codifican las diferentes categorías funcionales de eggNOG, indicaron que, aunque presentaban diferentes patrones, la cepa IPMB12^T se agrupó con *O. hercynius* CN3^T (aisladas de un supergusano y heces de jabalí, respectivamente y con crecimiento facultativamente anaeróbico o aeróbico), y ambas cepas se separaron de los géneros *Gilliamella* y *Frischella* (aisladas del intestino de abejas o abejorros y con crecimiento microaeróbico o anaeróbico). Esta diferencia en la distribución de la función génica la atribuyeron a la correlación entre las fuentes de aislamiento de estas

cepas bacterianas, incluido el entorno de crecimiento y sus requisitos de crecimiento (Kuo et al., 2021).

Como parte de la clasificación taxonómica, se obtuvo el árbol filogenético de la Figura 4.21. Las hojas del árbol están etiquetadas con el nombre de la especie NCBI RefSeq con sus identificadores GCF. Los valores de soporte de rama que indican la confiabilidad de cada división se muestran con color rojo, entre 0 y 1, para cada nodo y los diez *bins* se encuentran resaltados en amarillo.

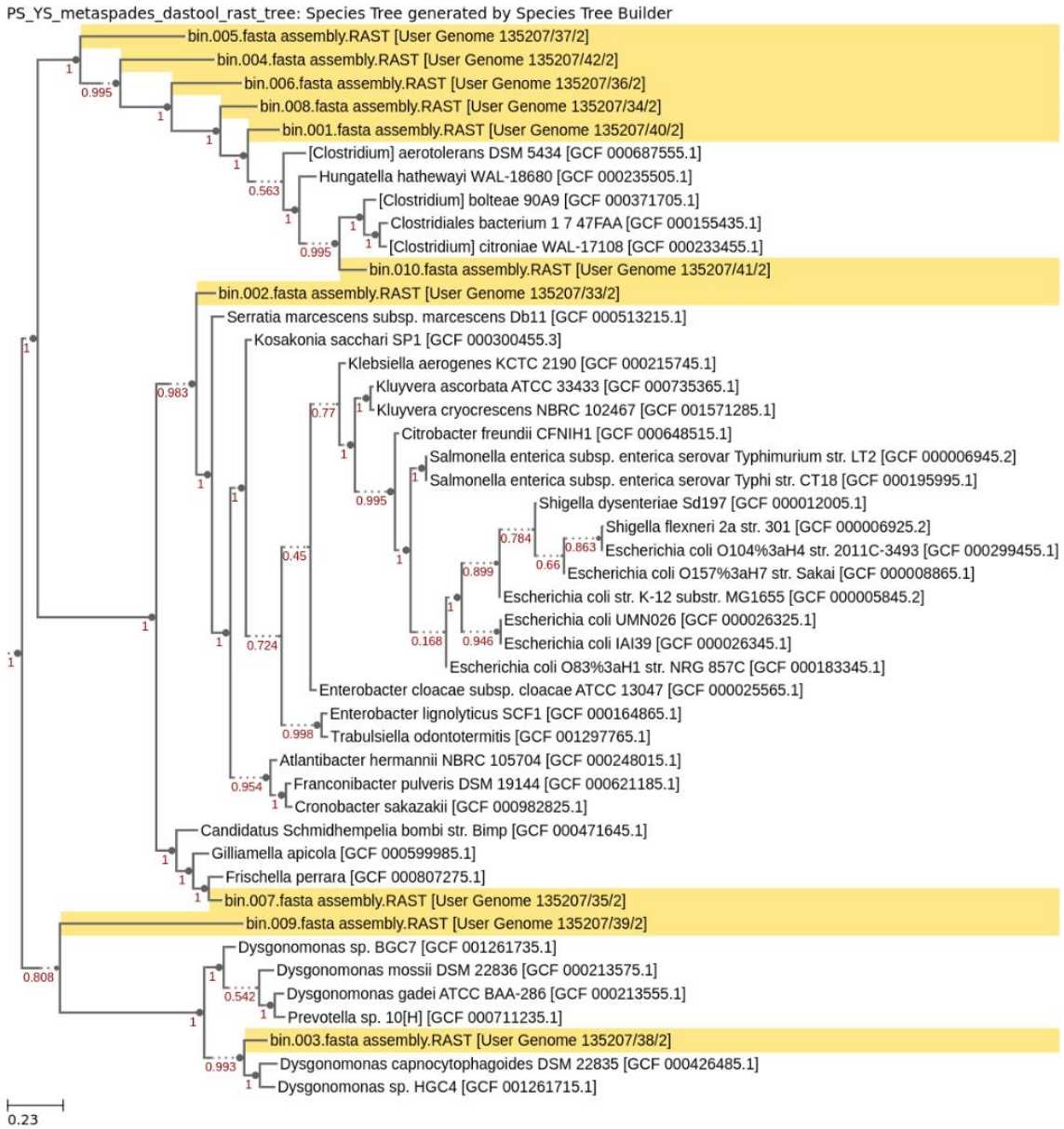


Figura 4.21. Árbol de especies generado en *Species tree*

Se puede notar que, el árbol generado se resolvió completamente pues sus nodos internos presentan dos ramas. Los microorganismos más cercanos encontrados para el *bin* 007, dentro de la familia Orbaceae, fueron de géneros *Frischella* y *Gilliamella*, representados por una cepa tipo de cada especie. Lo cual coincide con lo indicado por Wilharm (2019) y con los resultados obtenidos tanto en *Kaiju* como en *RASTtk* y *GTDB*, indicando que el *bin* 007 corresponde a *Orbus* sp. IPMB12 y al género *Orbus*, respectivamente; motivo por el cual se puede asegurar que la bacteria pertenece a la familia Orbaceae.

Además, el árbol de especies también incluye una relación entre el *bin* 007 y *Candidatus Schmidhempelia bombi*, otra gammaproteobacteria de la familia Orbaceae (NCBI, 2019).

En la Figura 4.22 se presenta el árbol filogenético construido en BV-BRC, donde se puede notar una vez más que el *bin* 007 está asignado a la familia Orbaceae, no obstante, debido a las distancias filogenéticas se puede apreciar que no corresponde ni a *Orbus* sp. IPMB12^{T} ni a *Orbus hercynius*, por lo que se podría pensar que es una especie aún no caracterizada del género Orbus.

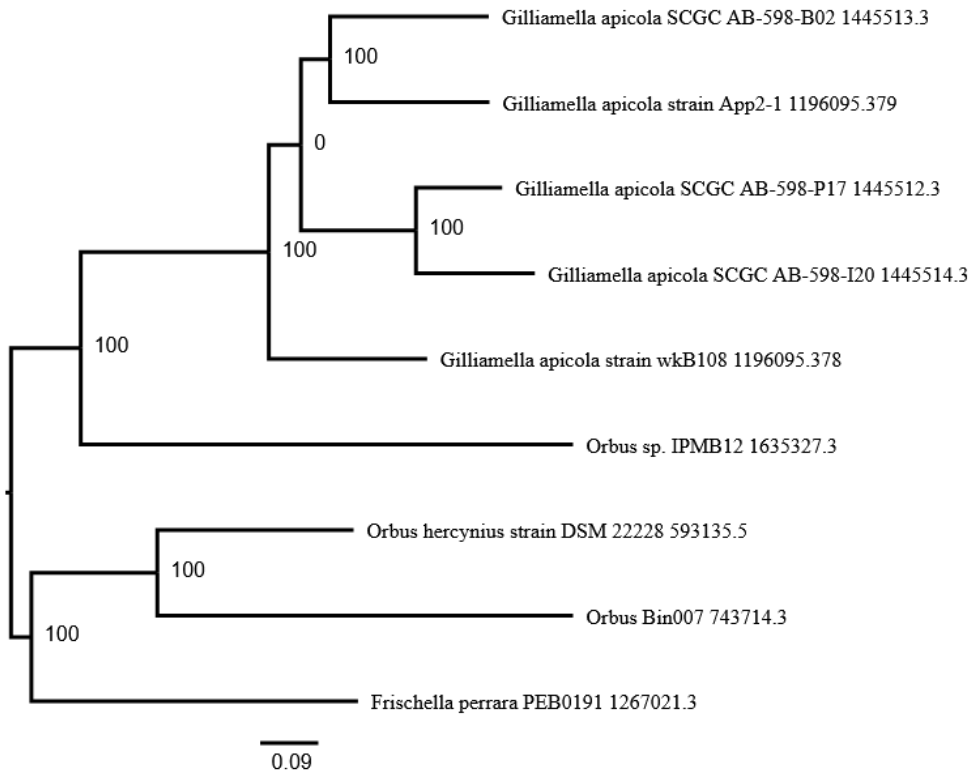


Figura 4.22. Árbol filogenético generado en BV-BRC

4.3. BÚSQUEDA DE GENES Y RUTAS METABÓLICAS DE INTERÉS

4.3.1. PERFIL FUNCIONAL O METABÓLICO DEL ORGANISMO

Los resultados de la búsqueda de genes y rutas metabólicas de interés en el genoma anotado de una proteobacteria perteneciente a la familia Orbaceae se presenta a continuación como un resumen del perfil funcional o metabólico del organismo, el cual fue proporcionado por DRAM v0.1.2. Cabe mencionar que, DRAM también proporciona información importante del genoma ensamblado, indicando que el *bin* 007 se obtuvo con 17 *scaffolds* y logró identificar 45 genes de ARNt y uno de ARNr de tipo 5S.

