UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

Inferencia y comparación de perfiles metabólicos entre muestras perturbadas y no perturbadas de metagenomas de tipo shotgun

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Maestro en Ciencias

PRESENTA:
Biól. María del Carmen Sánchez Olmos

Dra. Rosa María Gutiérrez Ríos
Instituto de Biotecnología

MIEMBROS DEL COMITÉ TUTOR
Dra. Blanca Itzetl Taboada
Instituto de Biotecnología
Dra. Eria Rebollar Caudillo
Centro de Ciencias Genómicas

Ciudad de México. Febrero, 2022

El presente proyecto se realizó bajo la asesoría de la Dra. Rosa María Gutiérrez Ríos, en el Laboratorio de Genómica Computacional adscrito al Departamento de Microbiología Molecular del Instituto de Biotecnología de la Universidad Nacional Autónoma de México campus Morelos y con el apoyo económico brindado por el Consejo Nacional de Ciencia y Tecnología(CONACyT) correspondiente al CVU: 1034022 bajo el programa de Maestría y Doctorado en Ciencias Bioquímicas. Agradezco al proyecto PAPIIT-DGAPA IN202821.

Índice

Índice de figuras

1. Agradecimientos

A la Dra. Rosa María Gutiérrez Ríos por el apoyo brindado durante este proceso. Por la confianza, dedicación y determinación.

A el Dr. Antonio Loza por guiarme en el área de la estadística para lograr realizar este proyecto.

Al los miembros del grupo de Genómica Computacional del Instituto de Biotecnología del Depto. de Microbiología Molecular por el apoyo brindado.

A las Dras. Blanca Taboada y Eria Rebollar miembros de mi comité tutoral por sus consejos

A los miembros del jurado que ayudaron a mejorar este manuscrito

A Ricardo Farrera por el cariño y apoyo invaluable que permitió terminar este proyecto.

2. Abstract

The use of metagenomics to understand the great variety of natural ecosystems across the globe has produced a large quantity of data stored in public databases, which gives an excellent opportunity to compare several niches. The analysis of the sequences derived from shotgun metagenomes can shed light on the composition of microbial communities and their functions.

Calculating the relative abundance of genes in a metagenomic sample is based on tools derived from RNA-seq analysis, which may not be appropriate considering the high dimensionality and composition of the data. These methods essentially manipulate the way genes are quantified and its normalization, so new approaches have to be developed. Once assigned, gene abundances are related to its functions, and those codifying for enzymes served to define the metabolic potential of metagenomic samples. The metabolic pathways can be recognized, mainly by the presence of marker genes that the researcher has to supposed to be present, taking into account the environmental variables, which are not fully measured. For these reasons, we implemented a statistical method that with minimal environmental information can find overrepresented enzymes defining metabolic pathways describing the metabolic potential of shotgun metagenome samples. In this work, the method has the main goal of describing the metabolic profile of a contaminated sample, distributed in the Gulf of Mexico, when contrated with reference metagenomes distributed worldwide.

For this purpose, the implemented method delineates each enzyme's under, over, or equal representation. To do this, we used the abundance pattern of every enzyme annotated in a group of samples with similar environmental characteristics (superficial sediments from all around the world), which we called the reference. Over each enzyme found on the reference, a probability density was calculated that was used to compare the abundance of the same enzyme found in the sediments collected in the Gulf of Mexico that helps to estimate thier level of representation in the metagenome.

This novel technique shed light on the metabolic pathways that are taking place in

the three çontaminated"samples analyzed in this project and prove the value of creating methods that describe the different representation of enzymes in terms of z-score as a tool to have access to the most relevant pathways found in the samples.

3. Resumen

A partir de ADN ambiental, la matagenómica ha permitido conocer la diversidad microbiana de una gran variedad de ecosistemas, incluyendo así a aquellos organismos difíciles de cultivar en laboratorio.

La metagenómica no sólo permite conocer la diversidad taxonómica de los ambientes en estudio sino también el potencial funcional de las mismas. Durante los análisis metagenómicos es rutinario comparar los perfiles metabólicos de dos o más muestras, con la finalidad de reconstruir y describir el potencial funcional de las comunidades microbianas de los ambientes en estudio.

En general, cuando se hacen comparaciones a nivel funcional de muestras metagenómicas los métodos estadísticos para evaluar diferencias entre muestras derivan de métodos o técnicas utilizadas para analizar datos de tipo transcriptómicos. Sin embargo, estos métodos no son siempre apropiados para el análisis de datos metagenómicos debido a la naturaleza de los mismos.

De esta forma, el método implementado en este proyecto no se basa en metodologías extrapoladas de otras áreas como la transcriptómica. Por el contrario, es una técnica nueva que por un lado hace uso de toda la información funcional para realizar la reconstrucción de rutas metabólicas y no solamente genes marcadores de ciertas vías metabólicas. Y por otro lado, pero no menos importante el método brinda una medida que refleja la representatividad estadística de las enzimas encontradas en metagenomas perturbados, en este caso del Golfo de México contaminados por derrames de petróleo.

Para el desarrollo del método se construyó un grupo de datos de referencia usando las abundancias de las enzimas provenientes de metagenomas obtenidas de sedimentos

distribuidos alrededor del mundo sin evidencia de tener algún tipo de contaminación. Para cada enzima de la referencia se define un función de densidad de probabilidad que sirve como base de comparación para ubicar las abundancias de las enzimas del Golfo y calificar su representatividad.

a través de la asignación de una calificación Z.

Esta nueva herramienta permitió conocer las vías metabólicas más relevantes para las tres muestras analizadas y permitió probar el valor del método en la descripción del potencial metabólico.

4. Introducción

4.1. La metagenómica actual

La visión actual de la diversidad y dinámica funcional de comunidades microbianas en distintos ambientes del planeta se ha enriquecido sobremanera debido al desarrollo de nuevas tecnologías de secuenciación (*e.g Nanopore*, *PacBio*), así como la creciente implementación de técnicas de *single cell* y de los enfoques para acceder al fenotipo de comunidades sin alteración física [22].

En conjunto, tanto el mejoramiento de tecnologías y el uso rutinario de la metagenómica ha brindado la capacidad de comprender la respuesta de las comunidades microbianas ante alteraciones antropogénicas, cambios en la productividad del ecosistema, así como, acelerar el descubrimiento de nuevas funciones de genes que contribuyan al incremento en el tamaño del acervo funcional de los ambientes del planeta.

Entre otras de las aportaciones de la metagenómica a la comprensión de la diversidad microbiana, se encuentra el descubrimiento de nuevos taxa procarióticos (*e.g* Brokarchaeota) [4], que de otra forma, ya sea con el uso de técnicas de cultivo en laboratorio o mapeo de rRNA 16S sería imposible vislumbrar.

Adicionalmente, la metagenómica muestra la variabilidad genética existente en una población natural conformada por eucariotes unicelulares, bacterias, archaeas y virus, reflejando la influencia de factores evolutivos, ecológicos y ambientales en dichas comunidades [3].

Finalmente, gracias al acceso proporcionado por esta herramienta a ecosistemas terrestres, marinos y de agua dulce, e incluso a ambientes complejos como son las ventilas hidrotermales o suelos de cultivo, es posible contribuir a una mejor interpretación y comprensión de los ciclos metabólicos y/o biogeoquímicos predominantes y su influencia en la dinámica planetaria, marcando un punto de partida para el manejo, monitoreo y remediación de ecosistemas [14].

Entre los alcances de la metagenómica encontramos su uso para acceder a la gran diversidad microbiana de un gran número de hospederos desde el ser humano, animales, plantas hasta protozoarios. Se ha observado que la microbiota desempeña un papel importante en la adecuación del organismo a su ambiente y tiene un rol igualmente relevante en la salud y enfermedad del mismo. Esto por un lado ha revelado una gran diversidad de bacterias y archaeas componentes importantes de muchos animales y plantas y por el otro ha abierto una pauta en la comprensión de las asociaciones ecológicas a nivel hospedero y su evolución[35]

En la Figura 1 se observa como el enfoque metagenómico ha impactado la manera convencional de hacer microbiología, y cuáles son las futuras direcciones en las que contribuirá dicha disciplina.

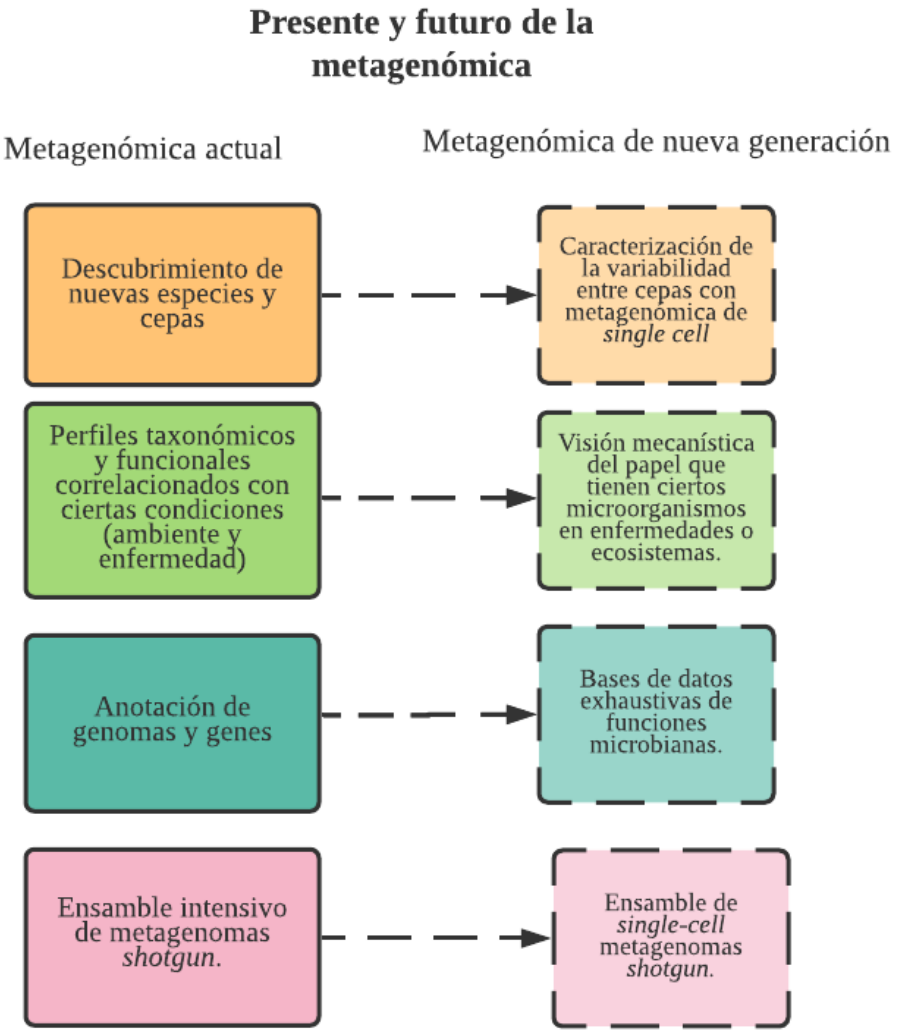

Figura 1: Impacto de la metagenómica en el campo de la microbiología y su dirección en el futuro. Obtenido y modificado de Laudadio I. (2019) [30].

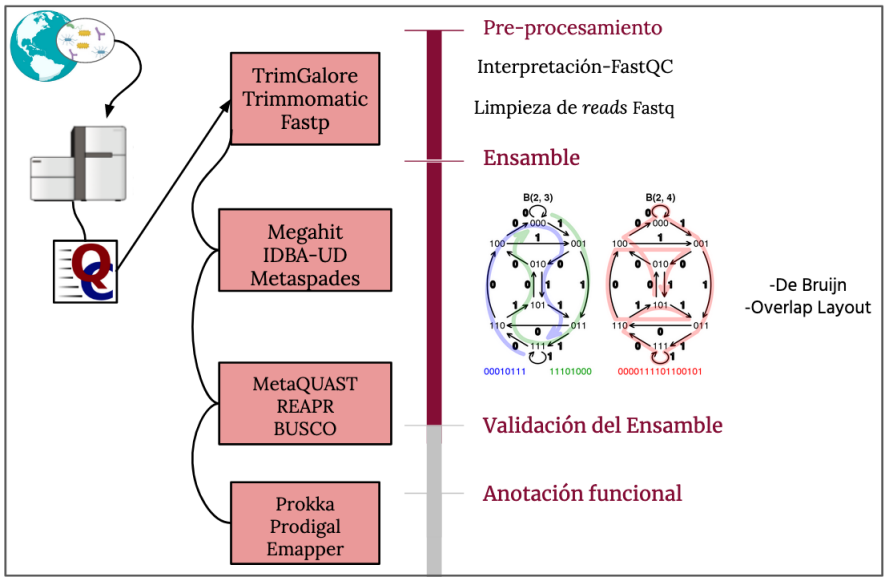

Figura 2: Representación del flujo convencional de análisis de muestras derivadas de metagenómica.

4.2. Flujo de análisis de datos metagenómicos

En la presente sección se describirán, cuáles son los pasos de análisis convencional de muestras de metagenomas *shotgun*, así como, los retos y dificultades a considerar en cada paso.

Durante un análisis metagenómico, los datos de las secuencias de DNA obtenidos directamente del ambiente se encuentran almacenados en formatos específicos como es el **fastq**.

El primer paso en el flujo de trabajo convencional que se puede observar en la Figura 2, consiste en la limpieza de los *reads* o secuencias, esto con el objetivo de eliminar errores propios de las plataformas de secuenciación, como son las secuencias de baja complejidad, bases con una calidad desfavorable según los *score* Phred, así como secuencias repetidas y/o adaptadores [12].

Una vez que se han limpiado las secuencias, el siguiente paso en el análisis de los datos, puede seguir dos enfoques en lo que concierne a la asignación de información biológica, que como se explicará en la sección de Métodos, es la que nos concierne para fines del presente proyecto.

El primero de estos consiste en el mapeo de los *reads* contra bases de datos con información a nivel funcional. Sin embargo, esta aproximación tiene la desventaja de generar sesgos debido a la falta de secuencias que permitan identificar correctamente la información encontrada en el metagenoma, ya que existe una gran cantidad de organismos no cultivados en el ambiente y por lo tanto, muchos genes sin una función asignada.

El segundo enfoque hace referencia a la generación de fragmentos de secuencias más largas a partir del ensamble de los *reads* o lecturas de DNA, conocidas como *contigs* y/o *scaffolds*, es decir, la reconstrucción de genomas.

El ensamble permite una reconstrucción más completa de las comunidades microbianas, esto como resultado de una predicción de marcos de lectura abiertos (ORF, "open reading frames", por sus siglas en inglés) con mayor precisión y certeza, eliminando el sesgo de enfoques dependientes de bases de datos que muchas ocasiones pasan por alto un gran número de genes [40]

El ensamble de metagenomas también brinda la capacidad de reconstruir genomas completos de los organismos encontrados en el ambiente (*"near-complete-genomes"*), proporcionando información a un nivel más fino de las conformaciones de genes *e.g* operones, elementos extracromosomales y de sus funciones dentro de dicho contexto genómico, permitiendo la integración de las vías metabólicas que imperan en la comunidad microbiana en cuestión [3].

En este sentido, la sensibilidad de los métodos tanto de extracción de DNA, aquellos asociados a las plataformas de secuenciación, así como del post-procesamiento de datos metagenómicos, pueden generar falsos positivos, es decir, recuperación de DNA de organismos que realmente no se encuentran en la comunidad ambiental (DNA de organismos muertos) o bien, falsos negativos, en donde el DNA de ciertos organismos resulta imposi-

ble de recolectar. Sesgos de este tipo deben tomarse en consideración al momento de la interpretación biológica [47].

4.2.1. Ensamble de metagenomas

En una muestra metagenómica encontramos ADN proveniente de una gran variedad de organismos, al ensamble de estos genomas se le conoce como ensamble de genomas a partir metagenomas.

El ensamble de genomas a partir de metagenomas es un proceso costoso y complejo computacionalmente. Se ha identificado que la complejidad en resolver ensambles obedece a la proporción entre el tamaño de las lecturas de DNA, los posibles errores de secuenciación y el tamaño de secuencias repetidas, que en especial para los metagenomas incluyen secuencias intergénicas e intragénicas.

En consecuencia, el escoger de entre distintas plataformas de secuenciación, presenta una serie de ventajas y desventajas importantes ya que cada una tiene distintos resultados en lo que se refiere al tamaño de las lecturas de DNA, profundidad de la secuenciación, rendimiento, tasa de error entre otros como se muestra en la Tabla 1.

Plataformas de secuenciación	Reads(M)	Longitud de *reads*	Error	Precisión
Sanger	96	800	<0.1	99.9%
Roche 454	96	1,25	700	98%
Illumina MiSeq	25	300	0.1	99.9%
Illumina NextSeq 500	400	150	0.1	99.9%
Illumina HiSeq 4000	5000	150	0.1	99.9%
Illuimina HiSeq X	6000	150	0.1	99.9%
PacBio	38,9	1000-40000	13-15	99.9%
Oxford Nanopore	Variable	5000-15000	3-5	70-90%

Tabla 1. Características de distintas plataformas de secuenciación. Datos obtenidos de Dick (2019) [12], Bleidorn (2017). [7] y Ghurye J.S (2016) [15]

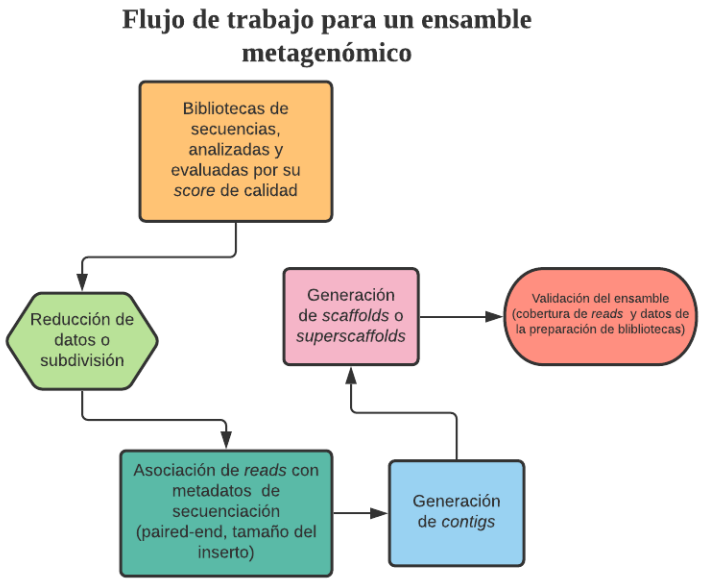

Figura 3: Flujo de trabajo convencional en la generación de ensambles metagenómicos. Obtenido de Howe A. (2017) [9]

Es importante mencionar que también existen otros factores que contribuyen a la complejidad para resolver el ensamble, como la presencia de variantes entre especies [40].