En la Figura 4.23 se presenta un mapa de calor interactivo que muestra la cobertura de los módulos y la cobertura de los complejos de la cadena de transporte de electrones (CTE), mientras que en la Figura 4.24 se muestra la presencia de funciones metabólicas. En la Tabla AIII.1 de los anexos se puede encontrar el número de pasos del módulo de cada ruta metabólica y el número de pasos presentes en el genoma ensamblado. Así mismo, en cuanto a los complejos de la CTE, la Tabla AIII.2 muestra el número de subunidades del módulo y el número de subunidades presentes, así como los genes.

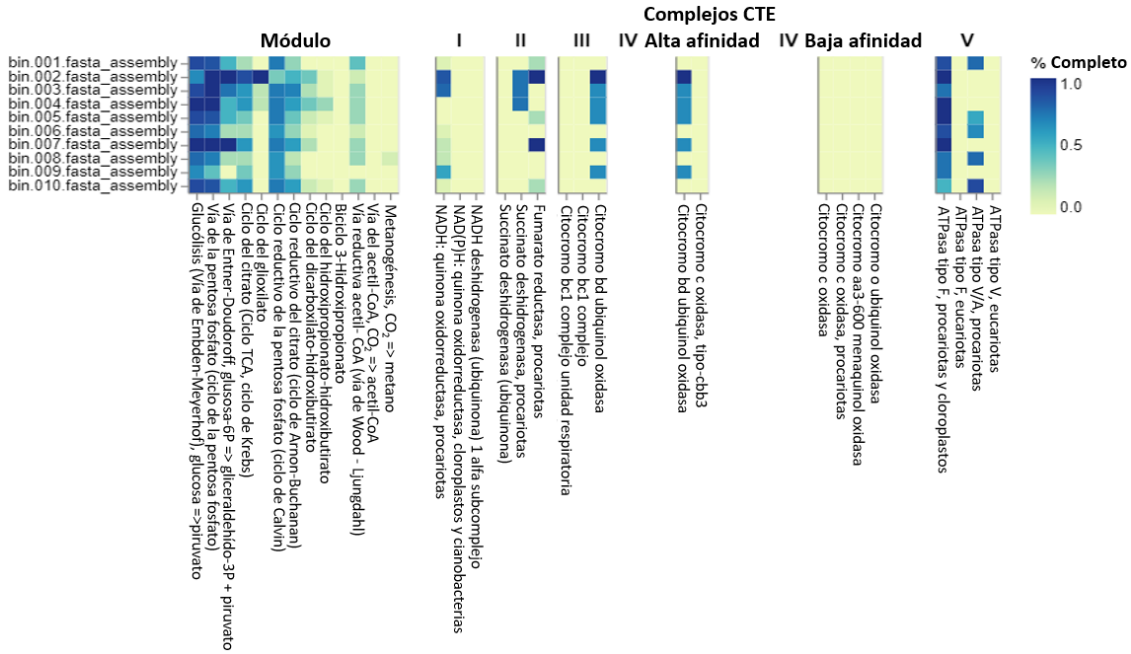


Figura 4.23. Mapa de calor de módulos y cobertura de los componentes de la cadena de transporte de electrones

La presencia de vías metabólicas para la respiración (Figura 4.23) y las funciones para el metabolismo de nitrógeno (Figura 4.24) permitieron diferenciar a las bacterias del género *Orbus* de las bacterias de géneros cercanamente relacionados como son *Gilliamella* (microaeróbica o anaeróbica) y *Frischella* (anaeróbica), pues a diferencia de estas, las de *Orbus* son capaces de desarrollarse en condiciones aeróbicas y pueden reducir el nitrato (Wilharm, 2019). Lo que concuerda con lo obtenido en la Tabla 6 y los árboles de especies de las Figura 4.22 y 4.23.

En la Figura 4.23 se puede notar que esta gammaproteobacteria cuenta con un metabolismo central que incluye las vías metabólicas de Embden-Meyerhoff (glucólisis), de la pentosa fosfato y Entner-Doudoroff, lo que le permite generar adenosín trifosfato (ATP) y precursores biosintéticos a partir del catabolismo de carbohidratos; esto es importante debido a que el sustrato de crecimiento utilizado para las larvas fue puré de banano (Pazmiño, 2021). Estas vías metabólicas también se encuentran en la bacteria *G. apicola* (de la familia Orbaceae), presente en el intestino de abeja y abejorro (Kwong et al., 2014).

Como resultado de la glucólisis se genera piruvato, que se descarboxila y oxida para la formación de la acetilcoenzimaA (acetil-CoA), razón por la cual se encuentran presentes en la bacteria las funciones metabólicas piruvato → acetil-CoA, que se observan en la Figura 4.24.

Otra ruta metabólica presente, relacionada con el metabolismo de los carbohidratos, es la del Ciclo del citrato o de Krebs (ciclo TCA), importante para los pasos finales de la oxidación de carbohidratos y ácidos grasos. En el ciclo se oxida la acetil-CoA, derivada de la glucólisis antes mencionada y la oxidación del piruvato para carbohidratos y de la oxidación beta para ácidos grasos, para formar CO_2 y suministrar NADH para la fosforilación oxidativa y otros procesos metabólicos (KEGG, 2023).

En cuanto a los complejos de la cadena de transporte de energía, se puede notar en la Figura 4.23 la presencia de citocromo bd ubiquinol oxidasa. El citocromo bd es una ubiquinol:oxígeno oxidorreductasa de la cadena respiratoria de procariotas cuya función es acoplar la reducción de oxígeno molecular, incluso en concentraciones submicromolares, al agua con la generación de una fuerza motriz de protones utilizada para la producción de ATP, lo que concuerda con la presencia de ATP-asa tipo F (Figura 4.23). Otra de sus funciones es otorgar a las bacterias resistencia a factores estresantes, como la protección contra el peróxido de hidrógeno, el óxido nítrico, el peroxinitrito y el sulfuro de hidrógeno

(Borisov et al., 2021), lo que facilita su supervivencia en condiciones hostiles como las que se presentan en los intestinos.

De acuerdo con Wilharm (2019), las gammaproteobacterias del género Orbus son quimioheterotróficas con metabolismo aeróbico y anaeróbico facultativo (Wilharm, 2019). Las bacterias quimioheterotróficas son aquellas que oxidan compuestos orgánicos en ausencia de luz. Aunque varias bacterias utilizan oxígeno para la oxidación, se conoce que algunas pueden usar nitrato, sulfato e iones férricos en lugar de oxígeno molecular para este proceso (Yamanaka, 2008). Esto concuerda con la presencia de funciones para el metabolismo de nitrógeno (Figura 4.24). Según Yamanaka (2008), este tipo de bacterias son capaces de descomponer cadáveres y excrementos animales, así como también plantas muertas (Yamanaka, 2008), lo que estaría relacionada con la eficacia de degradación de residuos por parte de *H. illucens* y concuerda también con la presencia de la función para celulosa amorfa (Figura 4.24) detectada con CAZy, una base de datos de información sobre enzimas activas en carbohidratos (CAZy, 2023); pues la celulosa es un polímero lineal que representa aproximadamente la tercera parte de los componentes de tejidos vegetales, principalmente de fibras naturales. Las zonas cristalinas son más resistentes, mientras que las partes amorfas son susceptibles a cualquier reacción química (Labrador & Osto, 2021).

Si bien las bacterias quimioheterotróficas no son capaces de fijar el carbono, de acuerdo con información de la Figura 4.23, también están presentes, aunque no por completo, los ciclos de Calvin (ciclo reductivo de las pentosas fosfato), el de Arnon-Buchanan (citrato reductor), vía Wood–Ljungdahl (vía reductora de acetil-CoA) y en menor grado el ciclo de dicarboxilato-hidroxibutirato, las cuales son vías metabólicas relacionadas con la fijación de carbono (Campbell et al., 2020; KEGG, 2022). Además, dentro de los complejos de la CTE se presenta la fumarato reductasa que, como se muestra en la Figura AIII también está involucrada en la fijación de carbono. De acuerdo con Khademian e Imlay (2021) la enzima fumarato reductasa proporciona una ruta para la respiración anaerobia y cede electrones al oxígeno, la cual es la principal diferencia con la enzima succinato deshidrogenasa que es su equivalente en la respiración aerobia que impulsa el ciclo TCA (Khademian & Imlay, 2021).

En la Figura AIII.1 se puede encontrar un diagrama con todas las rutas metabólicas presentes en la bacteria seleccionada; mientras que, en la Figura AIII.2 se resaltan el ciclo de Arnon-Buchanan (citrato reductor; M00173), la vía Wood–Ljungdahl (vía reductora de acetil-CoA;

M00377), el ciclo de dicarboxilato-hidroxibutirato (M00374) y la enzima fumarato reductasa (1.3.1.6).

Los organismos procariotas también pueden llevar a cabo una respiración no asimilativa de arsénico, en la cual, se reduce el As(V) a As(III) utilizando diferentes compuestos inorgánicos y orgánicos (materia orgánica) como donadores de electrones y como aceptor final el arseniato para la producción de energía; aunque también pueden utilizarse otras moléculas como aceptores, tales como fumarato, nitrito, nitrato, oxígeno, entre otras (Rodarte, 2017).

En algunas investigaciones se ha reportado que la reducción del arseniato está asociada con el uso de acetato o lactato como agente reductor (Kruger, Bertin, Heipieper, & Arsène-Ploetze, 2013; Macy, Santini, Pauling, O'Neill, & Sly, 2000; Newman et al., 1997; Niggemyer, Spring, Stackebrandt, & Rosenzweig, 2001). Por lo que, la presencia de las funciones: reducción de arsenato pt1 (reductasa), acetato pt2 y lactato D (Figura 4.24) pueden relacionarse con una resistencia al arsénico ya sea por este tipo de respiración (relacionada con genes arrA y arrB) en la que los microorganismos obtienen energía o por una desintoxicación donde la reducción del arseniato se da sin la generación de energía (relacionada con el gen arsC) (Kruger et al., 2013; Macy et al., 2000; Rodarte, 2017).

Otro proceso para la adquisición de energía de las bacterias quimioheterotróficas es la fermentación, donde los compuestos orgánicos se oxidan anaeróbicamente por compuestos orgánicos para producir ATP. Por ejemplo, en la fermentación de alcohol, el gliceraldehído-3-fosfato se oxida a acetaldehído y este a su vez es reducido a etanol (Yamanaka, 2008). Esto concuerda con la presencia de la función de producción de alcohol (Figura 4.23). Así mismo, la fermentación contribuye a la descomposición de compuestos orgánicos y la limpieza del medio ambiente cuando el oxígeno no está disponible (Yamanaka, 2008). Lo cual refuerza la idea de que esta gammaproteobacteria contribuye en la bioconversión de residuos en las LMSN tanto con su metabolismo aeróbico como con su metabolismo anaeróbico facultativo.

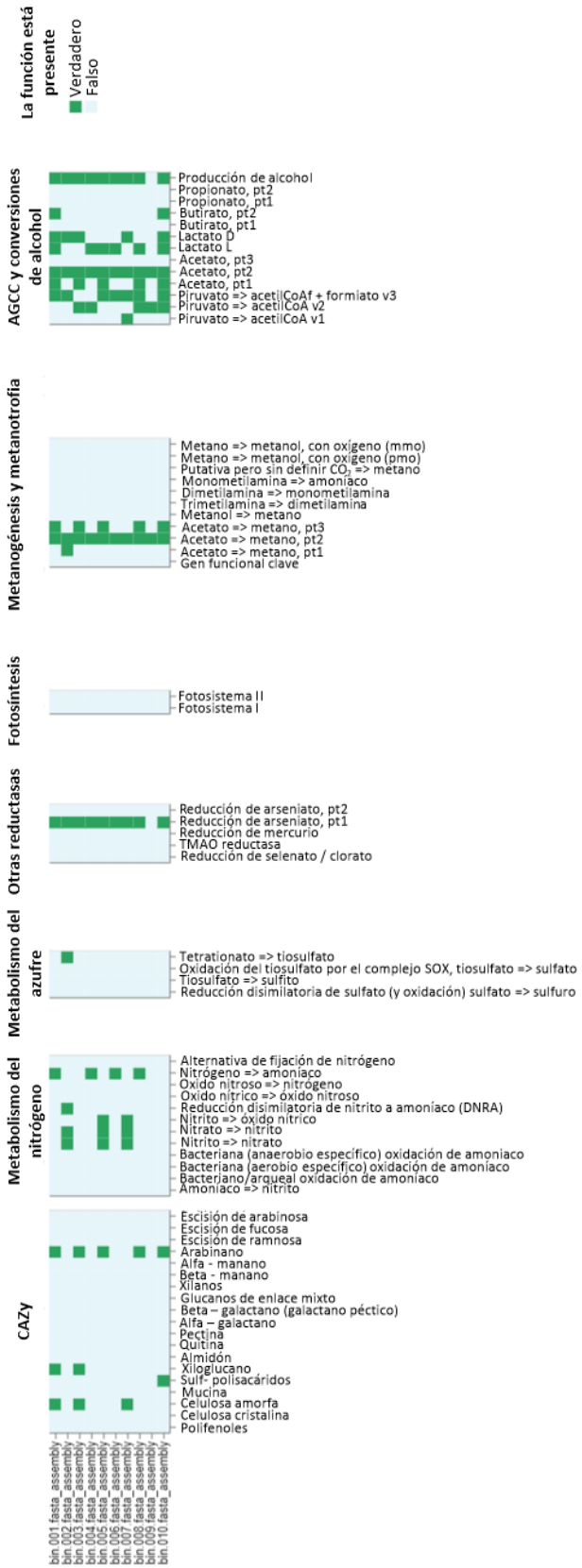


Figura 4.24. Funciones metabólicas presentes

* Ácidos grasos de cadena corta (AGCC)

4.3.2. GENES RELACIONADOS CON LA HIDRÓLISIS DE PET

Uno de los principales filos en los que se producen PET hidrolasas es el de las Proteobacterias (Danso et al., 2018) y de estas, las principales pertenecen a las Beta, Delta y Gamma proteobacterias. Así, el resultado de utilizar las herramientas *Build FeatureSet from Genome* v1.2.6 y *Merge FeatureSets* v1.7.4 fue un conjunto de características con los siguientes genes pertenecientes a las PETasas: Cut190 (W0TJ64), cut1 (E9LVI0), cut-2 (E5BBQ3), Tcur_1278 (D1A9G5), cut1 (E9LVH7), cut (H6WX58), cut2 (E9LVH9), ISF6_4831 (A0A0K8P6T7), Deinma_1209, liplAF5-2_UB, J057_15340_Mna, pbsA, est_Psyc OLEAN_C07960_Oant.

MUSCLE v3.8.425 generó una alineación de secuencias múltiple de secuencias de proteínas que puede descargarse en formato FASTA y Clustalw y se almacenó como un objeto de datos MSA, que contiene las secuencias de alineación, las etiquetas de las filas, el orden de las filas y la descripción.

Con Gblocks v0.91b se obtuvo una MSA recortada que mantuvo solo los bloques conservados, que pueden ser regiones más confiables para una posterior búsqueda de funciones en el genoma ensamblado.

En la Tabla 4.7 se presenta el informe con los aciertos de la búsqueda realizada con *HMMER Search from MSA (prot-prot)* v3.3.2, junto con los conjuntos de funciones a las que se accede.

Tabla 4.7. Coincidencias de la búsqueda con *HMMER Search from MSA*

Cobertura de alineación (HIT SEQ)	ID del gen	Función	Genoma	ALN_LON	Valor E	Puntuación	H_Ini-H_Fin
	fasta_assembly_NODE_2_length_650948_cov_26.124664_116	Proteína no caracterizada	bin.007.fasta_assembly_DRAM	125 (50.2%)	3.2e-08	30.6	67-191
	fasta_assembly_NODE_2_length_650948_cov_26.124664_180	Proteína no caracterizada	bin.007.fasta_assembly_DRAM	122 (49.0%)	1.6e-06	25.0	70-191
	fasta_assembly_NODE_64_length_93995_cov_27.803758_15	carboxilesterasa [EC:3.1.1.1]	bin.007.fasta_assembly_DRAM	92 (36.9%)	0.00021	18.1	14-105

Como se puede notar en la tabla, no se encontraron coincidencias con una alta significancia y es poco probable que existan PETasas, por lo menos de esa familia de proteínas, en el genoma ensamblado. Esto no descarta la posibilidad de que existan otras enzimas relacionadas con la degradación de PET. Aunque, también existe la posibilidad de que no hayan suficientes genes en el alineamiento para obtener un buen perfil.

5. CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

El 94,93 % de las lecturas (20 033 887) de los datos metagenómicos obtenidos a partir del tracto digestivo de larvas de mosca soldado negra (*Hermetia illucens*) pasaron el filtro de calidad, mientras que el restante 5,07 % de las lecturas fueron eliminadas por su baja calidad. La longitud de secuencia más corta fue de 50 pb, tanto para la secuencia forward como para la secuencia reverse. Para los 150 nucleótidos de ambas secuencias los puntajes de calidad fueron altos o de buena calidad (superior a 30), lo cual es adecuado para el análisis bioinformático.

De las 40 067 774 lecturas de datos metagenómicos, el 11,76 % estuvieron duplicadas, es decir el 12 % de los nucleótidos iniciales.

El 59,69 % de lecturas lecturas sin duplicar (21 102 394) se alinearon y mapearon al genoma de referencia del hospedero (*Hermetia illucens*).

Metaspades logró obtener 5 565 *contigs* mayores o iguales a 1 000 pb, 661 *contigs* mayores o iguales a 10 000 pb y 60 mayores o iguales a 100 000 pb. El contig más largo obtenido fue de 706 116 pb y un N50 de 22 421 pb.

De los 5 565 *contigs* obtenidos en el ensamblaje, el 38,52 % fueron agrupados y optimizados mediante desreplicación, agregación y puntuación en un total de 10 *bins* considerados como de alta calidad (BinScore > 0,7), pues únicamente se seleccionaron genomas considerados completos (> 90 %), con una contaminación menor al 3 % y genomas considerados borrador (> 70 %), con una contaminación entre 1,35 y 15,05 %.

Se logró ensamblar el genoma de una proteobacteria perteneciente a la familia Orbaceae a partir de datos metagenómicos libres de lecturas del hospedero, con un contenido de guanina citocina del 35,65 % y una longitud total de 3 084 216 bp (nucleótidos). El genoma se obtuvo con 17 scaffolds (*bin* 007) y se consideró completo (99,44 %), con una contaminación mínima (0,05 %).