En el diagrama de la Figura 3, se muestra el flujo de trabajo común en la generación de ensambles metagenómicos:

Para la generación de ensambles metagenómicos existen dos enfoques que se basan en la construcción de gráficas al conectar la información de las lecturas y resolver el camino más parsimonioso.

El primero de los algoritmos se conoce como OLC (*Overlap Layout Consensus*, por sus siglas en inglés), el cual consiste en la identificación de sobrelapamientos entre las lecturas a través de alineamientos pareados. Estos forman las conexiones en la gráfica

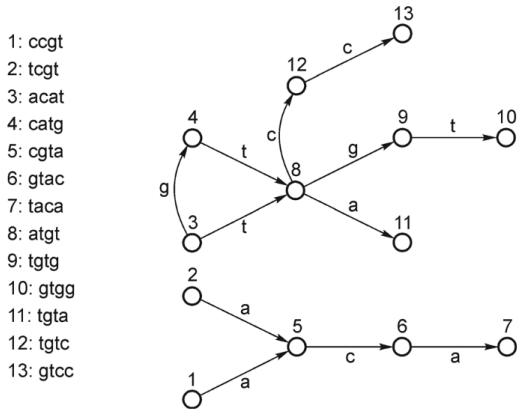

Figura 4: Ejemplo de una gráfica de de Bruijn, con la secuencia **ccgtac** y **catgtg**, los nodos son los *k-meros* de tamaño 4 a la izquierda de la imagen. Obtenido de Rizzi R. (2019) [49]

resultante, dando lugar a la secuencia continua más probable. Finalmente la secuencia consenso para cada *contig* se determina al elegir el nucleótido más representado en cada sobrelape de secuencias [59].

La construcción de gráficas de Bruijn representa el segundo paradigma, el cual fue introducido por Idbury y Waterman [26] y se caracteriza por ser eficiente y con alta precisión. Para la construcción de gráficas de de Bruijn, las lecturas de DNA se fragmentan en secuencias de tamaño *k* conocidas como *k-meros* los cuales representan los nodos de la gráfica, las conexiones entre nodos resultan del sobrelapamiento sin *missmatches* de k-1 bases entre los *k-meros*, como se muestra en la Figura 4.

De tal forma, que las conexiones encontradas entre los nodos de la red representan un posible camino. En donde, el objetivo es encontrar aquel que explique de manera consenso o resuelva mejor las conexiones entre los *k-meros*.

Debido a la existencia de múltiples caminos válidos que resuelven el ensamble ambos algoritmos generan reconstrucciones fragmentadas de los genomas. Así, mientras mayor sea la longitud del *k-mero* se obtendrá un menor número de nodos repetidos en la gráfica

y más fácil resultará la resolución de forma continua de los segmentos del genoma(s), amortiguando el impacto que tienen las secuencias repetidas.

Independientemente del método utilizado para resolver el ensamble de metagenomas y la posible complementación con técnicas como Hi-C, *single-cell*, *long-reads* y/o procesamiento de *scaffolds*, el ensamble resultante siempre tendrá cierto porcentaje de error, lo que hace necesario la validación del mismo. [40].

4.2.2. Validación de ensambles

En general, los métodos de validación de ensambles hacen uso de genomas de referencia o bien, calculan métricas de evaluación de calidad a partir de las características intrínsecas del ensamble (*de novo*), las métricas utilizadas se clasifican en cuatro categorías como se muestra en la Figura 5.

Aquellas métricas que hacen uso de un genoma de referencia intentan acceder a la *integridad* del ensamble, obteniendo la proporción de *contigs* totales alineados al genoma de referencia.

Por otra parte, para saber que tan completo resultó el ensamble existen métricas que reflejan la *continuidad* del mismo al medir la longitud de los *contigs* bajo la premisa de un sólo *contig* por cromosoma [60].

Además, con la intención de medir la *consistencia* del ensamble, se mapean los *reads* al mismo, para detectar *contigs* quimera.

Finalmente, la *precisión* del ensamble se refiere al número de *contigs* que corresponden al tamaño real del genoma cuando se tienen referencias completas.

Programas como MetaQuast hacen uso de genomas cerrados depositados en bases de datos para evaluar la integridad y precisión del ensamble, en cambio REAPR busca medir las inconsistencias propias del ensamble (ver figura 5).

Otra de las métricas fundamentales en la validación de ensambles metagenómicos,

Métricas por categoría para la validación de ensambles.

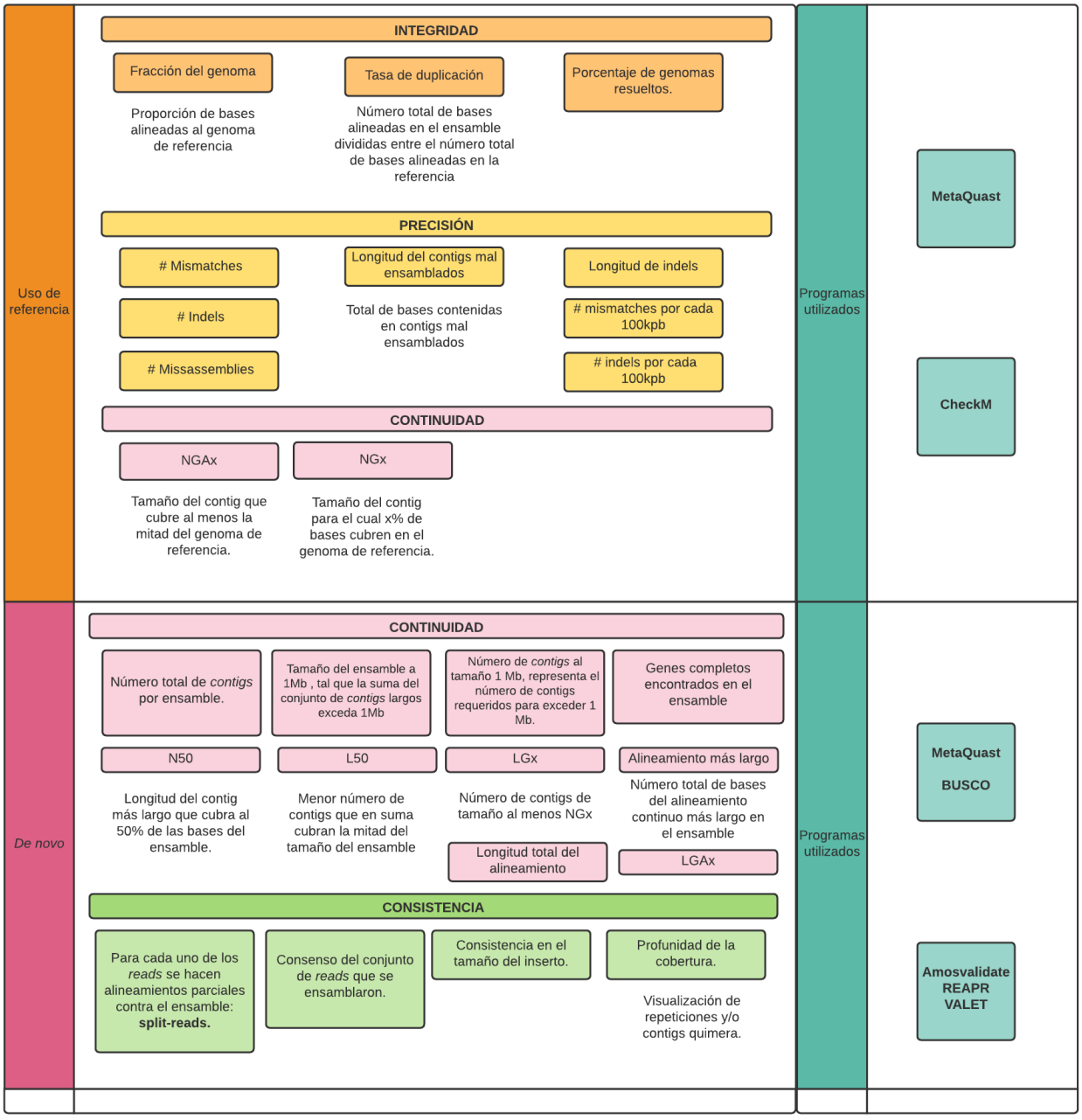

Figura 5: Métricas utilizadas para validar ensambles a partir de metagenomas. Obtenido y modificado de Olson (2019) [40] y Wang Z. (2020) [60]

es el número de marcos de lectura abiertos o ORF's, además de la densidad de genes (ORF/Mb). Una de las ventajas de usar métricas relacionadas con el número de ORF's es su capacidad en la detección de errores, al encontrar ORF interrumpidos a diferencia de métricas relacionadas con el tamaño de contigs, en las cuales este tipo de errores en ocasiones no se observan [40].

Por otro lado y de igual importancia durante la validación de ensambles, son los tipos de errores de los ensambles, los cuales pueden caer dentro de las siguientes cuatro categorías: 1. colapso de repeticiones, 2. inserciones, 3. deleciones y 4. inversiones.

Para poder acceder y evaluar dichos errores se necesita de un alineamiento o mapeo de las lecturas de DNA contra el ensamble resultante, a lo que se le conoce como *split-reads* (Figura 5, sección Çonsistencia del ensamble"). Esto permite evaluar la profundidad de secuenciación y su influencia en la construcción del ensamble. Además, permite conocer el grado de fragmentación del ensamble al observar la distancia y correcta orientación entre las lecturas [57].

4.3. Análisis estadístico de datos metagenómicos a nivel funcional

Las tecnologías de secuenciación de nueva generación (Next-Generation sequencing), han reducido el costo para la generación de datos genómicos, que resulta en un incremento en el almacenamiento y disponibilidad de dichos datos.

En lo referente a datos de metagenomas tipo shotgun este incremento de datos ha ido de la mano con el desarrollo de metodologías que permiten explicar los patrones observados a nivel taxonómico, funcional y ecológico en las muestras de ambientes estudiadas.

Existen distintos tipos de acercamientos para el análisis de datos *shotgun*, sin embargo, antes de explicar a detalle cada uno de estos, es importante resaltar las características propias de este tipo de datos.

En primer lugar, los datos *shotgun* muestran ciertas características que obedecen a la propia tecnología de secuenciación:

- La cobertura de secuenciación no es uniforme para las muestras metagenómicas, lo que resulta en una representación sesgada, tanto de genes como de genomas.
- La falta de representantes filogenéticos en las bases de datos, debido a la representación desigual de taxa en las muestras y en la capacidad de detección de los mismos por los métodos de secuenciación.
- Tanto los procesos experimentales como de control de calidad generan una alta variabilidad en los conteos finales de las muestras [8]

Si bien las anteriores características son resultado de la capacidad de las tecnologías de secuenciación, las siguientes propiedades describen la naturaleza composicional propia de datos metagenómicos.

- La **alta-dimensionalidad**, es decir, el gran número de categorías en las que se puede describir el conjunto de datos.
- La **escasez** de datos genera un gran número de ceros en las observaciones de la muestra analizada, como respuesta a observaciones infrecuentes (*e.g* genes o taxa)
- La **no determinación**, que hace referencia a el significativo número de variables que sobrepasa el número de muestras. [56].

Otra de las características de los datos metagenómicos es la incapacidad de obtener abundancias absolutas tanto de genes como de taxones, debido a que los cambios en un sólo componente de la muestra afecta directamente la abundancia de los demás, así el número total de conteos u observaciones está limitado al número de secuencias que el secuenciador proporciona y esto tiene una alta dependencia con los taxa más abundantes en la comunidad (por ende, el límite es la profundidad de secuenciación) [8].

Como consecuencia, las abundancias relativas de las observaciones en las muestras representan datos informativos que permiten realizar inferencias estadísticas y evaluaciones biológicas.

Como se explica anteriormente y de acuerdo a las propiedades listadas, los datos composicionales requieren de un tratamiento apropiado para evitar falsas correlaciones de los datos al momento de realizar análisis inferenciales. Cuando se habla de análisis de datos composicionales como lo son los datos metagenómicos, el flujo de trabajo a seguir consiste en:

- **Normalización de datos**. La más simple y ampliamente utilizada es la abundancia relativa, que divide las abundancias observadas entre el número total de observaciones. Otra manera de normalizar es disminuir la representación de muestras con la finalidad de obtener el mismo número de conteos finales entre muestras. Sin embargo, esto crea una gran desventaja debido a la pérdida de información, por lo cual es una práctica no recomendada [36].
- **Análisis exploratorio**, a través de la visualización de los datos con la finalidad de observar patrones en la estructura de los datos.
- **Análisis inferencial**. Muchos de estas pruebas buscan hacer asociaciones de los datos con variables determinadas. Pueden ser de naturaleza univariada o multivariada. Los primeros para determinar asociaciones con una variable de respuesta, los últimos con la finalidad de observar diferencias globales entre las muestras (*e.g* PERMANOVA).

Es importante mencionar que Aitchison [1] desarrolló un enfoque que hace prescindible la normalización de los datos y los hace más manejables en términos matemáticos, al aplicar logaritmo a las proporciones o frecuencias relativas de las observaciones. [8].

Ahora bien, tanto el enfoque de Aitchison como las normalizaciones de las abundancias, cualquiera que se elija enfrenta el problema de escasez de datos que resulta por un lado, en la alta dispersión de datos y en la alta abundancia de ceros. Como consecuencia, se recomienda realizar una manipulación a los datos por medio de dos alternativas: el uso de un pseudo-conteo o la adición de una constante que haga los datos manejables y menos dispersos entre si.

Una de las ventajas de realizar análisis de abundancias diferenciales es su capacidad para comparar distintas muestras a nivel funcional (identificación de vías mayor o menormente representadas) (ver Tabla 2). Por otro lado, existen acercamientos que determinan la presencia o ausencia de genes con funciones de interés o genes marcadores, identificando también la abundancia de los mismos para compararlos posteriormente entre muestras [54]. Sin embargo, uno de los problemas que resulta del análisis de la abundancia en la presencia o ausencia de un grupo de genes es la dificultad en distinguir la ausencia de determinado taxa en la muestra, a lo que se le denomia çeros esenciales o ceros estructurales"de aquellas ausencia causadas por la falta de muestreo, conocidos como çeros redondeados".

Cualquiera de los enfoques antes descritos, ya sea la búsqueda de la presencia o ausencia de genes marcadores o el análisis diferencial de la abundancia de genes, en un estudio metagenómico que tenga la intención de describir y evaluar las propiedades funcionales y/o taxónomicas entre comunidades requiere siempre de la construcción de una tabla o matriz de conteo que refleje el número de secuencias por muestra para un taxón o gen en específico.

En la Tabla 2 se ejemplifican una serie de softwares con distintos acercamientos para describir el potencial metabólico de una muestra metagenómica.

MEBS, por ejemplo, hace uso de las abundancias relativas a un grupo curado de marcadores [10] para describir el potencial metabólico. LEfSe, por otro lado, normaliza las abundancias para después aplicar pruebas estadísticas que permitan diferenciar potenciales metabólicos entre muestras [53], e incluso existen métodos como MetaComp que permite comparar las propiedades de las comunidades a través del uso de más de una prueba estadística [64].

Métrica utilizada	**Uso de un grupo de genes marcadores como referencia**	**Análisis de vías metabólicas completas**	**Referencia**
Contenido informacional de las vías metabólicas. **Medida de Entropía Relativa:** $$H' = P(i)\log 2\frac{P(i)}{Q(i)} \quad (1)$$	Grupo de microorganismos y dominios Pfam curados.	Analizan de vías metabólicas y su integración (KEGG, MetaCyc)	MEBS: [10]
Test para medir la similitud entre muestras con el uso de **abundancias relativas** de especies.	Mapeo contra un grupo de genes de copia única (uso de genomas de referencias agrupados por especie)	No	MIDAS: [38]
Método que incluye en la comparación la alta-dimensionalidad de los datos. Hace uso de **abundancias relativas** normalizadas, tanto de funciones como taxa.	No.	Analiza la abundancia de COG's, y KEGG para integrarlo en los subsistemas SEED	LEfSe: [53]
Uso de **abundacias relativas**, para después realizar *t-test* entre muestras	No.	Analiza la abundancia de COG's, para integrarlo en los subsistemas SEED	Metastats:[62]

Hace uso de matrices de **abundancias relativas**. Aplicación de 5 posibles métodos de análisis estadístico.	No.	No.	MetaComp:[64]
Los conteos de las familias de dominios Pfam resultantes, son representados como **presencia o ausencia**.	No. Hacen uso de características fenotípicas	No.	Traitar [61]

Cuadro 1: Métricas basadas en la abundancia relativa o presencia y ausencia de genes, que permiten describir el perfil metabólico de muestras metagenómicas ya sea refiriendose a un grupo de genes marcadores o bien reconstruyendo vías metabólicas completas.

4.4. Metagenómica y la dinámica de ecosistemas: el caso específico de los sedimentos marinos.

Nuestro planeta se entiende como un sistema complejo modelado por múltiples interacciones entre la atmósfera, biosfera y el océano. El entendimiento de la dinámica biogeoquímica modelada por las comunidades microbianas presentes en cada uno de los ecosistemas del planeta, hacen fundamental el estudio y comprensión de su naturaleza a nivel bioquímico, ecológico y evolutivo [58].