Los metadatos predeterminados, considerados esenciales del ensamblaje, fueron: código genético 11 (genomas de bacterias y arqueas), dominio bacteriano y el nombre científico "taxón desconocido". Con herramientas tales como *KAIJU, RASTtk,* (*GTDB)-Tk, Species tree* y BV-BRC se logró establecer que el genoma ensamblado pertenece a una Gammaproteobacteria de la familia Orbaceae y posiblemente corresponde a una especie aún no caracterizada del género Orbus.

La anotación del genoma ensamblado de la proteobacteria perteneciente a la familia Orbaceae resultó en la identificación de 2 881 genes, de los cuales 2 803 (97,29 %) fueron codificantes de proteínas, mientras que los 78 restantes (2,71 %) fueron no codificantes, con 32 (1,11 %) de repetición sin codificación y 46 (1,60 %) de ARN no codificante. De estos, 45 genes fueron reconocidos como ARNt y uno como ARNr de tipo 5S.

El metabolismo central de la gammaproteobacteria incluyó las siguientes vías de utilización de carbono: Embden-Meyerhoff (glucólisis), ciclo de las pentosas fosfato y Entner-Doudoroff con todos sus pasos presentes, 9, 7 y 4, respectivamente. También se encontraron vías metabólicas incompletas como la del ciclo de citrato o de Krebs (62,5 %) y cuatro vías de fijación de carbono: ciclo reductor de pentosa fosfato (ciclo de Calvin; 72,7 %), ciclo reductor de citrato (Arnon-Buchanan; 60 %), ciclo de dicarboxilato-hidroxibutirato (23 %), y la vía reductora de acetil-CoA (Wood–Ljungdahl; 28,6 %).

Se identificaron los 4 genes presentes de la enzima Fumarato reductasa (K00244, K00245, K00246, K00247) y los 8 de la ATPasa tipo-F (K02108, K02109, K02110, K02111, K02112, K02113, K02114, K02115), mientras que, de la Citocromo bd ubiquinol oxidasa se identificaron únicamente dos de los tres genes (K00425, K00426) y dos de los 11 genes de la NADH:quinona oxidorreductasa (K00340, K00343).

La presencia de rutas metabólicas y funciones vinculadas con la utilización de carbono, nitrógeno, celulosa amorfa y producción de alcohol está relacionada con la eficacia de degradación de residuos orgánicos por parte de esta bacteria en los intestinos de *H. illucens* en presencia o no de oxígeno, lo cual refuerza la idea de que esta gammaproteobacteria contribuye en la bioconversión de residuos en las LMSN tanto con su metabolismo aeróbico como con su metabolismo anaeróbico facultativo.

No se encontraron coincidencias de genes pertenecientes a las PETasas con una alta significancia y es poco probable que existan esas enzimas en el genoma ensamblado, pero no se descarta la posibilidad de que existan otras enzimas relacionadas con la degradación

de PET. Aunque, también existe la posibilidad de que no haya suficientes genes en el alineamiento para obtener un buen perfil.

5.2. RECOMENDACIONES

Realizar un análisis de correlación entre genes y rutas metabólicas con los microorganismos de mayor abundancia relativa.

Realizar un análisis comparativo de las rutas metabólicas de otros microorganismos de abundancia relativa alta en el intestino de las LMSN para determinar si existe o no una simbiosis en la interacción bacteria – bacteria y bacteria – hospedero.

Determinar la presencia de los genes arrA y arrB o arsC para definir si la resistencia al arsénico se debe a una respiración no asimilativa de arsénico o por desintoxicación.

6. REFERENCIAS

Allen, B. (2021). Genomics in KBase: Identifying Features of Interest in Genomes. Retrieved July 15, 2022, from https://narrative.kbase.us/narrative/83681#

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., … Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1146. https://doi.org/10.1038/nmeth.3103

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., … Yu, D. (2018). KBase: The United States department of energy systems biology knowledgebase. *Nature Biotechnology*, *36*(7), 566–569. https://doi.org/10.1038/nbt.4163

Babraham Institute. (2020). FastQC: Documentation. Retrieved September 4, 2022, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Beskin, K. V., Holcomb, C. D., Cammack, J. A., Crippen, T. L., Knap, A. H., Sweet, S. T., & Tomberlin, J. K. (2018). Larval digestion of different manure types by the black soldier fly (Diptera: Stratiomyidae) impacts associated volatile emissions. *Waste Management*, *74*, 213–220. https://doi.org/10.1016/j.wasman.2018.01.019

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borisov, V. B., Siletsky, S. A., Paiardini, A., Hoogewijs, D., Forte, E., Giuffrè, A., & Poole, R. K. (2021). Bacterial Oxidases of the Cytochrome bd Family: Redox Enzymes of Unique Structure, Function, and Utility As Drug Targets. *Antioxidants and Redox Signaling*, *34*(16), 1280–1318. https://doi.org/10.1089/ars.2020.8039

Brandon, A. M., Gao, S. H., Tian, R., Ning, D., Yang, S. S., Zhou, J., … Criddle, C. S. (2018). Biodegradation of Polyethylene and Plastic Mixtures in Mealworms (Larvae of Tenebrio molitor) and Effects on the Gut Microbiome. *Environmental Science and Technology*, *52*(11), 6526–6533. https://doi.org/10.1021/acs.est.8b02301

BRC. (n.d.). Bacterial and Viral Bioinformatics Resource Center. Retrieved January 31, 2023, from https://www.bv-brc.org/

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4), 1125–1139. https://doi.org/10.1093/bib/bbx120

Brent, R. (2000). Genomic Biology. *Cell*, *100*(1), 169–183. https://doi.org/10.1016/S0092-8674(00)81693-1

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., … Xia, F. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, *5*. https://doi.org/10.1038/srep08365

Bruno, D., Bonelli, M., De Filippis, F., Di Lelio, I., Tettamanti, G., Casartelli, M., … Caccia, S. (2019). The intestinal microbiota of Hermetia illucens larvae is affected by diet and shows a diverse composition in the different midgut regions. *Applied and Environmental Microbiology*, *85*(2), 1–14. https://doi.org/10.1128/AEM.01864-18

Campbell, M. A., Grice, K., Visscher, P. T., Morris, T., Wong, H. L., White, R. A., … Coolen, M. J. L. (2020). Functional Gene Expression in Shark Bay Hypersaline Microbial Mats: Adaptive Responses. *Frontiers in Microbiology*, *11*(November), 1–16. https://doi.org/10.3389/fmicb.2020.560336

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

CAZy. (2023). Carbohydrate Active Enzymes database. Retrieved February 10, 2023, from http://www.cazy.org/Citing-CAZy

Cedeño, A. (2021). *Efecto de la harina de mosca sobre el crecimiento de alevines de tilapia ngera (Oreochromis niloticus)*. Universidad de Guayaquil. Retrieved from http://repositorio.ug.edu.ec/handle/redug/52753

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Chivian, D., Clark, M., & Jungbluth, S. (2020). Metagenome-Assembled Genome Extraction from a Compost Microbiome Enrichment. Retrieved July 15, 2022, from https://narrative.kbase.us/narrative/33233

Cho, S., Kim, C. H., Kim, M. J., & Chung, H. (2020). Effects of microplastics and salinity on food waste processing by black soldier fly (Hermetia illucens) larvae. *Journal of Ecology and Environment, 44*(1), 1–9. https://doi.org/10.1186/s41610-020-0148-x

Chow E. (USFC). (2019). Next Generation Sequencing 1: Overview. Retrieved August 26, 2022, from https://www.ibiology.org/techniques/next-generation-sequencing/

Čičková, H., Newton, G. L., Lacy, R. C., & Kozánek, M. (2015). The use of fly larvae for organic waste treatment. *Waste Management, 35*, 68–80. https://doi.org/10.1016/j.wasman.2014.09.026

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/gigascience/giab008

Danso, D., Schmeisser, C., Chow, J., Zimmermann, W., Wei, R., Leggewie, C., … Streit, W. R. (2018). New insights into the function and global distribution of polyethylene terephthalate (PET)-degrading bacteria and enzymes in marine and terrestrial metagenomes. *Applied and Environmental Microbiology*, *84*(8). https://doi.org/10.1128/AEM.02773-17

Davis, J. J., Gerdes, S., Olsen, G. J., Olson, R., Pusch, G. D., Shukla, M., … Yoo, H. (2016). PATtyFams: Protein families for the microbial genomes in the PATRIC database. *Frontiers in Microbiology*, *7*(FEB), 1–12. https://doi.org/10.3389/fmicb.2016.00118

De Smet, J., Wynants, E., Cos, P., & Van Campenhout, L. (2018). Microbial community dynamics during rearing of black soldier fly larvae (Hermetia illucens) and impact on exploitation potential. *Applied and Environmental Microbiology*, *84*(9). https://doi.org/10.1128/AEM.02722-17

Del Hierro, A., Anrango, M., Ortiz, D., & Sánchez, L. (2021). Captura y cría de la mosca soldado negra (Hermetia Illucens) para la biodegradación de desechos orgánicos en Puerto Quito, Ecuador. *Ecuadorian Science Journal*, *5*(3), 341–354. https://doi.org/10.46480/esj.5.3.164

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

ESR International. (2016). Bio-Conversion of Putrescent Waste. Retrieved July 16, 2022, from https://arquivo.pt/wayback/20160516163727/http%3A//www.esrint.com/pages/bioconversion.html

FAO. (2012). *Pérdida y desperdicio de alimentos en el mundo - Alcance, causas y prevención* (Vol. 39). Roma. Retrieved from https://mrv.dnp.gov.co/Documentos de Interes/Perdida_y_Desperdicio_de_Alimentos_en_colombia.pdf

FAO. (2015). Food wastage footprint & climate change, (1), 1–4. Retrieved from https://www.uncclearn.org/learning-resources/library/10458

FAO. (2017). Hacia un Código Internacional de Conducta para la Prevención y Reducción de Pérdidas y Desperdicios de Alimentos. Retrieved from https://www.fao.org/3/i7338s/i7338s.pdf

FAO. (2019). The State of Food and Agriculture 2019. Moving forward on food loss and waste reduction.Jenkins, Willis Tucker,. Rome. https://doi.org/10.4324/9781315764788

FAO. (2022a). Pérdidas y desperdicios de alimentos en América Latina y el Caribe. Retrieved July 17, 2022, from https://www.fao.org/americas/noticias/ver/es/c/239393/#:~:text=La FAO calcula que dichos,-cosecha%2C almacenamiento y transporte.