En la actualidad no podemos dejar de lado la influencia que tienen las actividades humanas en la estructuración de las comunidades microbianas en los ecosistemas del mundo. Esto da pie a la formulación de preguntas como: ¿Qué tipo de patrones en las comunidades microbianas se originan ante perturbaciones?, ¿Es posible predecirlos y de qué manera influyen en la dinámica y estructura de las mismas?.

Ejemplificando y como respuesta a estas preguntas, De Anda y colaboradores,[10] [11] reportaron una serie de enfoques que permiten:

- Conocer la dinámica a nivel biogeoquímico de las comunidades encontradas en tapetes microbianos sometidos a perturbaciones antropogénicas, a través de acceder a la información funcional de las mismas.
- Evaluar las interacciones ecológicas entre organismos al implementar análisis de redes.
- Así como, la búsqueda de patrones que reflejen dichas interacciones entre comunidades microbianas.

De este modo se muestra que la metagenómica aunada a otros enfoques contribuye a la evaluación de la dinámica de un ecosistema y el efecto que tienen las actividades humanas en dichos ambientes.

Si bien es cierto, que cada uno de los ecosistemas sobre la tierra tienen influencia entre sí, y por lo tanto, no pueden contemplarse como entidades independientes, la complejidad que poseen los ambientes marinos, en específico los sedimentos profundos del mar los agrupa como ambientes que necesitan una análisis detallado de sus características a nivel fisicoquímico y biológico.

Los sedimentos profundos del mar se caracterizan por tener condiciones limitantes, como la falta de nutrientes. Por ejemplo, la energía disponible derivada del carbono orgánico fijado fotosintéticamente en la superficie es de apenas 1 %.

Aunado a esto, las altas temperaturas y presión ejercen condiciones difíciles para las comunidades que habitan estos ambientes.

Paradójicamente, más del 99.5 % del carbono total del planeta se encuentra en los sedimentos y rocas marinas, esto es 75×10^6 gigatoneladas de carbono, del cual 80 % es inorgánico, en suma a esto, la vida en los sedimentos marinos es característicamente abundante, con alrededor de 3×10^{29} células procariotas [33].

La estructura de un sedimento marino resulta de la constante deposición de materia orgánica proveniente de la superficie que viaja por la columna de agua, así, al incrementarse la profundidad también incrementa la edad y el aislamiento de comunidades que se ven expuestas a un cambio en las condiciones ambientales (e.g disponibilidad de energía), esto conlleva inminentemente a una estratificación selectiva de la diversidad microbiana encontrada, como se ejemplifica en la Figura 6 [41].

Derivado del incremento en las condiciones limitantes y de la profundidad del sedimento las tasas metabólicas de las comunidades microbianas decrecen 2 o 3 veces en comparación con las tasas metabólicas en condiciones de laboratorio. Es decir, las comunidades microbianas en estos ambientes sólo invierten energía para el mantenimiento de funciones básicas celulares (BPR - *Basal power requirement*, por sus siglas en inglés), en contraste a esto, se ha observado un aporte de biomasa constante en los mismos.

Así, la productividad asociada a la superficie de la columna de agua, la tasa de sedimentación de materia orgánica, la disponibilidad de energía y factores a nivel geoquímico, tienen como consecuencia directa la distribución diferencial de las comunidades microbianas a través del sedimento marino. Esta zonación resulta en la sucesión de especies aceptoras de electrones que tendrán un papel fundamental en la dinámica metabólica y biogeoquímica de las profundidades de los océanos.

En lo que concierne a las zonas del sedimento marino, la superficie del mismo se caracteriza por una bioturbación causada por la macrofauna (actúa 10-5 cmbsf [*centimeters below seaflor*]), dando lugar a una mayor captación y transporte de materia orgánica, creando al mismo tiempo un ambiente más heterogéneo y con mayor actividad microbiana.

Debajo de la zona de bioturbación, las condiciones limitantes y difusión de energía se mantienen con menos cambios. Sin embargo, a nivel bioquímico se pueden distinguir ciertas capas que obedecen a la concentración de especies tanto aceptoras como donadoras de electrones que permitirán la movilización de sustratos y productos por diferentes grupos de microorganismos [55].

La primera zona desde el inicio del sedimento marino, se conoce como zona de reducción de azufre (SR, *Sulfur reduction*, por sus siglas en inglés) y es la principal responsable de la remineralización de carbono. Debajo de esta zona encontramos la región de oxidación anaeróbica de metano dependiente de sulfato (SMT, *Sulfur methane transition*, por sus siglas en inglés) en donde se observan altas tasas de reducción de azufre (107-199 mbsf) y sólo cuando el azufre se acaba, la metanogénesis se vuelve el proceso de remineralización predominante [44] (ver Figura 6).

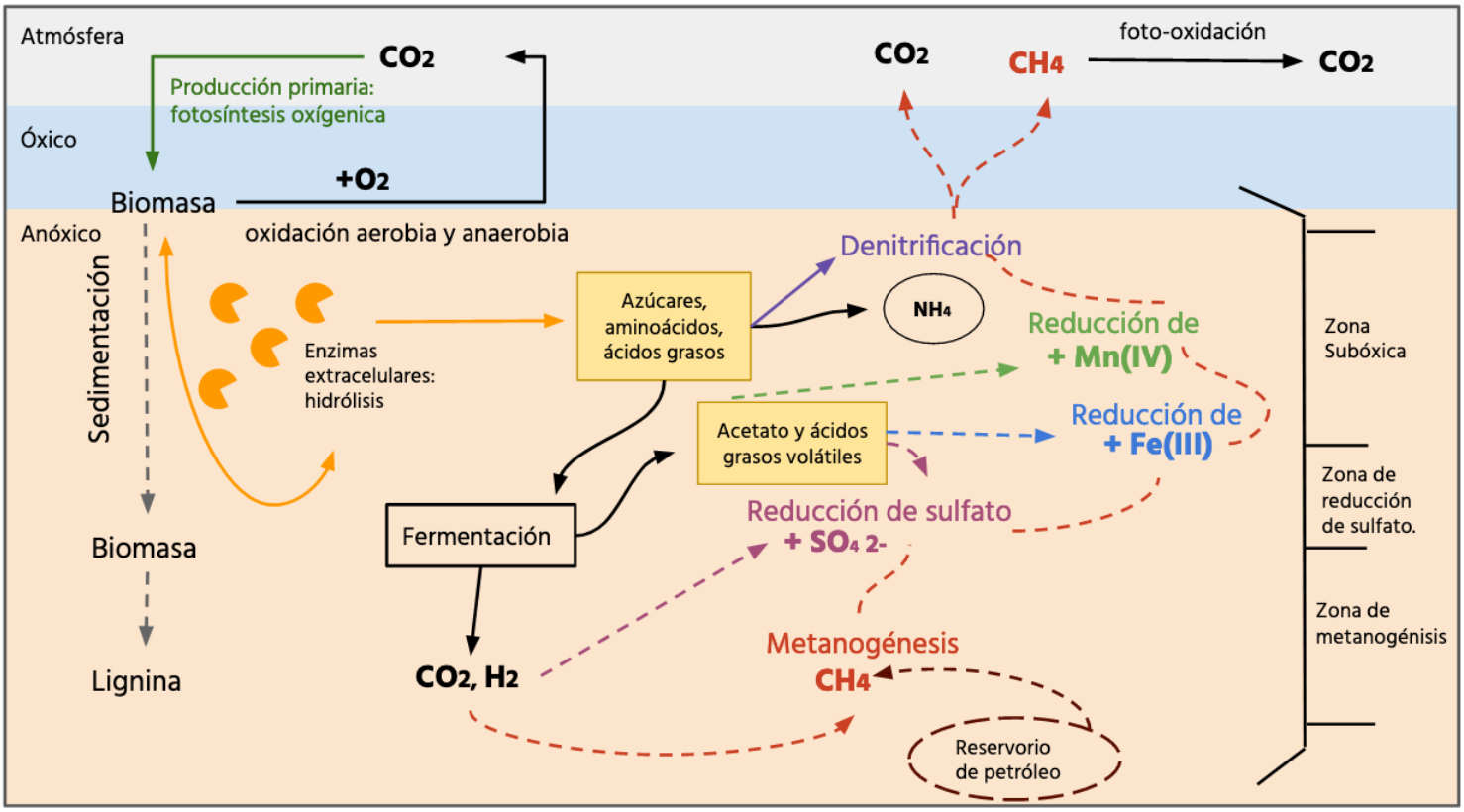

Figura 6: Diagrama de zonas a través del sedimento marino. Obtenido y modificado de Parkes J., 2014 [41]

Ahora bien, como se observa en la Figura 6 cerca de la superficie del sedimento, el oxígeno es rápidamente eliminado debido a la respiración aerobia, después los aceptores de electrones que se encuentran disponibles van disminuyendo en cuanto a su potencial de oxidoreducción, dando lugar a la zonación antes mencionada (*e.g* NO_3, MN^{4+}, Fe^{3+}, $SO_4{}^{2-}$).

Por otro lado, cuando no existen aceptores de electrones disponibles, la fermentación es la opción a la que recurren ciertos grupos de bacterias, produciendo H_2 y CO_2, que funcionan como donadores de electrones para movilizar las reacciones metabólicas necesarias. La acetogénesis, etanogénesis, propanogénesis y la reducción anaerobia de sulfato son vías del metabolismo predominantes en los sedimentos profundos [55].

Otra de las características determinantes para la naturaleza de la zonación geoquímica y por lo tanto con influencia en la diversidad funcional y taxonómica de las comunidades de los sedimentos, son tanto el tipo de roca madre del sedimento como la materia orgánica. Así, podemos encontrar sedimentos saprobios con alta concentración de materia orgánica en mares del Mediterráneo, sedimentos de arcilla que se caracterizan por un menor número de células, además de sedimentos ricos en hidratos de metano (*e.g* Ridge o Cascadia Margin).

Finalmente, existen sedimentos profundos, que podrían reflejar condiciones antiguas relacionadas con el origen de la vida, como aquellos en donde se ha observado síntesis de hidrocarburos de manera abiótica, esto en sedimentos donde se encuentran rocas ultramáficas y el calor de la corteza terrestre es moderado, así los procariontes que procesan estos hidrocarburos producen materia orgánica de manera abiótica, lo que puede tener implicaciones en el entendimiento del origen de la vida [41].

5. Justificación

La comparación de perfiles funcionales es un reto dentro del área de la metagenómica debido a la variación que representan los distintos diseños experimentales, el uso de dis-

tintas plataformas de secuenciación y el post-procesamiento que se le da a las secuencias. En consecuencia, es necesario desarrollar metodologías que permitan comparar muestras de metagenomas tipo *shotgun* sin importar la falta de información referente a las condiciones ambientales de donde las muestras fueron obtenidas y que adicionalmente tomen en consideración el nivel de representación de las enzimas que llevan a cabo las reacciones predominantes de dicho ambiente. Para este fin, se desarrolló una nueva metodología estadística que toma ventaja de la gran cantidad de muestras almacenadas en bases de datos públicas.

6. Hipótesis

La comparación de la representatividad estadística de cada enzima en un metagenoma perturbado contra la distribución de la misma enzima derivada de un grupo de metagenomas de referencia permitirá identificar aquellas que sirvan como base para describir el potencial metabólico diferencial del metagenoma perturbado.

7. Objetivo general

Proponer un método que permita estimar y contrastar los perfiles metabólicos de muestras metagénomicas.

7.1. Objetivos específicos:

- Seleccionar en bases de datos públicas metagenomas marinos de sedimentos.
- Agrupar las muestras de los metagenomas en tres categorías, la primera referente a ambientes perturbados (por ejemplo: exposición a derrames de hidrocarburos), ambientes no perturbados (metagenomas de referencia sin exposición aparente) y

consorcios enriquecidos en el laboratorio por algún tipo especifíco de compuesto (*e.g* xenobióticos)

- Procesar y anotar los datos crudos derivados de los metagenomas extraídos de las bases de datos públicas, e identificar a los genes que codifiquen para enzimas.
- Definir la distribución de probabilidad de cada enzima en las muestras de referencia.
- Seleccionar enzimas enriquecidas en los metagenomas contaminados según el nivel de significancia establecido para cada una de las distribuciones de probabilidad de las enzimas de la referencia
- Identificar enzimas definidas como clave en las rutas, las cuales se usarán como base para la reconstrucción de vías metabólicas.
- Describir el potencial metabólico de muestras perturbadas.

8. Método

8.1. Recolección y Procesamiento de datos: limpieza de las secuencias de ADN

Para la recolección de los datos se realizó una búsqueda en las bases de datos SRA de NCBI y MG-RAST. Las muestras recolectadas deben cumplir con las siguientes características:

- Muestras de sedimentos marinos.
- Plataforma de secuenciación utilizada: Illumina y sus múltiples variantes (Illumina Hiseq(1000,2000,2500), Illumina Miseq (1000),NextSeq 500 y Illumina Genome Analyzer IIx)
- Construcción de bibliotecas de DNA tipo paired-end.

Las características listadas a continuación, se encontraron disponibles para cierto grupo de muestras de todas las recolectadas. Por lo tanto, es importante mencionar que debido a la falta de estandarización de los metadatos, no todas las muestras cuentan con la misma información disponible.

- Información geográfica: Latitud y longitud.
- Características fisicoquímicas, como temperatura, pH, etc.
- Fecha de recolección de la muestra
- Profundidad del sedimento
- Publicación que refiera a los datos

Como se observa en el panel A de la Figura 7, el primer paso después de la recolección de las muestras, consistió en su agrupación con la ayuda de la información contenida en

los metadatos de las bases de datos. Así, se generaron tres categorías: No Perturbados (NP), en donde el ambiente no se ha visto afectado por cambios derivados de actividades humanas, en general, son muestras donde se estudian las condiciones naturales en las que habitan las comunidades microbianas, para descubrir nuevos taxa o funciones y para la comprensión a nivel evolutivo, metabólico y biogeoquímico.

La siguiente agrupación se denominó Perturbados (P), en donde existe evidencia fisicoquímica de la presencia de sustancias que alteran las condiciones del ambiente y por lo tanto, de las comunidades en cuestión. Ejemplos de tales condiciones, son los derrames de hidrocarburos, contaminación por Sb (antimonio), metales pesados, etc.

Finalmente, la categoría de Consorcios (C), agrupa muestras de metagenomas que fueron o enriquecidas *in situ*, o bien recolectadas y llevadas a laboratorio para enriquecerlos bajo ciertas condiciones controladas, *e.g* enriquecimiento de bacterias anaerobias, enriquecimiento para la reducción de perclorato, etc.

La división entre muestras Perturbadas y Consorcios se realizó con la finalidad de observar diferencias a nivel de representación de enzimas de las vías metabólicas encontradas en las muestras a analizar. Una vez definidos los grupos de muestras por categoría, según sus metadatos disponibles, el siguiente paso consiste en la evaluación de la calidad de las bibliotecas de DNA. Para completar dicho objetivo, se utilizó el programa FASTQC v.3.

La visualización de la calidad en de las secuencias tiene como finalidad, remover secuencias de baja complejidad, es decir, segmentos no resueltos por el secuenciador (*NNNN*), remover secuencias de adaptadores, secuencias con baja calidad de acuerdo al *puntuaciones Phred*, secuencias sobrerrepresentadas, así como la detección de contaminación en la muestra.

Con la finalidad de realizar un análisis exploratorio y comparativo de distintas herramientas para el procesamiento de datos metagenómicos se realizaron dos flujos de trabajo, mostrados en el diagrama de la Figura 7, el primero consiste en el uso del programa Trim Galore v.3 para la limpieza de las secuencias. En el flujo de trabajo 2, se

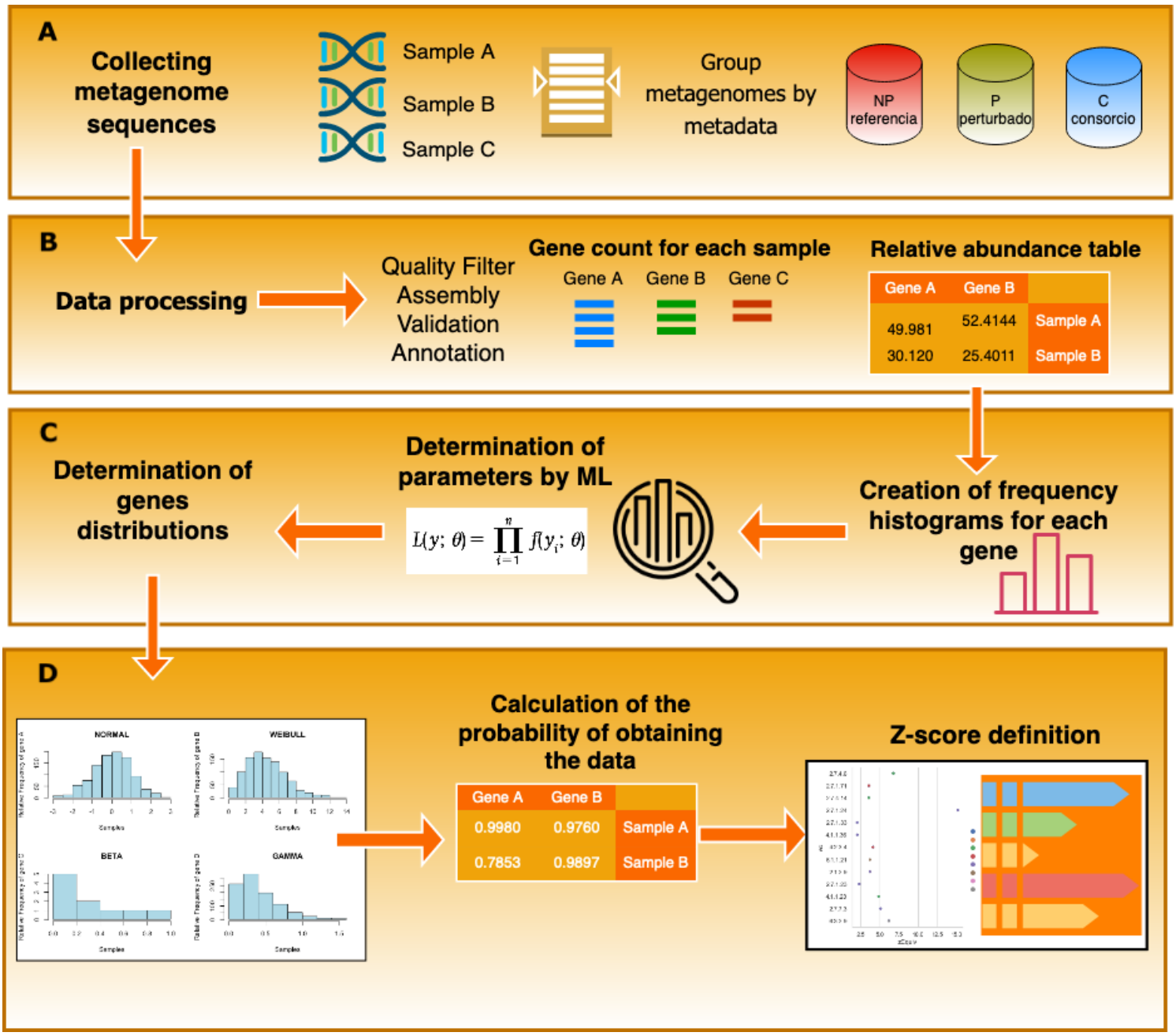

Figura 7: Estrategia Experimental

utilizó el programa fastp.