FAO. (2022b). Tackling food loss and waste: A triple win opportunity. Retrieved February 2, 2023, from https://www.fao.org/newsroom/detail/FAO-UNEP-agriculture-environment-food-loss-waste-day-2022/en

Garrity, G. M., Bell, J. A., & Lilburn, T. (2015a). Gammaproteobacteria class. nov. *Bergey's Manual of Systematics of Archaea and Bacteria*. https://doi.org/10.1002/9781118960608.cbm00045

Garrity, G. M., Bell, J. A., & Lilburn, T. (2015b). Proteobacteria phyl. nov. *Bergey's Manual of Systematics of Archaea and Bacteria*. https://doi.org/10.1002/9781118960608.pbm00022

HLPE. (2014). Las pérdidas y el desperdicio de alimentos en el contexto de sistemas alimentarios sostenibles, 23–25. Retrieved from http://www.fao.org/3/a-i3901s.pdf

Holalla, A. (2021). *Diseño de un sistema de crianza de Hermetia illucens (mosca soldado negro) para la producción de pie de cría*. Escuela Superior Politécnica Del Litoral. Retrieved from https://www.dspace.espol.edu.ec/handle/123456789/53192

Illumina. (2022). Introduction to NGS. Retrieved September 11, 2022, from https://www.illumina.com/science/technology/next-generation-sequencing.html

INEC. (2017). VDatos. Gestión de Residuos: Residuos. Retrieved July 13, 2022, from https://aplicaciones3.ecuadorencifras.gob.ec/VDATOS2-war/paginas/administracion/visualizador.xhtml

INEC. (2020). Modulo de Informacion Ambiental en Hogares - ESPND, 2019. *Boletín Técnico - INEC*. Retrieved from https://www.ecuadorencifras.gob.ec/documentos/web-inec/Encuestas_Ambientales/Hogares/Hogares 2019/BOL_TEC_AMB_ESPND_2019_11.pdf

INEC. (2021). Estadística de Información Ambiental Económica en Gobiernos Autónomos Descentralizados Municipales Gestión de Agua Potable y Saneamiento 2020 Resumen Estadístico. Retrieved from https://www.ecuadorencifras.gob.ec/gad-municipales/

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, *9*(1), 1–8. https://doi.org/10.1038/s41467-018-07641-9

Jiang, C. L., Jin, W. Z., Tao, X. H., Zhang, Q., Zhu, J., Feng, S. Y., … Zhang, Z. J. (2019). Black soldier fly larvae (Hermetia illucens) strengthen the metabolic function of food waste biodegradation by gut microbiome. *Microbial Biotechnology*, *12*(3), 528–543. https://doi.org/10.1111/1751-7915.13393

Johns Hopkins University. (2022). Bowtie 2. Retrieved September 6, 2022, from http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

Jung, H., Ventura, T., Sook Chung, J., Kim, W. J., Nam, B. H., Kong, H. J., … Eyun, S. Il. (2020). Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Computational Biology*, *16*(11), 1–25. https://doi.org/10.1371/journal.pcbi.1008325

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, *2015*(8), 1–15. https://doi.org/10.7717/peerj.1165

KBase. (2019). Compare Assembled Contig Distributions - v1.1.2. Retrieved January 6, 2023, https://kbase.us/applist/apps/kb_assembly_compare/run_contig_distribution_compare/release#:~:text=Compare Assembled Contig Distributions allows,contigs are typically more desirable.

KEGG. (2021). Carbon metabolism -Reference pathway. Retrieved January 27, 2023, from https://www.genome.jp/pathway/map01200

KEGG. (2022). Carbon fixation pathways in prokaryotes. Retrieved January 25, 2023, from https://www.genome.jp/pathway/ec00720+1.3.1.6

KEGG. (2023). KEGG PATHWAY. Retrieved January 12, 2023, from https://www.genome.jp/kegg/pathway.html

Khademian, M., & Imlay, J. A. (2021). How Microbes Evolved to Tolerate Oxygen. *Trends in Microbiology*, *29*(5), 428–440. https://doi.org/10.1016/j.tim.2020.10.001

Khamis, F. M., Ombura, F. L. O., Akutse, K. S., Subramanian, S., Mohamed, S. A., Fiaboe, K. K. M., … Tanga, C. M. (2020). Insights in the Global Genetics and Gut Microbiome of Black Soldier Fly, Hermetia illucens: Implications for Animal Feed Safety Control. *Frontiers in Microbiology*, *11*(July), 1–15. https://doi.org/10.3389/fmicb.2020.01538

Kim, W., Bae, S., Park, K., Lee, S., Choi, Y., Han, S., & Koh, Y. (2011). Biochemical characterization of digestive enzymes in the black soldier fly, Hermetia illucens (Diptera: Stratiomyidae). *Journal of Asia-Pacific Entomology*, *14*(1), 11–14. https://doi.org/10.1016/j.aspen.2010.11.003

Kruger, M. C., Bertin, P. N., Heipieper, H. J., & Arsène-Ploetze, F. (2013). Bacterial metabolism of environmental arsenic - Mechanisms and biotechnological applications. *Applied Microbiology and Biotechnology*, *97*(9), 3827–3841. https://doi.org/10.1007/s00253-013-4838-5

Kuan, Z. J., Chan, B. K. N., & Gan, S. K. E. (2022). Worming the Circular Economy for Biowaste and Plastics: Hermetia illucens, Tenebrio molitor, and Zophobas morio. *Sustainability (Switzerland)*, *14*(3), 1–13. https://doi.org/10.3390/su14031594

Kuo, C. H., Huang, P. Y., Sheu, S. Y., Sheu, D. S., Jheng, L. C., & Chen, W. M. (2021). Zophobihabitans entericus gen. Nov., sp. nov., a new member of the family orbaceae isolated from the gut of a superworm Zophobas morio. *International Journal of Systematic and Evolutionary Microbiology*, *71*(11). https://doi.org/10.1099/ijsem.0.005081

Kwong, W. K., Engel, P., Koch, H., & Moran, N. A. (2014). Genomics and host specialization of honey bee and bumble bee gut symbionts. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(31), 11509–11514. https://doi.org/10.1073/pnas.1405838111

Kwong, W. K., & Moran, N. A. (2013). Cultivation and characterization of the gut symbionts of honey bees and bumble bees: Description of Snodgrassella alvi gen. nov., sp. nov., a member of the family Neisseriaceae of the betaproteobacteria, and Gilliamella apicola gen. nov., sp. nov., a memb. *International Journal of Systematic and Evolutionary Microbiology*, *63*(PART6), 2008–2018. https://doi.org/10.1099/ijs.0.044875-0

Labrador, H., & Osto, S. (2021). Caracterización de la celulosa proveniente del lodo papelero y su esterificación. *Revista de La Facultad de Ciencias de La Universidad Nacional de Colombia*, *10*(2). Retrieved from http://portal.amelica.org/ameli/journal/115/1152771006/html/

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, D., Liu, C.-M., Ruibang, L., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and com- plex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1, 2. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25609793

Li, Q., Zheng, L., Cai, H., Garza, E., Yu, Z., & Zhou, S. (2011). From organic waste to biodiesel: Black soldier fly, Hermetia illucens, makes it feasible. *Fuel*, *90*(4), 1545–1548. https://doi.org/10.1016/j.fuel.2010.11.016

Lievens, S., Poma, G., Frooninckx, L., Van der Donck, T., Seo, J. W., De Smet, J., … Van Der Borght, M. (2022). Mutual Influence between Polyvinyl Chloride (Micro)Plastics and Black Soldier Fly Larvae (Hermetia illucens L.). *Sustainability (Switzerland)*, *14*(19). https://doi.org/10.3390/su141912109

Liu, Y. X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., & Bai, Y. (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein & Cell*, *12*(5), 315–330. https://doi.org/10.1007/s13238-020-00724-8

LPSN. (2021). Genus Zophobihabitans. Retrieved February 4, 2023, from https://lpsn.dsmz.de/genus/zophobihabitans

Luo, L., Wang, Y., Guo, H., Yang, Y., Qi, N., Zhao, X., … Zhou, A. (2021). Biodegradation of foam plastics by Zophobas atratus larvae (Coleoptera: Tenebrionidae) associated with changes of gut digestive enzymes activities and microbiome. *Chemosphere*, *282*(11), 131006. https://doi.org/10.1016/j.chemosphere.2021.131006