Trim Galore fue seleccionado principalmente porque autodetecta y elimina adaptadores. También y como consecuencia de la existencia de adaptadores, elimina el sesgo al inicio de los *reads*, que se observa como una sobrerrepresentación de nucleótidos. Ahora bien, aún cuando las estadísticas de calidad para algunas muestras parecían buenas, se aplicó un filtrado con los parámetros mínimos de corte al inicio y final de los *reads*. Los puntajes Phred de corte utilizados se ajustaron particularmente para cada muestra, variando de entre 8-15 como puntaje mínimo en la calidad de las bases por cada secuencia de ADN procesada.

Por otro lado, fastp es un software que ofrece la ventaja de distribuir el procesamiento de los datos por hilos, aumentando el rendimiento y velocidad de análisis, al mismo tiempo que los resultados de limpieza y filtrado son altamente confiables. Con esta herramienta se utilizó un puntaje mínimo de calidad de bases de 15, debido a los buenos resultados obtenidos para cada muestra.

La visualización de las estadísticas de calidad después de la limpieza con Trim Galore v.3 y fastp, permitieron corregir los parámetros para las muestras que lo requirieron. Para aquellas muestras, en las que no resultó favorable la filtración por calidades pues no mejoraba su calidad y se mantenían muy similares, se retuvo la biblioteca sin limpieza para el siguiente paso del flujo de trabajo: el ensamble.

Para una visualización general de los parámetros de limpieza utilizados para ambos programas y de las propiedades de las muestras después de la limpieza ver Anexo 1, sección *Limpieza de reads*.

8.2. Eliminación de duplicados

Durante la construcción de bibliotecas de DNA, antes del proceso de secuenciación, plataformas como Illumina realizan un paso de enriquecimiento de las secuencias a través de rondas de PCR, esto resulta en un número de *reads* repetidos o duplicados. Por

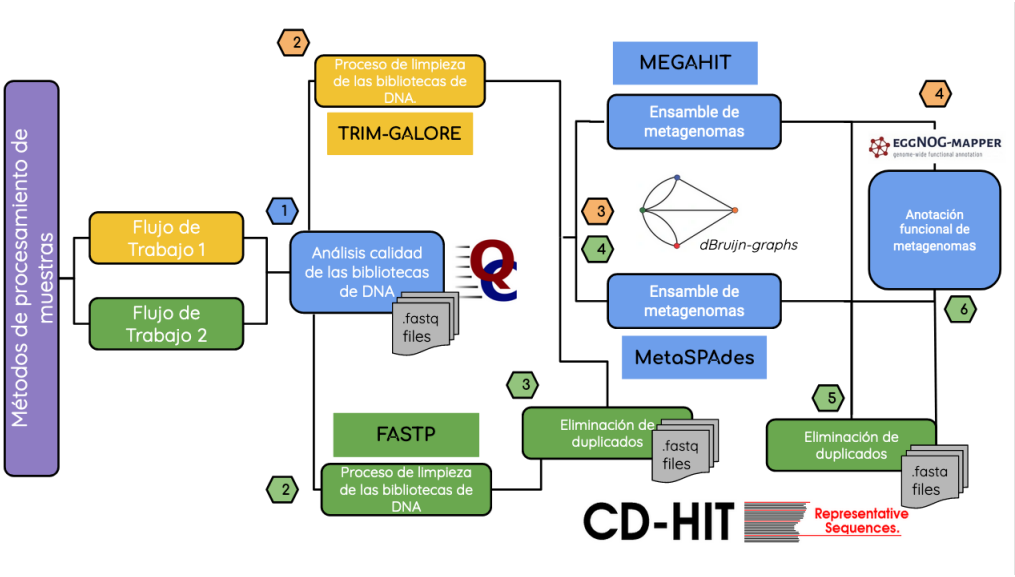

Figura 8: Metodologías utilizadas para el procesamiento de las muestras metagenómicas de sedimentos marinos

otro lado, los duplicados puede surgir también, de la identificación errónea de clusters separados provenientes de uno mismo por los software de captura de imágenes de las plataformas de secuenciación, a estos últimos se les llama duplicados ópticos [66]

Así, la eliminación de duplicados derivados de PCR, o bien duplicados ópticos, es obligatorio cuando se quiere evitar efectos negativos en la construcción del ensamble *de novo* y disminuir el uso de memoria computacional.

Como se observa en la Figura 8 para el flujo de trabajo número 2, se utilizó el software CD-HIT-DUP, con la finalidad de eliminar los duplicados ópticos y/o duplicados producto de PCR.

8.3. Ensamble de muestras metagenómicas

Existen una serie de factores que aportan cierta complejidad a las muestras metagenómicas, tales como:

- variación intraespecie en las comunidades microbianas
- la presencia de organismos no caracterizados

Esto, aunado a la dificultad de capturar información durante pasos como la extracción de DNA y la secuenciación (además de la subsecuente fragmentación del DNA en secuencias cortas) hacen del proceso de ensamble un reto, no sólo para obtener información de organismos no tan abundantes, sino también para resolver secuencias repetidas resultado de la variación entre cepas de una especie.

Los ensambladores que se fundamentan en la construcción de gráficos de *de Bruijn*, son los que mejor resuelven genomas a partir de bibliotecas de ADN, donde los *reads* son cortos y pareados como son los que resultan de la secuenciación con plataformas tipo Illumina.

De tal forma que, para las muestras analizadas en el presente proyecto, se utilizaron en paralelo y para cada una de las muestras por separado los programas MetaSpades y Megahit. El primero implementa mejor la captura de la variación intraespecie en la muestra y el segundo tiene la ventaja de aumentar el rendimiento computacional y resolver de igual forma el ensamble de datos metagenómicos [39].

- **Megahit:** Software que crea gráficas *de Bruijn* para resolver ensambles *de novo* de una manera muy eficiente, al usar gráficas sucintas de Bruijn. Reduce la necesidad de memoria y disminuye el sesgo derivado de errores de secuenciación. Además elimina *contigs* que presentan una baja cobertura durante la creación de los grafos [31]

  Git-hub: Megahit v1.2.9:

- **MetaSpades:** Ensamblador basado en la construcción de gráficos de *de Bruijn*. Detecta y elimina las quimeras que surgen en los *reads*, además de capturar información relacionada con la variación intraespecie o de cepas en las muestras [39] Spades v3.13.1

Recientemente, se ha reportado que estos dos *software* resultan ser eficientes y recomendados para esta tarea [60]. Las características importantes a resaltar para cada uno de estos ensambladores residen, en que MetaSpades tiene buenos resultados en términos de integridad y continuidad del ensamble, aún cuando las muestras tienen baja cobertura de secuenciación, además resuelve mejor genomas a nivel de cepas y especie, seguido de Megahit.

Al contrario, Megahit tiene mejores resultados cuando las muestras tienen una alta cobertura de secuenciación. Por otro lado, una de las ventajas del uso de Megahit, es que ha mostrado ser un ensamblador muy eficiente, en lo que se refiere al tiempo computacional requerido para llevar a cabo el ensamble, esto como resultado del cálculo de gráficos sucintos de de Bruijn (SDBG).

8.4. Validación de ensambles

Debido a la existencia de múltiples factores que contribuyen a la alta heterogeneidad de las muestras, tales como la complejidad del ambiente y variaciones entre muestras debido a la categoría asignada (referencia o perturbadas), fue necesario realizar una comparación entre los métodos utilizados para ensamblar, con la finalidad, de obtener aquél ensamble que refleje con mayor veracidad el contenido funcional.

Para la realización de este paso, se caracterizaron las siguientes métricas:

- Continuidad de los ensambles metagenómicos
- Integridad
- Precisión

con el uso de MetaQuast

A través de mapear los ensambles resultantes a genomas cerrados de referencia, MetaQuast evalúa la calidad del mismo. La selección de está herramienta, reside en su capacidad de proporcionar estadísticas de calidad en lo referente a las tres métricas (ver Figura 5), continuidad, precisión e integridad del ensamble, para evaluar finalmente, que ensamblador tuvo un mejor desempeño en la resolución del ensamble.

8.5. Eliminación de redundancia en el ensamble

A la porción de *contigs* en un conjunto de datos que son iguales con al menos otro *contig* dentro del mismo conjunto de datos o ensamble se le conoce como redundancia. Si bien es cierto, que para análisis de datos metagenómicos no se ha reportado la eliminación de redundancia del ensamble, el presente trabajo intenta realizar una comparación entre métodos, con la finalidad de elegir el ensamble con una mejor resolución y de acuerdo a las métricas mencionadas anteriormente. Para este paso y como se muestra en el diagrama de la Figura 7 en el flujo de trabajo no. 2, la eliminación de redundancia de los ensambles se realizó con el software CD-HIT-DUP.

8.6. Anotación funcional

Con la finalidad de anotar a nivel funcional los ensambles de las muestras, se utilizó el software eggNOG-Mapper, (Evolutionary genealogy of genes Non-supervised Orthologous Groups, por sus siglas en inglés). Este software tiene la ventaja de asignar genes ortólogos para hacer las inferencias funcionales al usar clusters y filogenias pre-computadas, almacenadas en la base de datos de eggNOG, a diferencia de las anotaciones realizadas por transferencia de homología [23],[24]. Aunado a esto, la herramienta de alineamiento contra las base de datos que utiliza eggNOG esDIAMOND, el cual tiene la ventaja de ser de 100 a 10,000 veces más rápido y de recuperar verdaderos positivos en comparación con BLAST e INTERPROSCAN [24] (para mayor detalle en el uso de parámetros de esta

herramienta, ver anexo 1, sección: Anotación Funcional)

8.7. Estimación del potencial metabólico

En el campo de la estadística, al conjunto de observaciones que describen el comportamiento global de una población se le denomina muestra.

En este caso concreto, las anotaciones funcionales derivadas de los metagenomas analizados representan la(s) muestra(s) que describe(n) de manera general el metabolismo que llevan a cabo grupos de procariontes habitantes de sedimentos marinos.

Para cada una de las anotaciones asignadas, de cada muestra de los 3 conjuntos de metagenomas se construyó una tabla de abundancias. Cada una de las abundancias observadas se dividió entre el número total de enzimas dentro de cada muestra, con la finalidad de reducir el sesgo por las diferencias en la profundidad de secuenciación de cada muestra.

La frecuencia relativa se calculó de la siguiente manera:

$$RF = n/N(10^6)^1 \tag{2}$$

Donde, *RF* se refiere a la frecuencia relativa para cada una de las enzimas anotadas en el metagenoma, *n* el número de veces que se encuentra la enzima en el metagenoma analizado, *N* el número total de enzimas en todo el metagenoma. Con la finalidad de hacer a las frecuencias relativas manejables en términos matemáticos se multiplicó por una cantidad fija, en este caso de 10^6.

Para este paso del cálculo de la frecuencia relativa se realizaron scripts *ad hoc* en el lenguaje de programación Python.

Es importante mencionar que las frecuencias relativas de las enzimas son cantidades muy pequeñas, con una varianza mayor a la esperada debido a efectos al azar en el

[1]Transformación de las frecuencias relativas al multiplicar por 10^6.

proceso de anotación, por lo tanto no se puede considerar como un valor constante sino como un valor obtenido de una distribución.

Para el siguiente paso que consiste en el cálculo de la distribución de probabilidad para cada una de las enzimas de las muestras de referencia, se asumió que la ocurrencia de cada una de las enzimas del grupo de datos de la referencia sigue una distribución de probabilidad multinomial o Poisson. Este supuesto se justifica debido a que las frecuencias de más de mil de enzimas anotadas en una muestra son muy bajas y que la expresión de la ec. (2) corresponde al estimador de máxima verosimilitud para el valor esperado. Dado que cada una de las frecuencias relativas de las enzimas anotadas representa una variable continua, para observar la tendencia central y la forma de la distribución de los datos se construyó un histograma de frecuencias. Después, los datos se ajustaron a una de las siguientes distribuciones: Gamma, Normal, Lognormal o Weibull, estas distribuciones candidatas se ajustaron a los datos a través del método de Máxima verosimilitud. Los métodos de máxima verosimilitud estiman los parámetros poblacionales como la Esperanza poblacional, al maximizar la verosimilitud de observar los datos.

En este punto, es importante tener en mente que la verosimilitud no es lo mismo que la probabilidad, asimismo la función de verosimilitud utilizada para el cálculo de parámetros, no es lo mismo que la distribución de probabilidad. Aclarando, en una distribución de probabilidad de una variable aleatoria, los parámetros de la distribución están fijos y los datos son las variables desconocidas. En cambio en una función de verosimilitud los datos son considerados como la variable permanente o fija y son los parámetros los que varían en múltiples valores. Sin embargo, la verosimilitud de los datos dado valores particulares de parámetros esta relacionado con la probabilidad de observar estos datos dado un valor específico para los parámetros[46].

La siguiente función de verosimilitud es la utilizada de forma iterativa hasta maximizar la verosimilitud de los datos:

$$L(\theta) = \ln[\prod_{i=1}^{n} f(Yi;\theta)] = \sum_{i=1}^{n} \ln[f(Yi;\theta)] \tag{3}$$

Cuando se analizan distribuciones no normales, los estimadores por máxima verosi-

militud necesitan ser calculados utilizando algoritmos iterativos complejos. Así, la verosimilitud de los datos dado un parámetro particular se relaciona con la probabilidad de obtener esos datos asumiendo los parámetros descritos [46].

Existen varias pruebas que nos hablan de la bondad del ajuste de los datos a la distribución de probabilidad dada, entre ellos se encuentra la llamada Anderson-Darling, la cual se utilizó para darle soporte estadístico a la elección de distribuciones de probabilidades elegidas para los datos.

Con la finalidad de conocer la representatividad estadística de cada enzima en las muestras perturbadas en contraste a las muestras de referencia o no perturbadas se calculó la probabilidad de observar los datos (frecuencias relativas), para cada una de las enzimas de las muestras dentro de la categoría de perturbados y consorcios en las distribuciones de probabilidad de las enzimas de las muestras de referencia.

A continuación se describe el flujo de trabajo estadístico con mayor detalle: El flujo de trabajo estadístico se desarrolló con el lenguaje de programación Python desarrollando scripts *ad hoc*, utilizando las siguientes librerías: pandas, numpy, scipy, matplotlib, seaborn, math y fitter.

Como primer paso para el análisis, se redujó el campo total de enzimas a través de las 19 muestras, a 1667 enzimas, en donde al menos el 80 % de las muestras compartieran cada una de las 1667 enzimas.

El siguiente paso consistió, en la asignación de una distribución de probabilidad a cada una de las enzimas de las muestras de referencia. La primera etapa, consistió en la prueba de hipótesis para probar si los datos son de tipo normal con el test de Shapiro-Wilk. Es importante mencionar, y como se explica en la sección de métodos, para ésta y las siguientes pruebas se utilizaron las frecuencias relativas al número total de enzimas anotadas (EC numbers) para cada muestra, multiplicado por 1,000,000.

Como se observa en la Figura 28, en el diagrama nombrado "Distribuciones de probabilidad de enzimas en la Referencia", el 16 % fueron asignadas a una distribución de tipo normal.

Las asignaciones a distribuciones de probabilidad para las enzimas diferente a la normal, fueron calculadas con la librería de python: Fitter. Resultando el 50 % de datos de enzimas ajustadas a una distribución tipo gamma, 4 % a lognormal y 30 % a una distribución de probabilidad tipo weibull.

Finalmente, para cada una de las enzimas y su distribución de probabilidad asignada de las muestras problema o perturbadas (mostradas en el Cuadro no.3), se calculó la probabilidad de observar cada frecuencia relativa dado los parámetros de su distribución. Para hacer más fácil la interpretación ésta serie de probabilidades P_k, fueron utilizadas posteriormente para su conversión a un punto quantil o puntuación z usando la ecuación:

$$P_k = \int_{-\infty}^{Z} \mathrm{d}x f(x), \tag{4}$$

en donde, $f(x)$ es la función de densidad de probabilidad de una variable normal con distribución $\mathcal{N}(0, 1)$ (normal con media cero y desviación estándar unitaria), de este modo, la probabiliad original se expresa en términos de la distancia de un punto a partir de la media en términos de desviaciones estándar.