Macy, J. M., Santini, J. M., Pauling, B. V., O'Neill, A. H., & Sly, L. I. (2000). Two new arsenate/sulfate-reducing bacteria: Mechanisms of arsenate reduction. *Archives of Microbiology*, *173*(1), 49–57. https://doi.org/10.1007/s002030050007

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538. https://doi.org/10.1186/1471-2105-11-538

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms11257

Mitra, B., & Das, A. (2023). The Ability of Insects to Degrade Complex Synthetic Polymers. In V. Shields (Ed.), *Arthropods - New Advances and Perspectives [Working Title]*. https://doi.org/10.5772/intechopen.106948

Morales, J. (2021). *Biotransformación de residuos orgánicos a partir del manejo ex situ de Hermetia illucens (L., 1758) (Diptera:Stratiomyidae) como una alternativa para la gestión sostenible de los desechos sólidos en el Distrito Metropolitano de Quito*. Universidad Central del Ecuador. Retrieved from http://www.dspace.uce.edu.ec/handle/25000/23015

Morán, S. (2020). Ecuador, ahogado en basura, está lejos de cumplir las metas de los ODS al 2030. Retrieved July 13, 2022, from https://www.planv.com.ec/historias/sociedad/ecuador-ahogado-basura-esta-lejos-cumplir-metas-ods-al-2030

NCBI. (2019). Candidatus Schmidhempelia bombi. Retrieved December 29, 2022, from https://www.ncbi.nlm.nih.gov/data-hub/taxonomy/1505866/

NCBI. (2020). Hermetia illucens reference genome iHerIll2.2.curated.20191125. Retrieved December 23, 2022, from https://www.ncbi.nlm.nih.gov/search/all/?term=GCF_905115235.1_iHerIll2.2.curated.20191125

Newman, D. K., Kennedy, E. K., Coates, J. D., Ahmann, D., Ellis, D. J., Lovley, D. R., & Morel, F. M. M. (1997). Dissimilatory arsenate and sulfate reduction in Desulfotomaculum auripigmentum sp. nov. *Archives of Microbiology*, *168*(5), 380–388. https://doi.org/10.1007/s002030050512

Niggemyer, A., Spring, S., Stackebrandt, E., & Rosenzweig, F. (2001). Isolation and Characterization of a Novel As(V)-Reducing Bacterium: Implications for Arsenic Mobilization and the Genus Desulfitobacterium. *Applied and Environmental Microbiology*, *67*(12), 5568–5580. https://doi.org/10.1128/AEM.67.12.5568–5580.2001

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Oliveira, F., Doelle, K., List, R., & Reilly, J. R. O. (2015). Assessment of Diptera: Stratiomyidae , genus Hermetia illucens (L ., 1758) using electron microscopy. *Journal of Entomology and Zoology Studies*, *3*(5), 147–152.

Oliver, P. (2004). Disposal apparatus and method for efficiently bio-converting putrescent wastes. US6780637B2. Washington. Retrieved from https://patents.google.com/patent/US6780637B2/en

Olm, M. (2017). Why genome completeness and contamination estimates are more complicated than you think. Retrieved April 22, 2023, from https://microbe.net/2017/12/13/why-genome-completeness-and-contamination-estimates-are-more-complicated-than-you-think/

ONU. (2021). Se desperdicia 17 % de todos los alimentos disponibles a nivel del consumidor. Retrieved April 29, 2023, from https://www.unep.org/es/noticias-y-reportajes/comunicado-de-prensa/onu-se-desperdicia-17-de-todos-los-alimentos-disponibles

Osimani, A., Ferrocino, I., Corvaglia, M. R., Roncolini, A., Milanović, V., Garofalo, C., … Clementi, F. (2021). Microbial dynamics in rearing trials of Hermetia illucens larvae fed coffee silverskin and microalgae. *Food Research International*, *140*(July 2020). https://doi.org/10.1016/j.foodres.2020.110028

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055.

https://doi.org/10.1101/gr.186072.114

Pazmiño, M. (2021). *Caracterización molecular y evaluación de la capacidad de degradación de microplástico de insectos del género Hermetia originarios del cantón Puerto Quito*. Universidad de las Fuerzas Armadas ESPE.

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

PNUD. (2015). Objetivos de desarrollo sostenible. Retrieved from https://www.undp.org/content/undp/es/home/librarypage/corporate/sustainable-development-goals-booklet.html

PNUMA, W. (2021). UNEP Food Waste Index Report 2021. *UN Environment Programme*, 80. Retrieved from https://www.unep.org/resources/report/unep-food-waste-index-report-2021

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3). https://doi.org/10.1371/journal.pone.0009490

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844. https://doi.org/10.1038/nbt.3935

Quinlan, A. (2021). bedtools2. Retrieved January 6, 2023, from https://github.com/arq5x/bedtools2/releases

Quinlan, A., & Kindlon, N. (2022). bedtools: a powerful toolset for genome arithmetic. Retrieved September 12, 2022, from https://bedtools.readthedocs.io/en/latest/

Rodarte, L. (2017). *Bioprospección de microorganismos con resistencia a metales de sitios contaminados con arsénico*. Instituto Potosino de Investigación Científica y Tecnológica. Retrieved from https://ipicyt.repositorioinstitucional.mx/jspui/bitstream/1010/1630/1/TMIPICYTR6B52017.pdf

Romero, M. (2022). *Modelización del ciclo de vida de la mosca soldado negro (Hermetia illucens) desarrollándose sobre desechos orgánicos*. Universidad Central del Ecuador. Retrieved from http://www.dspace.uce.edu.ec/handle/25000/25656

Salguero, Y., & Guevara, A. (2019). El primer banco de alimentos del Ecuador creado por docentes de la Escuela Politécnica Nacional. *MktDESCUBRE*, 38–49. https://doi.org/10.36779/mktdescubre.v13.193

Sangwan, N., Xia, F., & Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, *4*, 1–11. https://doi.org/10.1186/s40168-016-0154-5

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L. 0003 3527 8101, Solden, L. M., … Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*, *48*(16), 8883–8900. https://doi.org/10.1093/nar/gkaa621

Shan, J., Su, T., Zhao, J., & Wang, Z. (2021). Isolation, Identification, and Characterization of Polystyrene-Degrading Bacteria From the Gut of Galleria Mellonella (Lepidoptera: Pyralidae) Larvae. *Front Bioeng Biotechnol*, *9*(736062). https://doi.org/10.3389/fbioe.2021.736062

Shelomi, M., Wu, M. K., Chen, S. M., Huang, J. J., & Burke, C. G. (2020). Microbes associated with black soldier fly (Diptera: Stratiomiidae) degradation of food waste. *Environmental Entomology*, *49*(2), 405–411. https://doi.org/10.1093/EE/NVZ164

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843. https://doi.org/10.1038/s41564-018-0171-1

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), 1–11. https://doi.org/10.1002/cpmb.59

SourceForge. (2012). SAMtools. Retrieved January 6, 2023, from https://samtools.sourceforge.net/

Spranghers, T., Noyez, A., Schildermans, K., & De Clercq, P. (2017). Cold Hardiness of the Black Soldier Fly (Diptera: Stratiomyidae). *Journal of Economic Entomology*, *110*(4), 1501–1507. https://doi.org/10.1093/jee/tox142

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

https://doi.org/10.1093/bioinformatics/btu033

Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, *57*(5), 758–771. https://doi.org/10.1080/10635150802429642

Sumba, M. (2016). *Diseño de un sistema de crianza de la mosca soldado negra (Hermetia illucens) para la producción de harina como ingrediente proteico en la elaboración de balanceados*. Escuela Superior Politécnica Del Litoral. Retrieved from https://www.dspace.espol.edu.ec/handle/123456789/51701

Sun, J., Prabhu, A., Aroney, S. T. N., & Rinke, C. (2022). Insights into plastic biodegradation: Community composition and functional capabilities of the superworm (Zophobas morio) microbiome in styrofoam feeding trials. *Microbial Genomics*, *8*(6), 1–19. https://doi.org/10.1099/mgen.0.000842

Volkmann, M., Skiebe, E., Kerrinnes, T., Faber, F., Lepka, D., Pfeifer, Y., … Wilharm, G. (2010). Orbus hercynius gen. nov., sp. nov., isolated from faeces of wild boar, is most closely related to members of the orders "Enterobacteriales" and Pasteurellales. *International Journal of Systematic and Evolutionary Microbiology*, *60*(11), 2601–2605. https://doi.org/10.1099/ijs.0.019026-0

Wang, Y. S., & Shelomi, M. (2017). Review of black soldier fly (Hermetia illucens) as animal feed and human food. *Foods*, *6*(10). https://doi.org/10.3390/foods6100091

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., … Sobral, B. W. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*, *42*(D1), 1–11. https://doi.org/10.1093/nar/gkt1099

Wilharm, G. (2019). Orbus. *Bergey's Manual of Systematics of Archaea and Bacteria*. https://doi.org/10.1002/9781118960608.gbm01703

Wu, Y. W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638

Yamanaka, T. (2008). *Chemolithoautotrophic Bacteria: Biochemistry and Environmental Biology*. (S. S. & B. Media, Ed.). Tokio: Springer. Retrieved from https://books.google.com.ec/books?id=dWvBGm5FUzQC&dq=chemoheterotrophic&hl=es&source=gbs_navlinks_s