9. Resultados

9.1. Construcción de la base de datos e información de las muestras analizadas.

En primer lugar se muestran las características generales de las muestras de metagenomas de sedimento marino, utilizadas para la construcción de la base de datos que conforma la referencia y las muestras perturbadas. Donde la Tabla 2 contiene la información de las muestras de referencia y la Tabla 3 acerca de las muestras de sedimentos con evidencia de perturbación.

Cuadro 2: Información general de las muestras dentro de la categoría de Referencia.

SRA ID	Origen	Método de secuenciación	Proyecto ID	Prof.(m)	Latitud	Longitud.	Fecha de Recolección	Anotaciones	Prof. de secuencia-ción
SRR11582139	Costa Rica	NextSeq 500	PRJNA627197	NA	10.301239	85.610549	2017-02-17	682,027	26,905,740
SRR1627907	Costa Rica	llumina HiSeq 1000	PRJNA264715	93.98	8.59234666	84.07735833	2011-04-03	46,486	17,120,033
SRR1971625	Noruega	Illumina HiSeq 1000	PRJNA248084	NA	72.003	14.731	2001-2003	163,573	15,402,223
SRR2133847	USA	Illumina HiSeq 2000	PRJNA290197	0.0-0.12	44.6695	-125.0983	2011-09-00	696,161	65,316,491
SRR3095933	Dinamarca	Illumina HiSeq 2500	PRJNA297401	NA	56.1033	10.4578	2013-05	1,615,881	25,597,048
SRR4026060	Grecia	Illumina Genome Analyzer IIx	PRJNA336970	NA	36.5	25.45	NA	148,938	10,566,474
SRR5229884	China	Illumina HiSeq 2500	PRJNA355347	5	24.83442	121.96201	2015-05-26	537,878	18,841,269
SRR5242450	Japón	Illumina MiSeq	PRJNA373808	NA	37.3104	-137.2311	2015-10-00	70,447	247,962
SRR5242455	Japón	Illumina MiSeq	PRJNA373808	NA	37.3602	-137.4147	2015-05-00	73,739	954,176
SRR5881625	China	Illumina HiSeq 2500	PRJNA364899	NA	47.5233	-125.0078	2014-10-16	107,447	25,655,580
SRR6193124	USA	Illumina HiSeq2000	PRJNA366549	NA	37.474067	-121.973033	2011-12-09	956,704	36,450,510
SRR6201607	Puerto Rico	Illumina HiSeq2000	PRJNA366555	NA	17.951083	-67.193167	2011-12-14	1,257,514	3,139,806
SRR6344986	Océano Pacífico	Illumina HiSeq 1000	PRJNA420998	14.96	44.873504	-125.151235	2002-07-14	129,339	7,891,137
SRR6660647	Océano Atlántico, Noruega	Illumina HiSeq 2000	PRJNA431796	NA	72.005	14.733333	2003-06-30	164,006	27,710,507
SRR7051260	México	Illumina Hiseq	PRJNA445016	NA	27.0114	-110.5956	2016-12-24	689,311	25,385,828
SRR8581483	Reino Unido	Illumina HiSeq 2500	PRJNA522699	0.015	NA	NA	2015-04-00	192,217	6,338,879
SRR8709623	Océano Pacífico	Illumina HiSeq 2500	PRJNA526329	0.01	11.91	144.93	2014-11-00	711,280	23,589,982
SRR9649755	Golfo de México	Illumina HiSeq 2500	PRJNA553005	NA	NA	NA	NA	1,213,338	21,308,190
ERR2031032	Namibia	NextSeq 500	PRJEB21737	4.5	-26.01	12.573	2008-04-00	721,556	24,765,266

Cuadro 3: Información general de las muestras dentro de la categoría de Perturbados.

SRA ID	Origen	Método de secuenciación	Proyecto ID	Prof. (m)	Latitud	Longitud	Fecha de recolección	Anotaciones	Prof. de secuencia-ción
ERR1992809	Semi-Synthetic marine	Illumina HiSeq 2500	PRJEB20198	NA	NA	NA	2020-01-06	236,923	12,500,000
SRR8457023	Golfo de México	Illumina MiSeq 1000	PRJNA515837	93.98	27.206659	-91.010554	2017-04-01	8,696	1,737,901
SRR11308316	Golfo de México	NextSeq 500	PRJNA609564	1295	19.93113	94.3407	2016-03-18	747,257	39,650,585
SRR11308317	Golfo de México	NextSeq 500	PRJNA609564	2967	25.87972	94.6686	2016-04-06	1,625,338	73,575,564

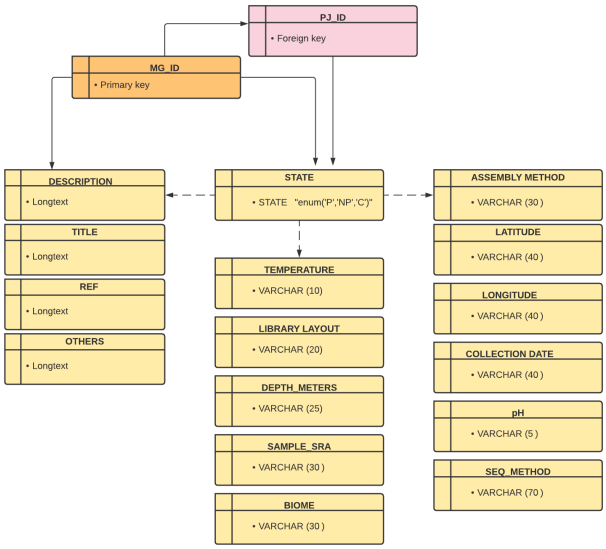

Figura 9: Instancias que conforman la tabla Metadatos, de la base de datos de las muestras recolectas, base realizada con la herramienta Mysql.

Los criterios mencionados en la sección de metodología, para la selección de las muestras en las bases de datos públicas, tienen como principal objetivo analizar muestras con características homogéneas, para evitar sesgos introducidos por las distintas plataformas de secuenciación, ver Tabla 2 y 3, de la sección de Método. Así, cada una de las muestras deriva de secuenciación tipo Illumina (con sus variantes) y son bibliotecas tipo paired-end.

Además, con la finalidad de construir una base de datos lo más completa posible para las subsecuentes interpretaciones a nivel biológico de los resultados, los metadatos de cada muestra fueron obtenidos. Sin embargo, es importante mencionar la existencia de muestras con falta de metadatos (*e.g* profunididad del sedimento, fecha de recolección, etc.) debido su inexistencia en las bases de datos públicas.

En la Figura 9 se muestra la estructura de la tabla principal, nombrada **Metadata** de la base de datos construida.

Figura 10: Mapa de la ubicación geográfica de los metagenomas de sedimento marino analizadas.

Por otro lado, en el mapa de la Figura 10 se observa la localización geográfica de cada una de las muestras metagenómicas analizadas en el presente proyecto.

9.2. Análisis de datos metagenómicos

Como se ha mencionado anteriormente, los datos metagenómicos se caracterizan por tener un gran número de sesgos que afectan las interpretaciones biológicas. De tal forma que, con la intención de eliminar los efectos de éstos sesgos, que pueden provenir de las plataformas de secuenciación, *e.g* secuencias duplicadas o bien, de la redundancia en los ensambles finales debido a el procesamiento de cada ensamblador, como resultado se realizaron dos protocolos en el análisis de datos (como se muestra en el diagrama de la Figura 8), con el objetivo de comparar la resolución final de cada ensamble por cada muestra.

Como primer paso, la limpieza de las secuencias de DNA fue realizada con parámetros

ajustados a las características de cada una de las muestras, tales como la cantidad de bases de baja complejidad, el sesgo al inicio de los *reads*, la presencia o no de adaptadores o secuencias repetidas, etc., con los programas Trim Galore y fastp. Los parámetros utilizados para cada una de las muestras se pueden consultar en el Anexo 1, en la sección de Limpieza de *reads*.

De acuerdo a las características anteriormente mencionadas, para cada una de las muestras, la cantidad total de *reads* por biblioteca puede o no disminuir considerablemente después de la limpieza.

En los gráficos de barras de las Figuras 11 y 13, se muestra el número de secuencias de DNA antes (barras color azul) y después (barras color naranja) de la limpieza por calidades con Trim Galore v.3, para el flujo de trabajo no. 1 (ver Figura 7).

Por otro lado, en las Figuras 12 y 14, se muestra tanto el filtrado por calidades, realizado con Fastp, antes (barra color naranja) y después (barra color azul), aunado a la eliminación de redundancia derivado de posibles duplicados producto de PCR (barras color verde), en el flujo de trabajo no. 2 (ver Figura 8).

Como se puede observar en las gráficas de las Figuras 11,12,13 y 14 el filtrado por calidades (barras color naranja), realizado con Trim-Galore y Fastp genera resultados similares tanto en el tamaño final del las secuencias, como en el número de secuencias totales, las gráficas de las Figuras 46 y 47 del Anexo 1, muestran estas características para un grupo de muestras que fueron limpiadas tanto con Trim-Galore como con Fastp.

Para el caso específico del Flujo de trabajo no. 2 la eliminación de redundancia después del filtrado por calidades de las bibliotecas de DNA con la herramienta CD-HIT-DUP no tienen un efecto significativo.

Esto indica que la eliminación de redundancia en las bibliotecas de DNA depende en gran parte de los protocolos seguidos durante la secuenciación para cada muestra en particular, de allí que no se observe ningún patrón específico.

Sin embargo, existen muestras como la ERR2031032, SRR13515399, SRR1971625,

SRR4026060 y SRR5396644 (ver Figura 12), con una secuenciación en plataformas: Nextseq 500, Illumina NovaSeq 6000, Illumina HiSeq 1000, Illumina, NextSeq 500, respectivamente, en las cuales el número de secuencias después de la eliminación de duplicados disminuyó.

A pesar de que se ha reportado la presencia de duplicados producto de PCR, en estudios en donde se utiliza secuenciación tipo 454 o pirosecuenciación [20], [13], se sabe que en un protocolo de secuenciación para tecnologías tanto 454 como Illumina, realiza rondas de PCR para aumentar la cantidad de DNA en las muestras a secuenciar, por lo que, la eliminación de duplicados es altamente recomendado para evitar este tipo de sesgos.

Desafortunadamente, no es posible asociar una variante del tipo de secuenciación Illumina con la presencia o no de duplicados, pues esto depende directamente de la concentración inicial de DNA de la muestra y de las rondas de PCR elegidas durante el protocolo de secuenciación para el enriquecimiento de DNA.

Ahora bien, la dificultad para distinguir entre duplicados naturales y producto de PCR es mayor, en comparación con la detección de duplicados ópticos. En consecuencia, los parámetros utilizados para este paso, se ajustaron con la intención de detectar *reads* exactamente iguales, donde el parámetro *-e* = 0 de Fastp, se refiere al número exacto de mistmatches permitidos, como los duplicados ópticos son producto de clusters iguales, la probabilidad de que dos secuencias sean exactamente iguales es muy baja y estos son eliminados con esta opción.

Además el parámetro *-m* = false, que se refiere a si la longitud de los *reads* deben ser exactamente iguales para considerarlos duplicados, permite tener una ventana de posibilidades en la detención de duplicados naturales o producto de PCR, los cuáles son más difíciles de distinguir, para este aspecto el hecho de que las bibliotecas de DNA sean tipo PE, disminuye la probabilidad de que dos secuencias sean exactamente iguales.

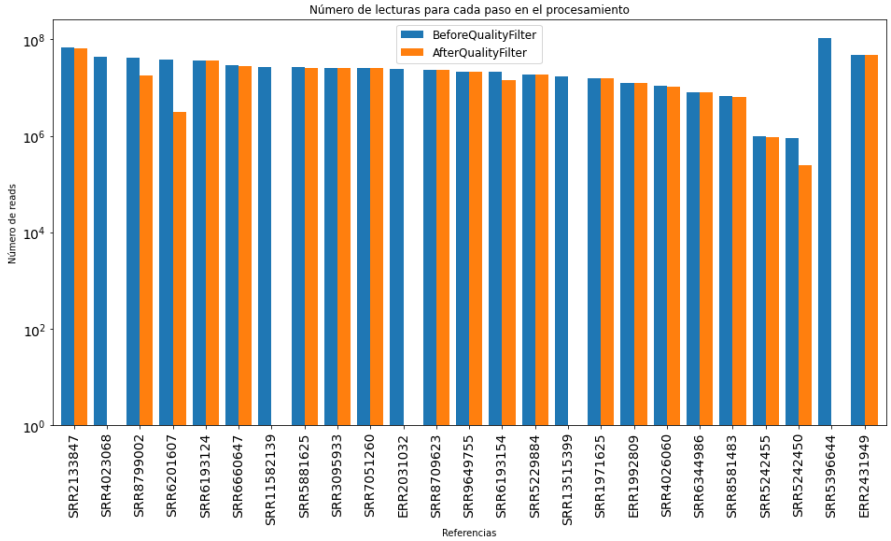

Figura 11: Número de secuencias después de cada paso en el procesamiento de filtrado por calidades, con el Flujo de trabajo 1, para las muestras clasificadas como Referencias. Es importante resaltar que aquellas barras sin su par,representan muestras con una calidad buena y por lo tanto no fue necesario realizar un filtrado de calidades

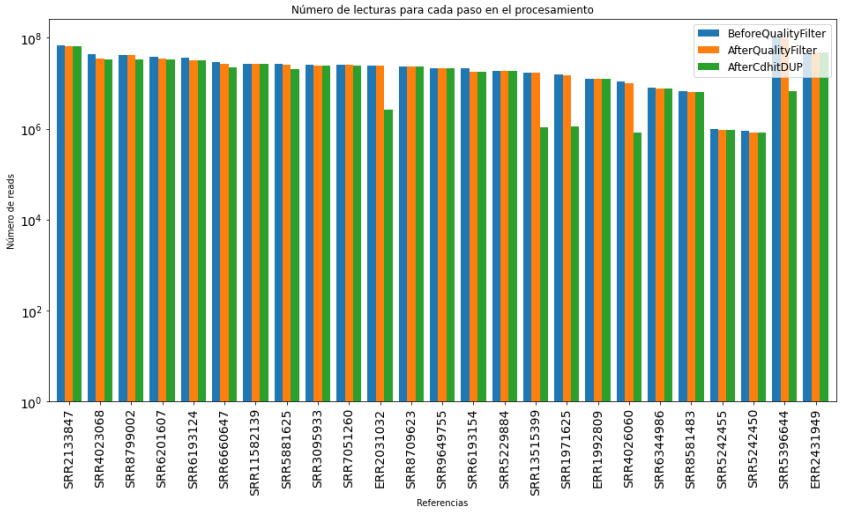

Figura 12: Número de secuencias después de cada paso en el procesamiento de filtrado por calidades, con el Flujo de trabajo 2, para las muestras clasificadas como Referencias.

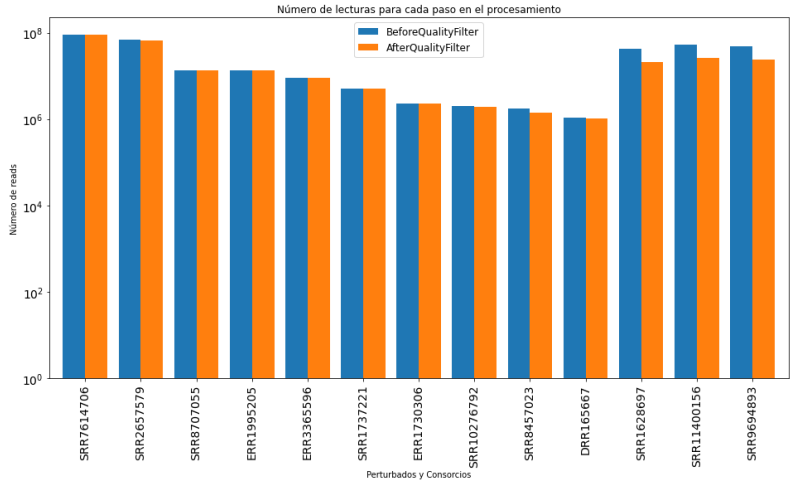

Figura 13: Número de secuencias después de cada paso en el procesamiento de filtrado por calidades, con el Flujo de trabajo 1, para las muestras clasificadas como Perturbadas y consorcios. Nota: Las últimas tres muestras de izquierda a derecha del gráfico representan los consorcios

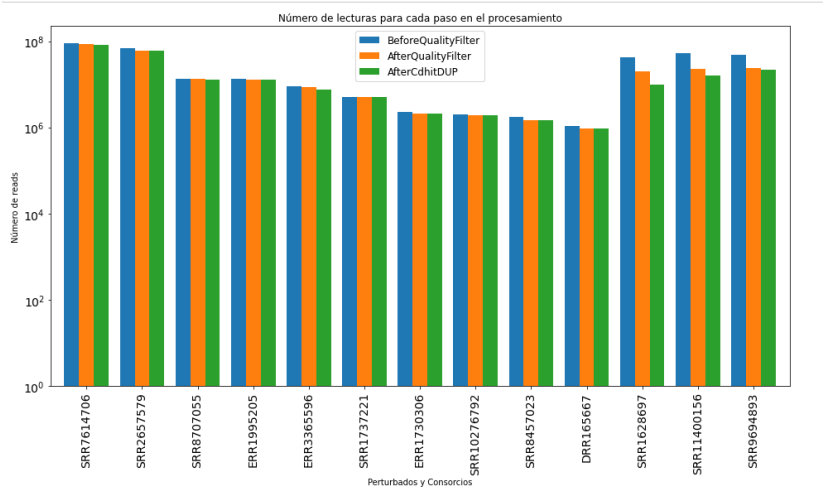

Figura 14: Número de secuencias después de cada paso en el procesamiento de filtrado por calidades, con el Flujo de trabajo 2, para las muestras clasificadas como Perturbadas y consorcios. Nota: Las últimas tres muestras de izquierda a derecha del gráfico representan los consorcios

9.3. Ensamble y validación

Para el ensamble de los metagenomas, se utilizaron dos herramientas: Megahit y MetaSpades.