Yang, J., Yang, Y., Wu, W. M., Zhao, J., & Jiang, L. (2014). Evidence of polyethylene biodegradation by bacterial strains from the guts of plastic-eating waxworms. *Environmental Science and Technology*, *48*(23), 13776–13784. https://doi.org/10.1021/es504038a

Yang, Y., Chen, J., Wu, W. M., Zhao, J., & Yang, J. (2015). Complete genome sequence of Bacillus sp. YP1, a polyethylene-degrading bacterium from waxworm's gut. *Journal of Biotechnology*, *200*, 77–78. https://doi.org/10.1016/j.jbiotec.2015.02.034

Zhan, S., Fang, G., Cai, M., Kou, Z., Xu, J., Cao, Y., … Huang, Y. (2020). Genomic landscape and genetic manipulation of the black soldier fly Hermetia illucens, a natural waste recycler. *Cell Research*, *30*(1), 50–60. https://doi.org/10.1038/s41422-019-0252-6

Zheng, H., Nishida, A., Kwong, W. K., Koch, H., Engel, P., Steele, M. I., & Moran, N. A. (2016). Metabolism of toxic sugars by strains of the bee gut symbiont Gilliamella apicola. *MBio*, *7*(6). https://doi.org/10.1128/mBio.01326-16

Zhineng, Y., Ying, M., Bingjie, T., Rouxian, Z., & Qiang, Z. (2021). Intestinal microbiota and functional characteristics of black soldier fly larvae (Hermetia illucens). *Annals of Microbiology*, *71*(1). https://doi.org/10.1186/s13213-021-01626-8

7. ANEXOS

ANEXO I. CÓDIGO EJECUTADO EN EL BASH DE LINUX, A TRAVÉS DEL CLÚSTER DE CEDIA, PARA LA ELIMINACIÓN DE LAS LECTURAS DEL HOSPEDERO

Ingresar al clúster de CEDIA y solicitar el uso del nodo de cómputo:

salloc -n 1 --mem=128G -c 12 --gpus=1 #usar este

Entrar al nodo:

ssh dgx-node-0-0

Análisis de calidad

Instalar FastQC

wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.9.zip

unzip fastqc_v0.11.9.zip # Descomprimir el archivo

Correr FastQC en cada archivo que contiene las secuencias crudas:

export PATH=$PATH:/home/yadira.salguero/programs/FastQC/ # Establecer ruta del FastQC

fastqc /home/yadira.salguero/tesis/datoscrudos/PS/PS_forward/forward1.txt # Control de calidad secuencias crudas *forward*

fastqc /home/yadira.salguero/tesis/datoscrudos/PS/PS_reverse/reverse1.txt # Control de calidad secuencias crudas *reverse*

Limpieza de secuencias con Trimmomatic

Instalación de trimommatic:

wget http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.39.zip # Descargar trimmomatic

unzip Trimmomatic-0.39.zip

Ejecución sobre los archivos *forward* y *reverse*

java -Xmx1G -jar /home/yadira.salguero/programs/Trimmomatic-0.39/trimmomatic-0.39.jar PE -phred33 -threads 8 -trimlog logfile /home/yadira.salguero/tesis/datoscrudos/PS/PS_forward/forward1.txt /home/yadira.salguero/tesis/datoscrudos/PS/PS_reverse/reverse1.txt /home/yadira.salguero/tesis/datoscrudos/PS_leftP.fastq.gz /home/yadira.salguero/tesis/datoscrudos/PS_leftU.fastq.gz /home/yadira.salguero/tesis/datoscrudos/PS_rightP.fastq.gz /home/yadira.salguero/tesis/datoscrudos/PS_rightU.fastq.gz ILLUMINACLIP:/home/yadira.salguero/programs/Trimmomatic-0.39/adapters/TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:5:20 LEADING:5 TRAILING:5 MINLEN:50

Control de calidad sobre archivos limpios (FastQC)

export PATH=$PATH:/home/yadira.salguero/programs/FastQC/ # Establecer ruta del FastQC

fastqc /home/yadira.salguero/tesis/datoscrudos/PS_leftP.fastq.gz # Control de calidad

fastqc /home/yadira.salguero/tesis/datoscrudos/PS_rightP.fastq.gz

Eliminación de secuencias duplicadas con BBMap

Instalar BBMap

wget https://sourceforge.net/projects/bbmap/files/BBMap_38.34.tar.gz

tar xvzf BBMap_38.34.tar.gz

Establecer ruta del BBMap

export PATH=$PATH:/home/yadira.salguero/programs/bbmap/

Eliminar secuencias duplicadas utilizando BBMap

dedupe.sh in1=/home/yadira.salguero/tesis/datoscrudos/PS_rightP.fq in2=/home/yadira.salguero/tesis/datoscrudos/PS_leftP.fq out=/home/yadira.salguero/tesis/datoscrudos/PSded.fq outd=/home/yadira.salguero/tesis/datoscrudos/PSdup.fq ac=f

** elimina duplicados, pero genera archivo intercalado

reformat.sh in=/home/yadira.salguero/tesis/datoscrudos/PSded.fq
out1=/home/yadira.salguero/tesis/datoscrudos/PSded1.fq
out2=/home/yadira.salguero/tesis/datoscrudos/PSded2.fq **de-interleave file

Eliminación de lecturas del hospedero

Instalar Bowtie2

cd programs

wget https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.4.2/bowtie2-2.4.2-sra-linux-x86_64.zip/download

unzip download

Descargar el genoma del hospedero

wget
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/905/115/235/GCF_905115235.1_iHerIll2.2.curated.20191125/GCF_905115235.1_iHerIll2.2.curated.20191125_genomic.fna.gz

gunzip GCF_905115235.1_iHerIll2.2.curated.20191125_genomic.fna.gz

descomprimir y renombrar el archivo para tener un nombre más corto

Crear ruta a Bowtie2

export PATH=$PATH:/home/yadira.salguero/programs/bowtie2-2.4.2-sra-linux-x86_64:$PATH

Ir a la carpeta deseada (comando cd) y crear una base de datos con el genoma del hospedero

bowtie2-build /home/yadira.salguero/tesis/datoscrudos/GCF_Hermetia.fna Hermetia_DB

Mapear secuencias a la base de datos del hospedero

bowtie2 -x /home/yadira.salguero/tesis/datoscrudos/Hermetia_DB -1
/home/yadira.salguero/tesis/datoscrudos/PSded1.fq -2
/home/yadira.salguero/tesis/datoscrudos/PSded2.fq -S
/home/yadira.salguero/tesis/datoscrudos/mapped_and_unmapped.sam

Solicitud de accesos root en el clúster e instalación de sammtols

enroot import docker://nvcr.io#nvidia/cuda

enroot create --name micudapersonal nvidia+cuda.sqsh

enroot list

enroot start --mount $HOME --root --rw micudapersonal sh -c '/bin/bash' # solo esto en futuras conexiones al clúster

apt update

apt install samtools

cd home/yadira.salguero/

Selección de secuencias no mapeadas

Convertir file .sam a .bam

samtools view -b -f 12 -F 256 /home/yadira.salguero/tesis/datoscrudos/mapped_and_unmapped.sam /home/yadira.salguero/tesis/datoscrudos/bothEndsUnmapped.bam

-f 12 Extraer solo (-f) las secuencias con lecturas pareadas no mapeadas: <read unmapped><mate unmapped>

-F 256 No (-F) extraer alineamientos secundarios (secuancias que han sido mapeadas a más de una referencia): <not primary alignment>

Organizar lecturas

samtools sort -n /home/yadira.salguero/tesis/datoscrudos/bothEndsUnmapped.bam -o /home/yadira.salguero/tesis/datoscrudos/bothEndsUnmapped_sorted.bam

Exit # para salir del root

División de lecturas en archivos paired end con bedtools

Instalar bedtools

wget https://github.com/arq5x/bedtools2/releases/download/v2.25.0/bedtools-2.25.0.tar.gz

tar -zxvf bedtools-2.25.0.tar.gz

cd bedtools2

```
make

# Crear ruta a bedtools

export PATH=$PATH:/home/yadira.salguero/programs/bedtools2/bin

bedtools bamtofastq -i
/home/yadira.salguero/tesis/datoscrudos/bothEndsUnmapped_sorted.bam -fq
/home/yadira.salguero/tesis/datoscrudos/r1.fastq -fq2
/home/yadira.salguero/tesis/datoscrudos/r2.fastq
```

ANEXO II. RESUMEN DE LOS INFORMES DE CALIDAD

Figura AII.1. Resumen de los informes de calidad de las secuencias forward (izquierda) y reverse (derecha) antes de la limpieza

Figura AII.2. Resumen de los informes de calidad de las secuencias forward (izquierda) y reverse (derecha) después de la limpieza

ANEXO III. RUTAS METABÓLICAS

Tabla AIII.1. Número de pasos del módulo de las rutas metabólicas encontradas

Vía metabólica	**Pasos del módulo**	**Pasos presentes**
Embden-Meyerhoff (glucólisis)	9	9
Ciclo de la pentosa fosfato	7	7
Entner-Doudoroff	4	4
Ciclo de citrato o de Krebs (TCA)	8	5
Ciclo de calvin	11	8
Arnon-Buchanan (ciclo del citrato reductor)	10	6
Ciclo dicarboxilato-hidroxibutirato	13	3
Wood–Ljungdahl	7	2