Cuando se utilizan ensambladores basados en la construcción de gráficos de *de Brujin*, la elección del tamaño de *k-meros*, se debe tomar en cuenta, así mientras más pequeño su tamaño, se pueden generar más conexiones, sin embargo, se dificulta la resolución de secuencias repetidas. Por otro lado, a mayor longitud de *k-mero*, se crean *contigs* más largos. Para darle respuesta a éste problema, la mayoría de los ensambladores construyen gráficos de manera iterativa y con múltiples tamaños de *k-meros*.

Por un lado, MetaSpades, extrae los *contigs* generados con el tamaño mínimo de *k-mero* establecido y luego los utiliza para la construcción de *contigs* de mayor tamaño, con el siguiente tamaño de *k-mero*, sin dejar de lado la información propia de los *reads* y de los *contigs* pre-ensamblados, ya que se centra en capturar la mayor cantidad de información de especies raras o poco abundantes, a su vez que ayuda a capturar la variación intraespecie.

Megahit, también utiliza un rango de tamaño de *k-meros* para iterar y mejorar el ensamble, pero este sólo mantiene *k-meros* con alta precisión, es decir, que aparecen más de una sola vez o que tienen mayor cobertura e implementa una estrategia para recuperar conexiones con baja profundidad, al tomar *reads* con alta calidad e incrementando el tamaño del *contig* en las regiones con baja cobertura, esto se ve reflejado en la eficiencia y rapidez a nivel computacional del programa.

Así, los parámetros elegidos en lo referente al tamaño de *k-meros* para los dos programas utilizan rangos que mejoran su rendimiento y resolución de ensambles, como se muestran en el Anexo 1 en la sección de Ensamble de muestras.

Por consiguiente, con el objetivo de evaluar la calidad y resolución de los ensambles derivados de los dos *software* mencionados anteriormente, la validación de los mismos, se enfocó en la obtención de métricas basadas en el mapeo del ensamble hacia una referen-

cia de genomas cerrados, con la herramienta MetaQuast, que proporciona información referente a la continuidad, precisión e integridad del ensamble.

Esto permitirá elegir el ensamble de acuerdo a su consistencia derivada del rendimiento propio del ensamblador utilizado.

La primera métrica utilizada para validar los ensambles hace referencia a la integridad del mismo. Para obtener esta métrica los ensambles obtenidos de los metagenomas se alinean contra una base de datos (en este caso MetaQUAST lo hace contra SILVA 16S rRNA) y así se puede conocer la porción de bases totales del ensamble alineadas a los genomas de referencia identificados en el ensamble metagenómico.

En la Figura 15 se muestran las fracciones de bases alineadas a la referencia de los ensambles de las muestras, tanto de referencia como perturbadas con cada uno de los ensambladores utilizados para el flujo de trabajo 1 y 2. Se puede observar que independientemente del procesamiento anteriormente descrito, en ambos flujos de trabajo y para ambos ensambladores la integridad de los ensambles es muy similar. Algo interesante a resaltar, es que la eliminación de redundancia puede estar jugando un papel importante en la porción de bases alineadas a la referencia ya que existen muestras con mayor fracción de genoma para los dos ensambladores en el Flujo de trabajo que elimina redundancia.

La siguiente propiedad analizada es la continuidad del ensamble, la cual permite conocer que tan resuelto se encuentra el ensamble en cuestión. Dentro de ésta propiedad, existen métricas que hablan de las propiedades intrínsecas del mismo, como son: longitud del ensamble, número de *contigs* y longitud de *contig* más largo. También existen aquella métricas que reflejan la continuidad del ensamble al comparar con genomas cerrados encontrados en las bases de datos, tales como: longitud total alineada al genoma de referencia, y el alineamiento continuo más largo. En las Figuras 16-20, se muestran esta serie de métricas con la finalidad de comparar la resolución a nivel de continuidad entre ambos flujos de trabajo y entre ambos ensambladores. Las Figuras muestran que las métricas de la longitud y el número de contigs totales de los ensambles, aumentan en el segundo flujo de trabajo. Métricas de continuidad, como la longitud el contig más

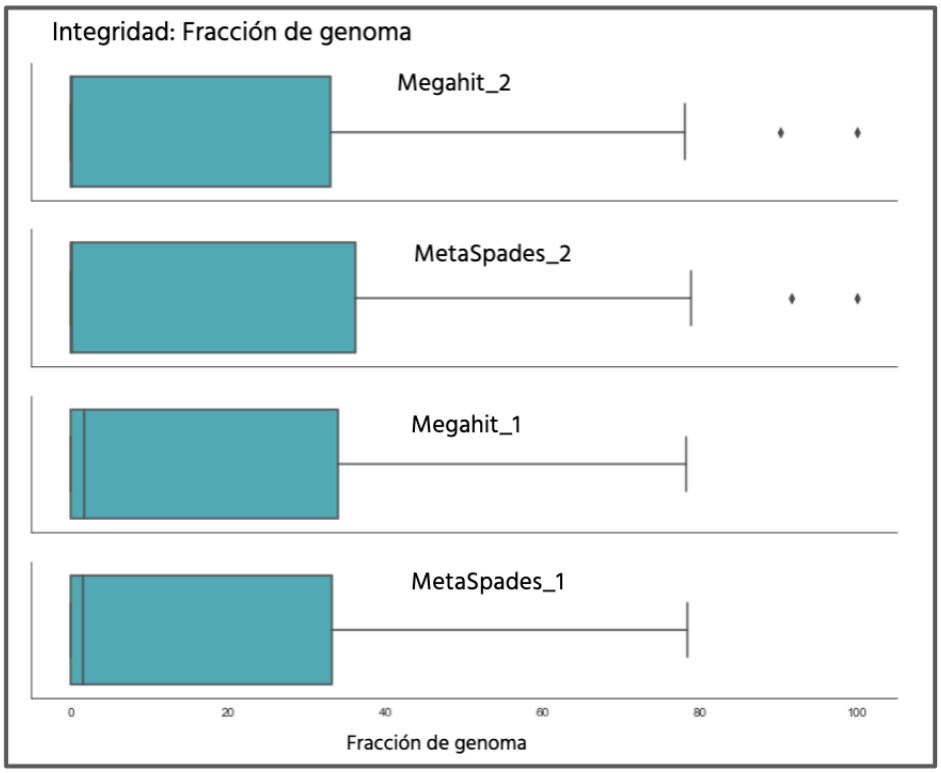

Figura 15: Integridad del ensamble, en términos de la fracción de genomas recuperados en cada uno de los dos métodos para los dos ensambladores: Megahit_1 y MetaSpades_1, Megahit_2 y MetaSpades_2, hacen referencia al flujo de trabajo no. 1 y 2

largo, longitud del ensamble alienado a la referencia y el del alineamiento a la referencia más largo, también aumentan en el segundo flujo de trabajo, esto refleja que las propiedades de los ensambles mejoran para este flujo de trabajo, en donde, MetaSpades parece generar ensambles con mayor continuidad seguido de Megahit.

El número de inserciones o deleciones que el ensamblador generó durante el proceso, hace referencia a la precisión del ensamble. Tanto las inserciones como las deleciones se identifican cuando se hace una comparación del ensamble con genomas de referencia (SILVA 16S rRNA). Como se observa en las figuras 21 y 22, se puede apreciar la cantidad de errores que los ensambladores generaron durante el proceso tanto en el grupo de muestras del flujo de trabajo número 1 (Megahit_1 y MetaSpades_1), como para las muestras del flujo de trabajo no. 2 (Megahit_2 y MetaSpades_2).

El número de deleciones e inserciones en los ensambles, para los dos ensambladores del segundo flujo de trabajo aumentan, como consecuencia del aumento en su tamaño (que se vió reflejado en la continuidad de los ensambles), con un mayor número para los ensambles realizados con MetaSpades, seguido de Megahit.

Ahora bien, es importante mencionar que no existe ningún valor de referencia para medir la precisión del ensamble como buena o de mala calidad. Todo esto debido a una serie de variables como la complejidad del ambiente, el ensamblador utilizado y la naturaleza de la muestra. Wang Z. y colaboradores muestran que para algunos ensambladores como Faucet el aumentar la cobertura de secuenciación mejoran las métricas que evaluan la presición. MetaSpades y Megahit tienen métricas de presición similares a Faucet.[60]

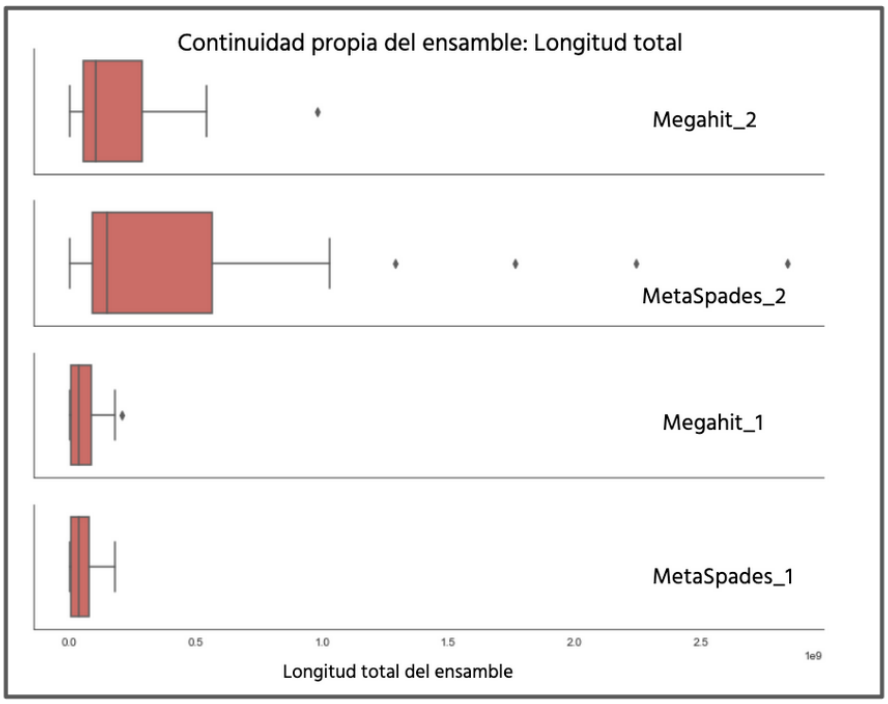

Figura 16: Longitud total del ensamble.

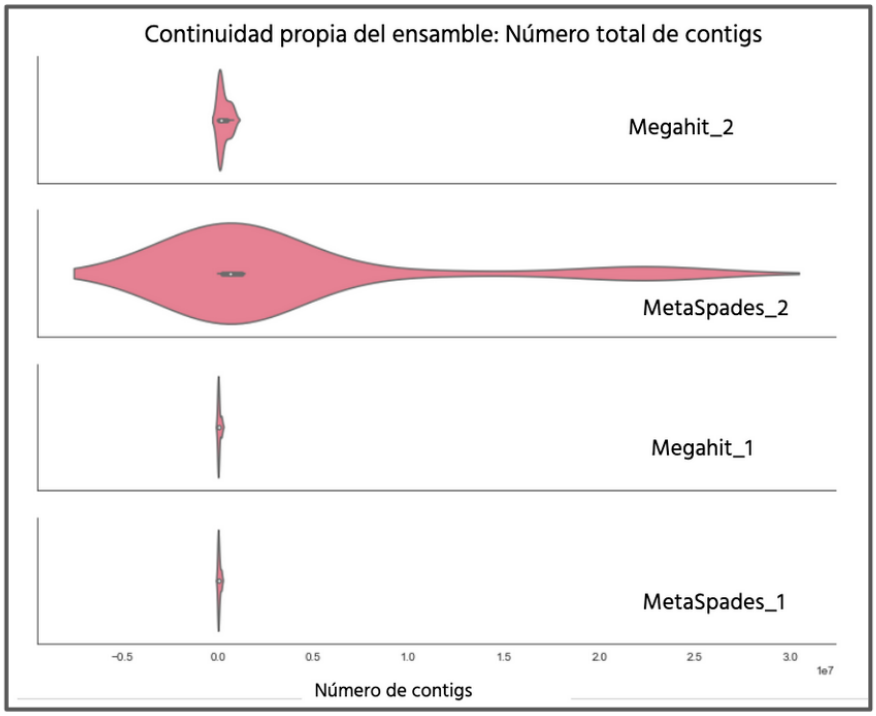

Figura 17: Número de *contigs* totales por ensamble.

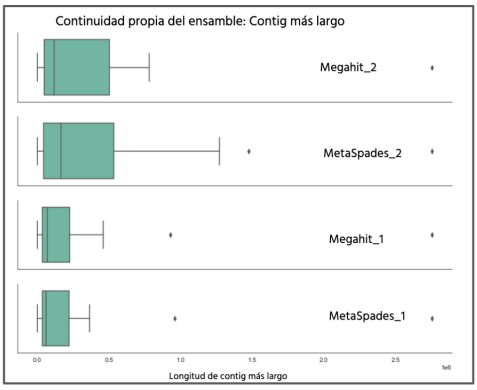

Figura 18: Longitud del *contig* más largo.

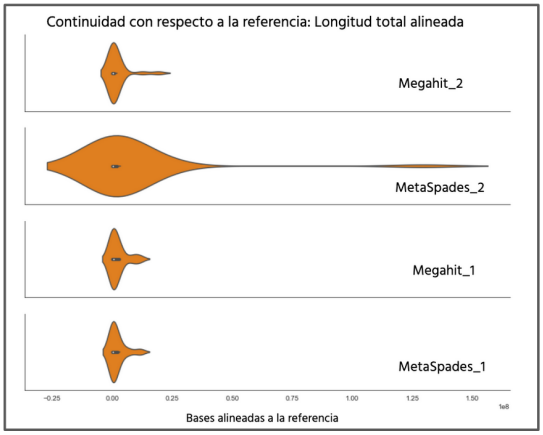

Figura 19: Longitud total del ensamble alineada a la referencia.

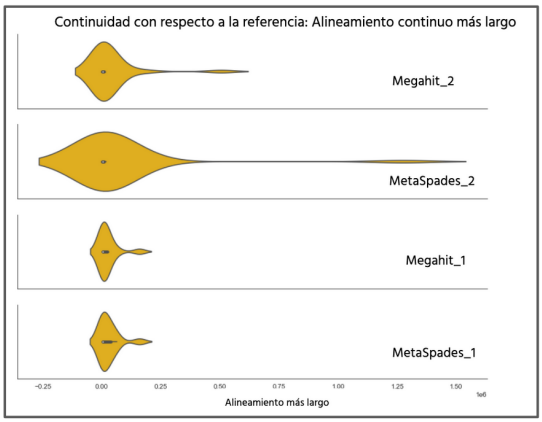

Figura 20: Longitud del alineamiento a la referencia más largo.

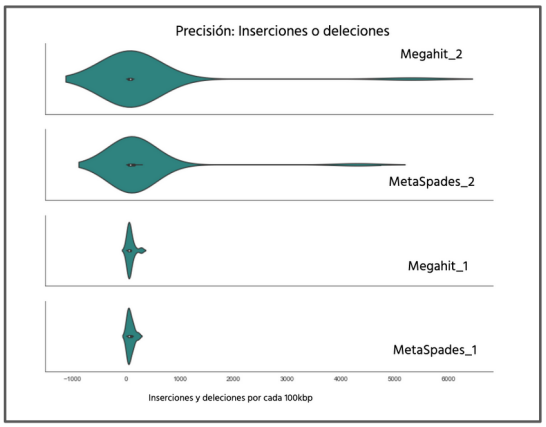

Figura 21: Número total de inserciones y/o deleciones por cada 100kpb

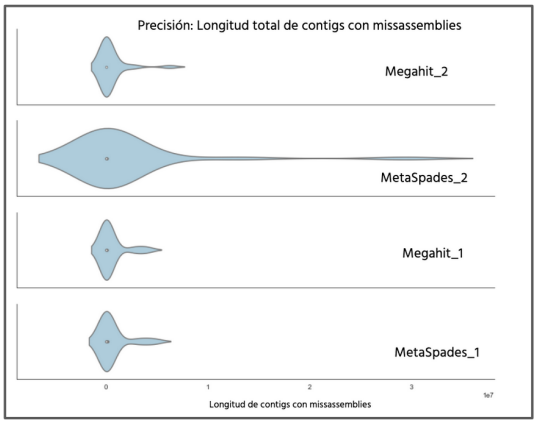

Figura 22: Longitud total de *contigs* con inserciones y/o deleciones.

Las métricas antes mencionadas influyen directamente en el número de anotaciones totales obtenidas, de tal forma que, este conteo final sirve como puntaje de evaluación para conocer la calidad del ensamble. En la siguiente imagen se observa el número total de anotaciones obtenidas en cada uno de los ensambles para los dos flujos de trabajo.

Es importante mencionar que aquellas muestras con un número de anotaciones menores a 1000 fueron eliminadas para los subsecuentes análisis.

Debido a que la metodología del presente proyecto requiere de las anotaciones a nivel funcional, el siguiente paso de evaluación consistió en medir la existencia de diferencias o similitudes en la proporción o abundancia de los genes anotados entre los dos ensambladores utilizados, para el flujo de trabajo número 2, ya que en este flujo de trabajo se obtuvieron globalmente mejores resultados en las propiedades analizadas anteriormente (continuidad, precisión e integridad).

Para lograr dicho objetivo, se realizó un análisis de correlación entre las proporciones de los mismos genes anotados entre la muestras ensambladas con Megahit y con MetaSpades. Las figuras 24 y 25 muestran los mapas de calor de los coeficientes de correlación para las abundancias de genes compartidos de los pares de muestras analizadas (ensam-

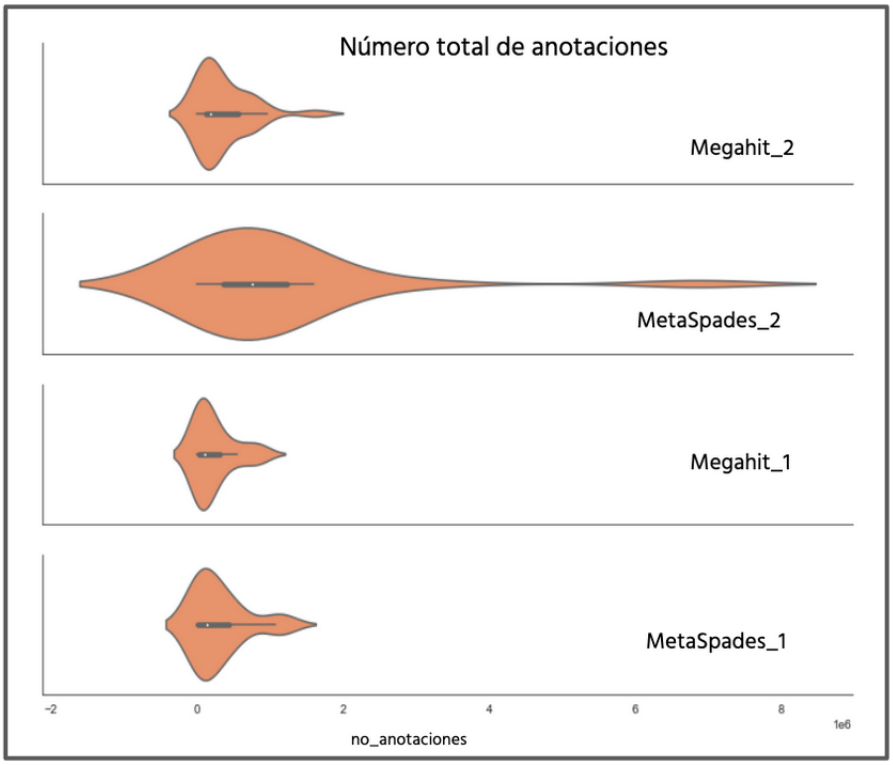

Figura 23: Número total de anotaciones.

ble Megahit vs. ensamble MetaSpades).

Como se observan en las figuras 24 y 25, la abundancia de los genes compartidos que se anotaron con MetaSpades y Megahit, tiene coeficientes de correlación altos en su mayoría y a través de las muestras analizadas.

Ahora bien, con el objetivo de realizar una análisis complementario para la evaluación de los ensambles, se realizó una búsqueda en la literatura de aquellos genes esenciales para la supervivencia en organismos modelo.

Estos genes esenciales tienen sus principales funciones en procesos de traducción, transcripción y replicación molecular, además de procesos de transporte de membrana y conservación de energía [28], [18].

Los mapas de calor de la figuras 26 y 27 muestran la proporción de estos genes esenciales a través de las muestras tanto del grupo de las referencias como para las muestras perturbadas. Es importante resaltar que las muestras del grupo de datos de referencia muestran proporciones de enzimas esenciales conservadas. Por ejemplo, las enzimas 2.7.7.7 y 2.7.7.6 correspondientes a la DNA-polimeasa y RNA-polimerasa, respectivamente, muestran una mayor proporción o abundancia a través de las muestras (ocurre lo mismo en las muestras perturbadas, ver Figura 27) en comparación con el resto de las enzimas esenciales.

Independientemente de este hecho, se puede observar un patrón de conservación en lo referente a la abundancia del resto de las enzimas esenciales. Por ejemplo, la categoría con mayor cantidad de genes o enzimas conservadas es la de el procesos de síntesis de proteínas o traducción, llevada a cabo por las tRNA sintetasas y proteínas ribosomales.

En lo que se refiere a las tRNA sintetasas- 6.1.1.1:Tyrosyl-tRNA sintasa, 6.1.1.10:Methionyl-tRNA sintasa, 6.1.1.11:Seryl-tRNA sintasa, 6.1.1.16:Cysteinyl-tRNA sintasa, 6.1.1.19:Arginyl-tRNA sintasa, 6.1.1.2:Tryptophanyl-tRNA sintasa, 6.1.1.20:Phenylalanyl-tRNA sintasa, 6.1.1.21:Histidyl-tRNA sintasa, 6.1.1.3:Threonyl-tRNA sintasa, 6.1.1.4:Leucyl-tRNA sintasa, 6.1.1.5:Isoleucyl-tRNA sintasa, 6.1.1.7:Alanyl-tRNA sintasa y 6.1.1.9:Valyl-tRNA sintasa, sus proporciones son constantes en tanto las muestras de referencia como en

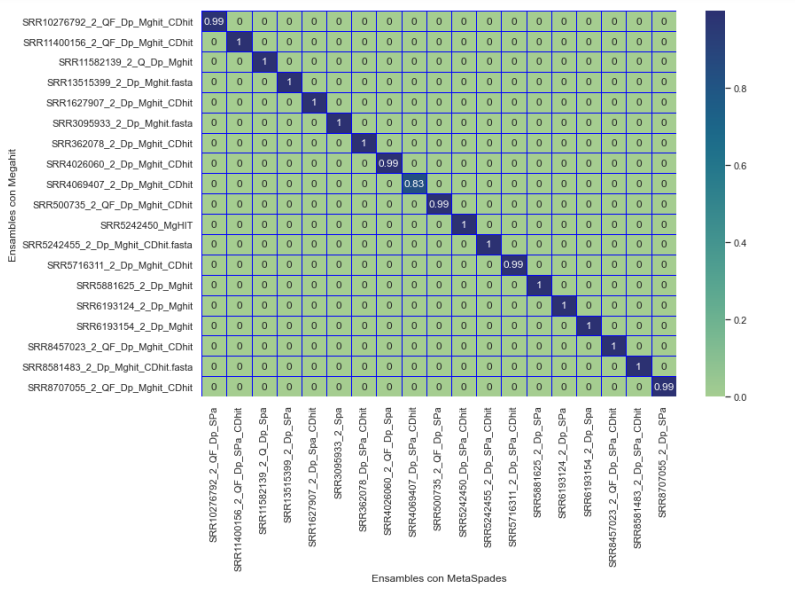

Figura 24: Coeficientes de correlación de las proporciones de genes anotados para las muestras clasificadas como referencias o no perturbadas (NP), analizadas con el flujo de trabajo número 2.

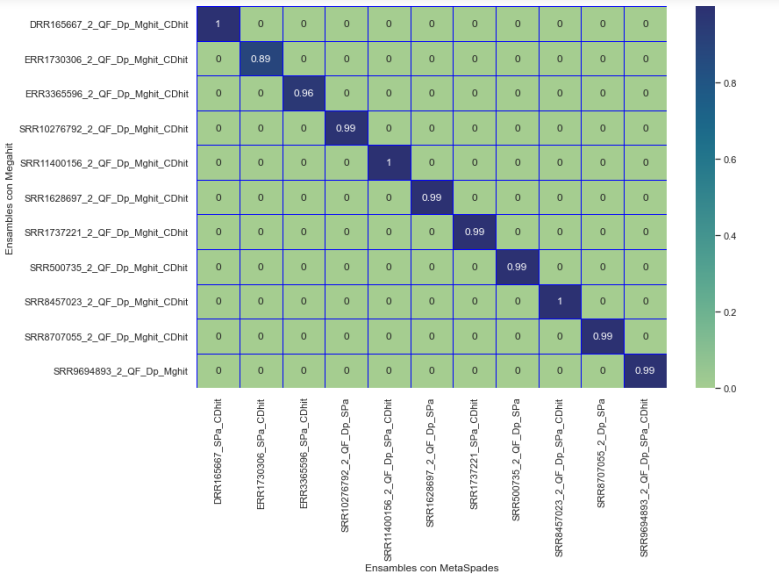

Figura 25: Coeficientes de correlación de las proporciones de genes anotados para las muestras clasificadas como perturbadas y consorcios (P y C), con el flujo de trabajo número 2.

las perturbadas (con excepción de las muestras ERR3365596 y ERR1730306).

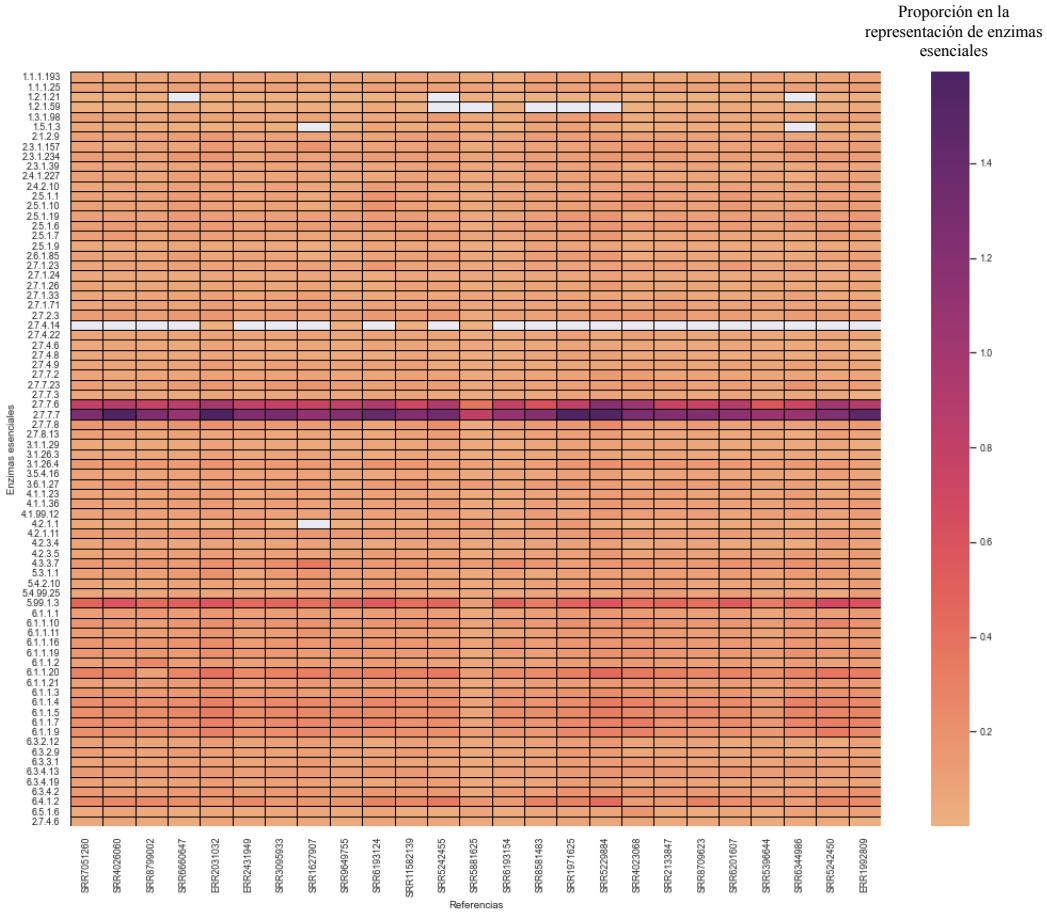

Figura 26: Proporción de genes esenciales a través de las muestras analizadas dentro de la categoría de referencias

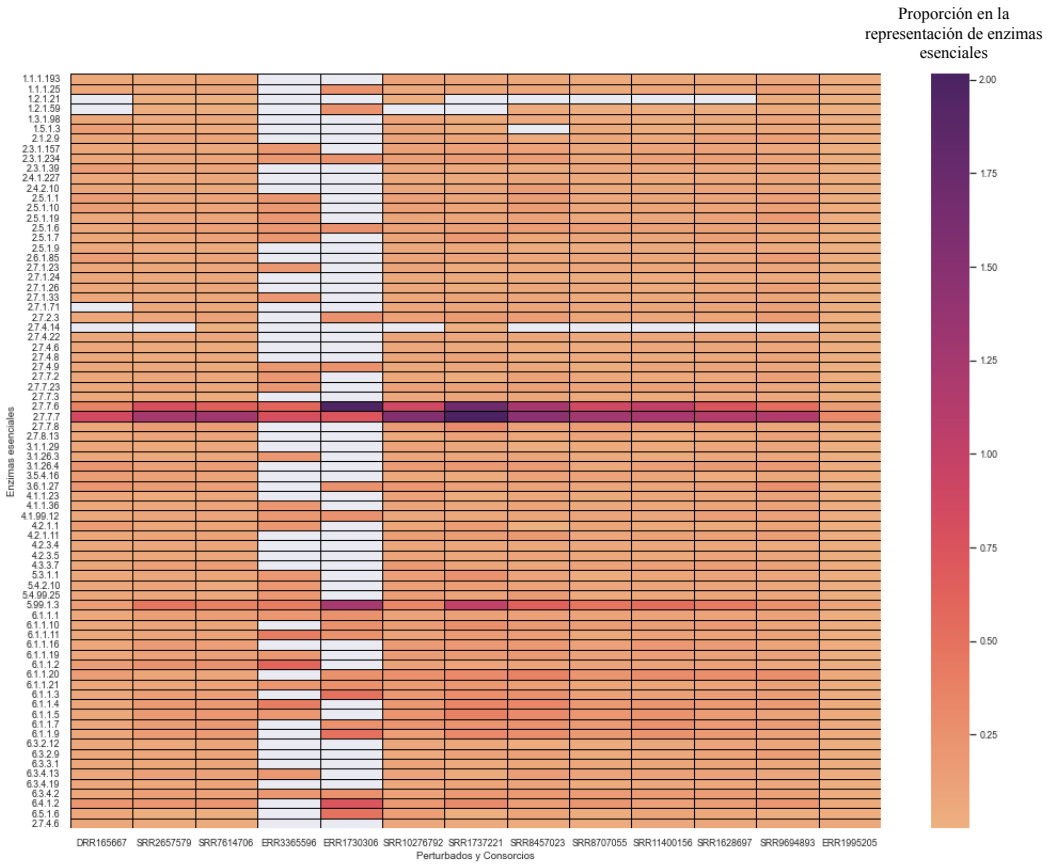

Figura 27: Proporción de genes esenciales a través de las muestras analizadas dentro de la categoría de perturbados y consorcios

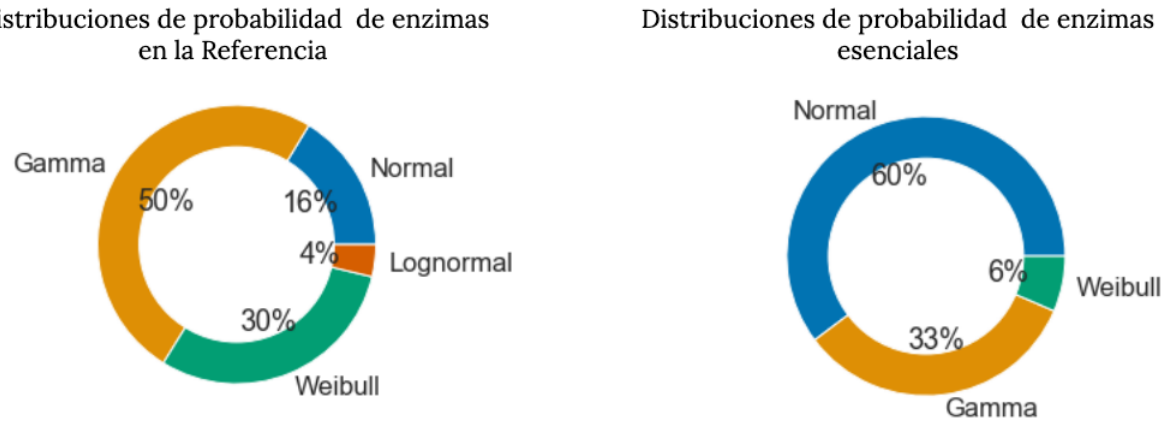

Figura 28: Porcentaje de enzimas por cada distribución de probabilidad asignada.

9.4. Optimización del método estadístico para la definición de perfiles metabólicos

La optimización del método estadístico requirió de la generación de una base de datos de enzimas robusta. Para lo cual, se eligieron 19 muestras sin evidencia de contaminación, a la que se le llamó referencia. Durante este paso de selección de muestras que conforman la referencia se eliminaron aquellas muestras que tuvieran un número de anotaciones por debajo 1000, debido a la dificultad que surge en el cálculo de distribuciones de probabilidad con alto sustento estadístico (índice de Anderson-Darling).

Las 19 muestras de metagenomas de sedimento marino que conforman la base de datos para la referencia se muestra en la Tabla número 2.

Una vez definidas las distribuciones de probabilidad de cada una de las enzimas del grupo de datos de la referencia, las frecuencias relativas de cada enzima en el grupo de datos perturbados fueron utilizadas para ubicarlas en su curva de distribución de probabilidad y asignarles el valor de probabilidad respectivo. Estos valores de probabilidad fueron transformados a puntuaciones z, que más tarde permitieron la reconstrucción de rutas metabólicas. Es importante resaltar el significado asignado a estas puntuaciones

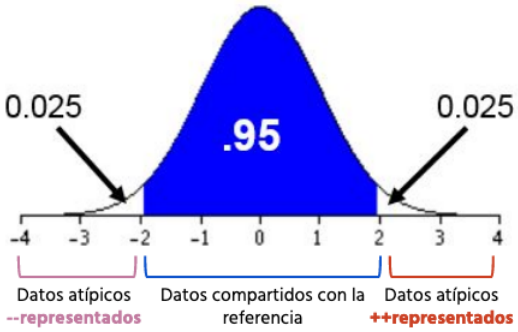

Figura 29: Representación de la asignación de *z-scores* y su interpretación en la curva estándar.

z, para fines de interpretación de los perfiles metabólicos reconstruidos.

Así, como se muestra en la Figura 29, el 97.25 % de los datos se encuentra entre el intervalo que va de -2 y 2, esto significa que valores que caigan dentro de este rango son compartidos con la referencia, al tener mayor probabilidad de caer dentro de éste intervalo. Por otro lado, valores que se encuentran por encima de 2 representan cantidades fuera de lo común, el 0.25 % de los datos se encuentran en este segmento de la curva y por lo tanto son menos probables de encontrarse, a estos valores se les interpreta como cantidades muy representadas en relación con la distribución de probabilidad asignada de la referencia.