Tabla AIII.2. Número de subunidades del módulo y subunidades presentes, así como del Complejo de la CTE

Complejo CTE	**Subunidades del módulo**	**Subunidades presentes**	**Genes presentes**	**Genes faltantes**
NADH:quinona oxidoreductasa	11	2	K00340, K00343	K00330, K00334, K00335, K00336, K00337, K00338, K00339, K13380, K15863
Fumarato reductasa	4	4	K00244, K00245, K00246, K00247	-
Citocromo bd ubiquinol oxidasa	3	2	K00425, K00426	K00424
ATPasa tipo-F	8	8	K02108, K02109, K02110, K02111, K02112, K02113, K02114, K02115	-

En la Figura AIII.1 se presenta una vista general de una ruta de referencia del metabolismo central del carbono, que indica en un círculo la cantidad de carbonos para cada compuesto, excluyendo un cofactor (CoA, CoM, THF o THMPT) que se reemplaza por un asterisco. En el mapa se resaltan las vías de utilización de carbono de la glucólisis (mapa00010), la vía de las pentosas fosfato (mapa00030) y el ciclo del citrato (mapa00020), y cuatro vías de fijación

de carbono: (1) ciclo reductor de pentosa fosfato (ciclo de Calvin) en plantas y cianobacterias que realizan la fotosíntesis oxigénica, (2) ciclo reductor de citrato en bacterias fotosintéticas de azufre verde y algunos quimiolitoautótrofos, (3) ciclo de dicarboxilato-hidroxibutirato, y (4) vía reductora de acetil-CoA en bacterias metanogénicas (KEGG, 2021).

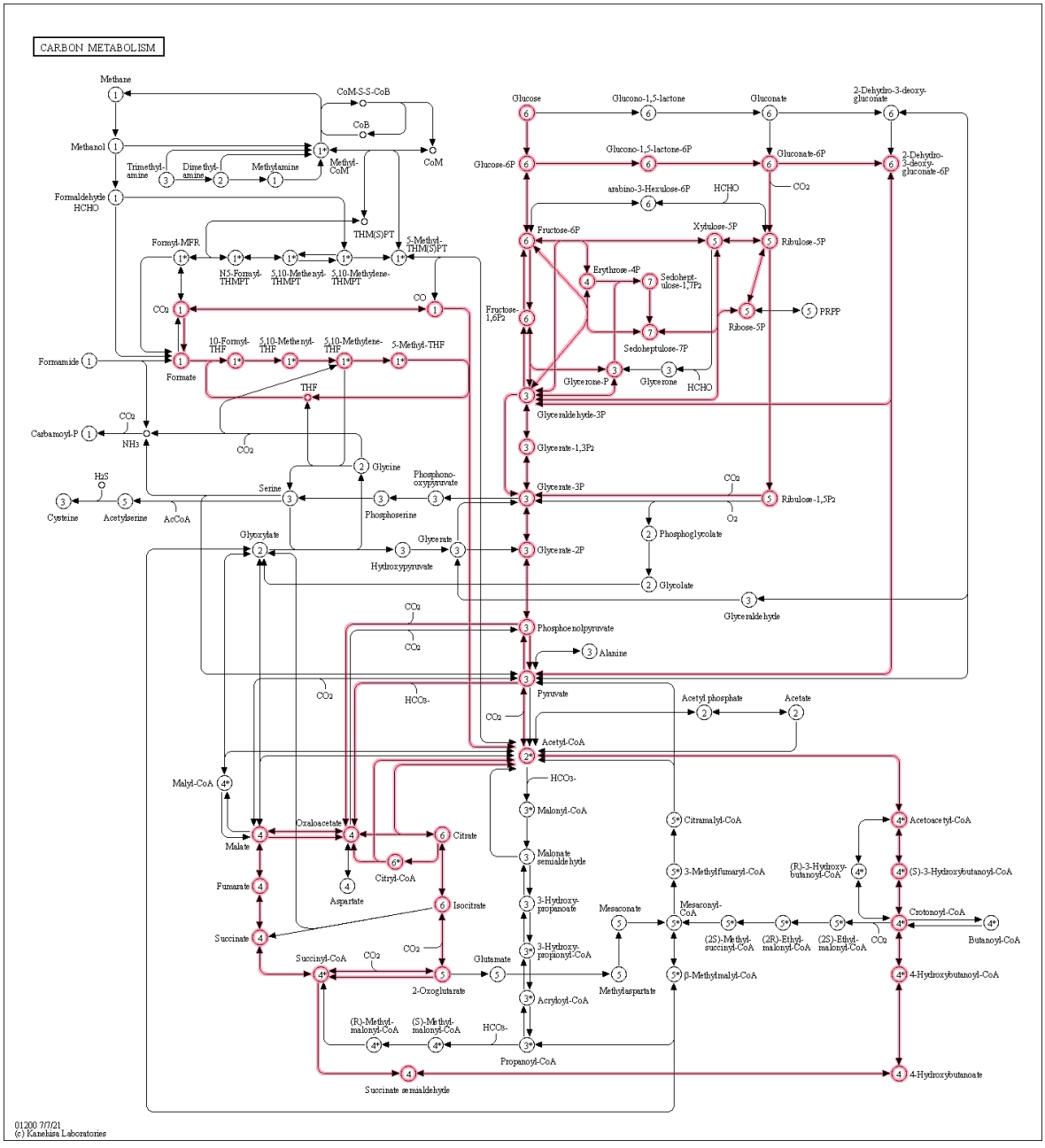


Figura AIII.1. Rutas metabólicas de la gammaproteobacteria del género *Orbus* (KEGG, 2021)

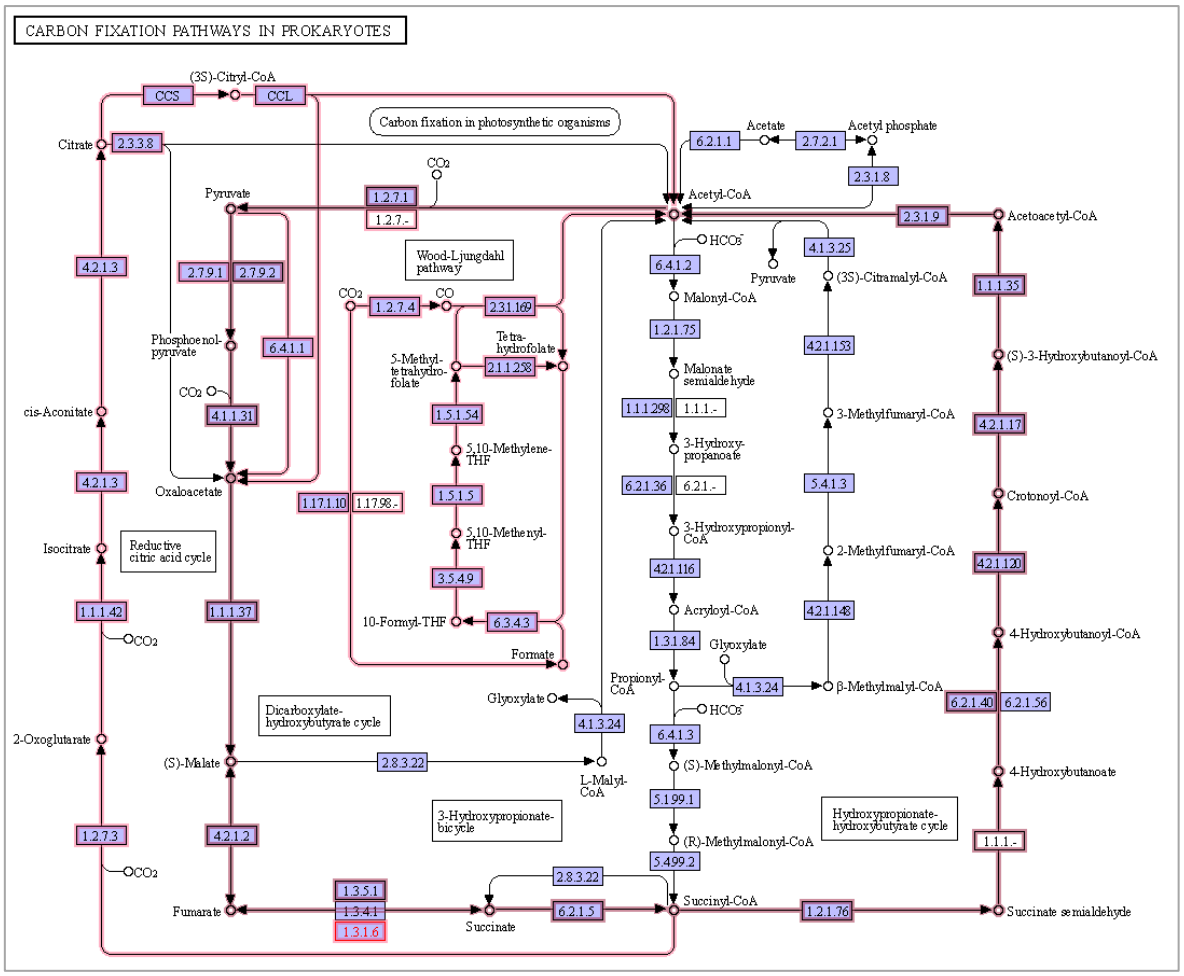


Figura AIII.2. Rutas metabólicas para la fijación de carbono en procariotas (KEGG, 2022)