Al contrario de lo anterior, aquellos valores que se encuentran por debajo de -2 (el 0.25 % de los datos), son atípicos pero se interpretan como menos representados cuando se realiza la reconstrucción de vías metabólicas.

Por consiguiente, para cada una de las enzimas de la referencia se tiene una curva de distribución de probabilidad asignada, de tal forma que los valores (frecuencia relativa*(1,000,000)) (ver anexo 1. Método estadístico), para cada enzima anotada en las muestras perturbadas, puede ubicarse de acuerdo a su probabilidad y finalmente de acuerdo a su *z-score* asignado en la curva estándar.

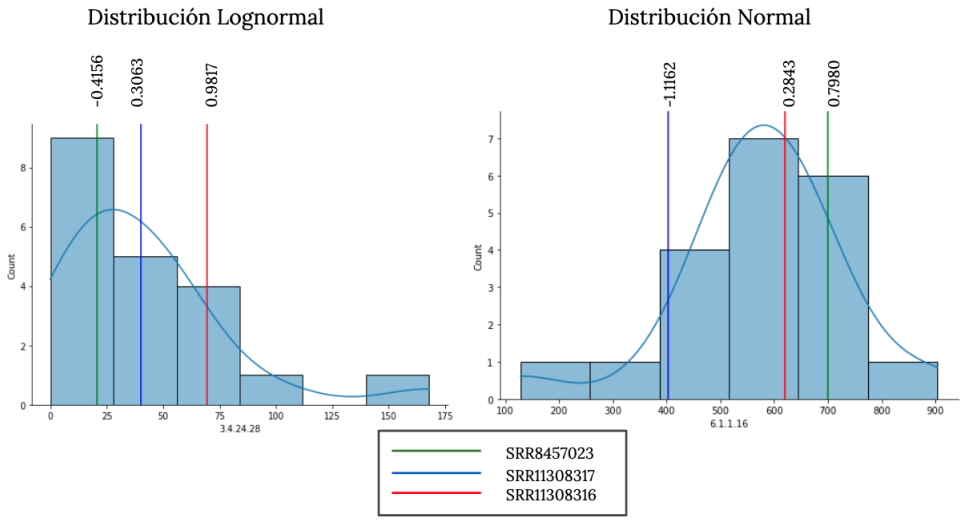

Figura 30: Histogramas de frecuencia con la curva de ajuste de una distribución (izquierda) lognormal y (derecha) normal. Las líneas representan la tasa en la que se encuentran las muestras analizadas y el número encima de cada línea corresponde al z-score calculado.

Las siguientes figuras 30 y 31 muestran los histogramas de frecuencias para enzimas seleccionadas al azar, con el objetivo de representar la curva de ajuste a la distribución de probabilidad de cada una de las distribuciones encontradas, weibull, normal, lognormal y gamma. De está forma una vez definida la distribución se puede hacer el cálculo de el z-score correspondiente, el cual se muestra encima de cada línea en los gráficos, donde las líneas representan la tasa en la que cae la observación problema para cada una de las muestras analizadas.

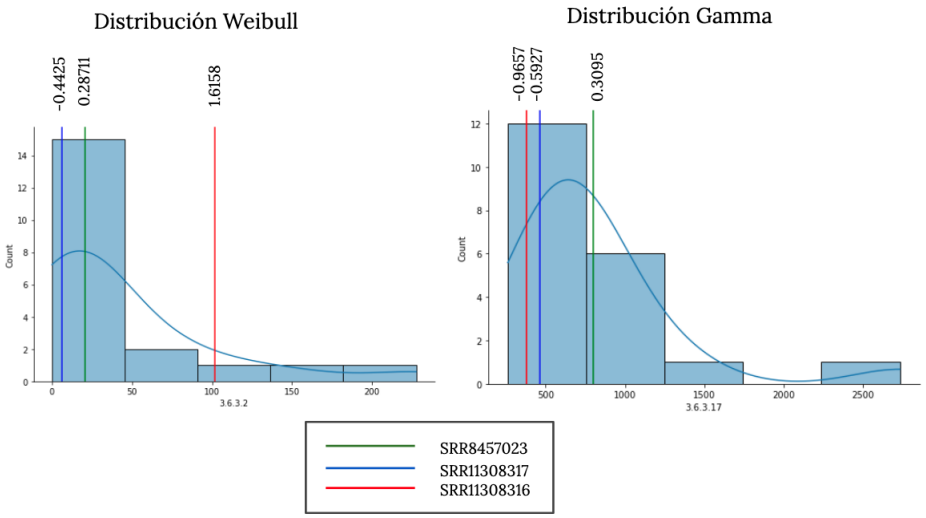

Figura 31: Histogramas de frecuencia con la curva de ajuste de una distribución (izquierda) weibull y (derecha) gamma. Las líneas representan la tasa en la que se encuentran las muestras analizadas y el número encima de cada línea corresponde al z-score calculado.

10. Análisis de vías metabólicas en metagenomas del Golfo de México

En ésta sección se muestra la reconstrucción de las vías metabólicas de las muestras problema, en donde cada enzima tiene relacionado su valor de representatividad o *z-score*. Es importante mencionar que la reconstrucción de las vías metabólicas se realizó con base a los mapas de vías metabólicas de KEGG con la herramienta: KEGG-Mapper – Color y con ayuda de la fuentes bibliográficas específicas para cada vía señalada.

10.1. Degradación de hidrocarburos

Los hidrocarburos son un importante componente de los sedimentos marinos siendo estos, el mayor reservorio de carbono del planeta. El origen de los hidrocarburos puede ser abiótico, derivado de procesos termogénicos en el fondo marino o producto de procesos biológicos. Sin embargo, las actividades humanas también pueden contribuir al aumento en la concentración de hidrocarburos en los mares, como resultado de derrames de petróleo [65].

La degradación de hidrocarburos puede llevarse a cabo en presencia o ausencia de oxígeno, dependiendo de la disponibilidad de este en el ambiente donde se encuentren. En el caso de los sedimentos marinos la degradación de hidrocarburos se lleva a cabo mayoritariamente por medio de vías anaeróbicas, debido a la baja disponibilidad del oxígeno en estos. Los microorganismos que se encargan de la degradación de hidrocarburos en los sedimentos marinos hacen uso de una serie de enzimas que llevan a cabo distintos mecanismos químicos para la ruptura de los enlaces C-H, los cuáles pueden tener energías de disociación que van desde 350 kJ/mol hasta 550 kJ/mol [50].

Entre los mecanismos de activación de enlaces C-H encontramos, 1) hidroxilación con agua, 2) la adición de fumarato, 3) carboxilación, 4) adición de agua en múltiples enlaces y por último 5) la metanogénesis reversa.

Muchas de las enzimas que realizan la activación de hidrocarburos por cualquiera de los mecanismos mencionados anteriormente, han sido ampliamente descritas.

En el caso específico de las muestras analizadas en el presente trabajo, la activación de hidrocarburos por hidroxilación (estudiada en la degradación de etilbenzeno y propilbenzeno por bacterias denitrificantes), adición de fumarato e hidratación de alquenos y alquinos, con el número de enzimas responsables, de acuerdo a la Enzyme Comission (EC): 1.17.99.2, 4.1.99.11, 4.2.1.112, respectivamente, no fueron encontradas, por lo tanto, el mecanismo de activación de hidrocarburos no puede definirse para estos marcadores periféricos con certeza [50].

Una vez que han sido activados los hidrocarburos como alcanos, alquenos, alquinos y aromáticos, su degradación prosigue por una serie de pasos que generan compuestos que se incorporan a la biosíntesis de materia orgánica o su degradación (beta-oxidación).

Para el caso particular de hidrocarburos aromáticos, los cuáles pueden tener un origen natural, como los compuestos lignoaromáticos y ácidos húmicos, aunados a los compuestos derivados de la contaminación antropogénica como el benzoato, fenol, *p*-cresol, anilina, tolueno, entre otros, la vía común de degradación es la del benzoil-CoA, cuyo catabolismo genera compuestos más pequeños como el acetil-CoA, el cual posteriormente puede incorporarse al ciclo de Krebs[21].

La vía del Benzoil-CoA puede darse por dos vías alternas, la primera se ha estudiado en *Thauera aromatica, Azoarcus* sp. y en *Geobacter metallireducens* y sirve como vía clave para identificar la degradación de compuestos aromáticos, en ambientes contaminados [45].

Está vía comienza con la reducción del benzoil-CoA a ciclohexa-1,5-diene-1-carbonil-CoA hasta el intermediario 3-hidroxipimelil-CoA [29], como se observa en la Figura 32, sección d, una de las enzimas clave descritas para la vía, es la 3.7.1.21 o *oah* (6-oxocyclohex-1-ene-1-carbonyl-CoA hidratasa), que para la muestra SRR11308316 el Z-score de la enzima es de 2.218 y para la muestra SRR11308317 es de 0.3699, lo que refleja en la primera muestra SRR11308316, una degradación de compuestos aromáticos por la vía

del Benozoil-CoA. Otra enzima que ha sido descrita como marcador en el catabolismo central de compuestos aromáticos, es la benzoyl-CoA reductasa-1.3.7.8, la cual tiene un z-score para la muestra SRR11308316 de 2.2684, y para la muestra SRR11308317 (ver Figura 36) de 0.5378.

Sin lugar a duda, la muestra SRR11308316 refleja una degradación de compuestos aromáticos por la vía descrita. Por otro lado, la muestra SRR11308317 tiene valores de z-score no sobrerrepresentados, pero sí similares a los valores de referencia para las enzimas clave de la ruta, aunado a los valores iguales a la referencia de la mayoría de las enzimas que conforman la vía del Benzoil-CoA. Como resultado, se puede hipotetizar que ésta vía para la degradación de compuestos aromáticos está en microorganismos presentes en ambas muestras, tomando en consideración, sólo marcadores del catabolismo central de estos compuestos, ya que los marcadores para el metabolismo periférico de activación de hidrocarburos no se hallaron en las muestras analizadas. Esto no descarta la posibilidad de observar degradación de compuestos aromáticos de origen natural como ácidos húmicos y lignoarómaticos, además de hidrocarburos arómaticos.

Existe otra vía de degradación para compuestos aromáticos descrita en *Rhodopseudomonas palustris*, en donde el ciclohexa-1,5-diene-1-carbonil-CoA se reduce hasta pimelil-CoA que después se transforma a 3-hidroxipimelil-CoA [16], sin embargo, en las muestras de sedimento analizadas no se encuentra.

En lo que se refiere a la degradación de alcanos, los rearreglos que sufren estos compuestos en su esqueleto de carbono dan lugar a ácidos grasos que entran a la beta-oxidación generando intermediarios como el propionil-CoA o acetil-CoA, este último intermediario es oxidado hasta dióxido de carbono completando así el ciclo global de carbono en estos ambientes (ver Figura 34: muestra SRR11308316). Como se puede observar en las Figuras 33 y 37 para las muestras SRR11308316 y SRR11308317, respectivamente, las vías de beta-oxidación para el catabolismo de hidrocarburos de cadena lineal se encuentran representadas confirmando así su incorporación al metabolismo central (por beta oxidación).

Este mecanismo ha sido descrito para bacterias reductoras de nitrato y sulfato, ade-

más de bacterias metanogénicas degradadoras de alcanos [63]. Vías que se encuentran representadas en la muestras como se puede apreciar en las Figuras 34,36 (metanogénesis). Siendo la metanogénesis una vía en donde las enzimas muestran valores similares a los de la referencia, confirmando su papel para poder completar la vía global de carbono.

La mayor parte de la mineralización de materia orgánica proveniente de la columna de agua, sucede en los sedimentos anóxicos y es llevada a cabo, por una gran variedad de organismos, desde bacterias desnitrificantes, reductoras de metales, hasta sulfato reductoras y archaeas metanogénicas. La versatilidad de estos organismos de utilizar compuestos como sustratos para la generación de biomasa, depende de la disponibilidad de donadores y aceptores de electrones, así por ejemplo, las bacterias sulfato reductoras no pueden respirar azúcares y tampoco aminoácidos. Para la degradación de estos compuestos en monómeros, existen una serie de organismos fermentadores que los convierten principalmente en ácidos grasos volátiles como el formato, acetato, propionato, butirato, lactato, H_2, CO_2 y alcoholes, utilizables para una segunda etapa de fermentación e incorporación al metabolismo microbiano.

En general, la secuencia de oxidantes utilizados para la mineralización de materia orgánica en sedimentos marinos es:

$$O_2, NO_3^-, Mn(IV), Fe(III), SO_4^2-, CO_2$$

(disminución del potencial redox del oxidante). Se sabe que aproximadamente de 25-50 % del carbono orgánico es mineralizado por bacterias reductoras de sulfato.

La reducción consecutiva de oxidantes a través de la columna del sedimentos se conoce como: *cascada redox*, la importancia de está, es la transferencia de energía del carbono orgánico hasta compuestos inorgánicos, alimentando así las redes tróficas de organismos litótrofos [52].

En lo que se refiere a la degradación de hidrocarburos en condiciones anóxicas, se necesita de estos aceptores de electrones, los cuáles dependen de las condiciones redox y de la disponibilidad de energía para que el flujo de reacciones tenga lugar. A través de la columna del sedimento marino, podemos encontrar zonas que se caracterizan por la

abundancia o ausencia de aceptores y donadores de electrones, además de otras características fisicoquímicas como la temperatura, presión y disponibilidad de oxígeno. En el caso específico de la mineralización de carbono derivado de hidrocarburos a metano y dióxido de carbono, el metabolismo es de tipo sintotrófico, es decir, existen organismos iniciadores en la utilización de ciertos sustratos (*e.g* fermentadores) y otros organismos de la comunidad oxidan los intermediarios generados en las reacciones anteriores (*e.g* metanógenos), manteniendo así bajas concentraciones de compuestos como el hidrógeno.

La biodegradación de hidrocarburos puede darse en condiciones de metanogenia o bien con la presencia de aceptores de electrones como el nitrato, hierro (III), o sulfato. Cuando la biodegradación se realiza en comunidades sintótrofas del sedimento, acoplada a la metanogénesis, existe una producción neta de hidrógeno y acetato, que a su vez puede ser consumida por metanógenos hidrogenotróficos y por metanógenos acetotróficos, respectivamente [17].

Como se observa en la Figura 32 de la muestra SRR11308316, la degradación de compuestos aromáticos que se describió anteriormente, se encuentra en condiciones donde existen aceptores de electrones, como el nitrato (reducción de nitrato) y sulfato (reducción desasimiladora de sulfato), ambas vías de reducción tienen una representatividad similar a la referencia utilizada.

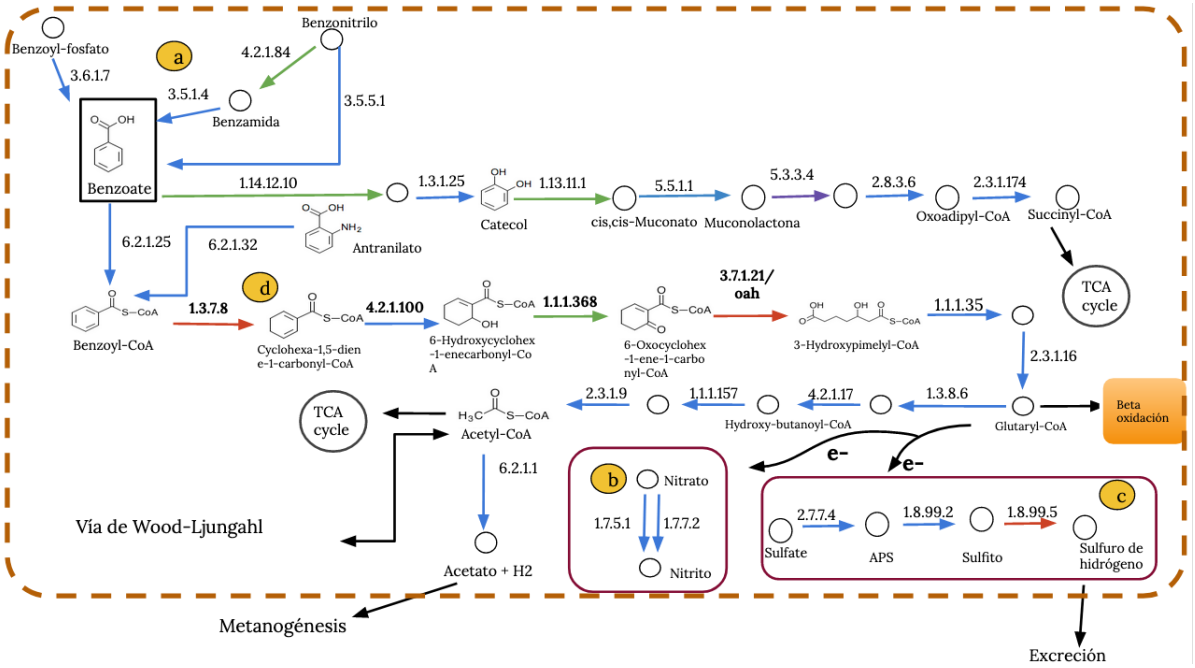

Figura 32: Metabolismo de degradación de hidrocarburos arómaticos en la muestra SRR11308316. a)Incorporación de compuestos aromáticos a la vía del Bezoil-CoA, b) Reducción de nitrato, c) Reducción desasimiladora de sulfato, d)Vía del Benzoil-CoA para degradación de hidrocarburos arómaticos. Código de color de flechas: rosa=z-score debajo de los valores de la referencia, azul=z-score con valores iguales que la referencia, rojo= z-score con valores mayores a los de la referencia, morado= enzima no encontrada, verde= enzima encontrada en la muestra perturbada o de análisis pero no en las muestras de la referencia, negro= conexión con otras vías dentro de metabolismo.

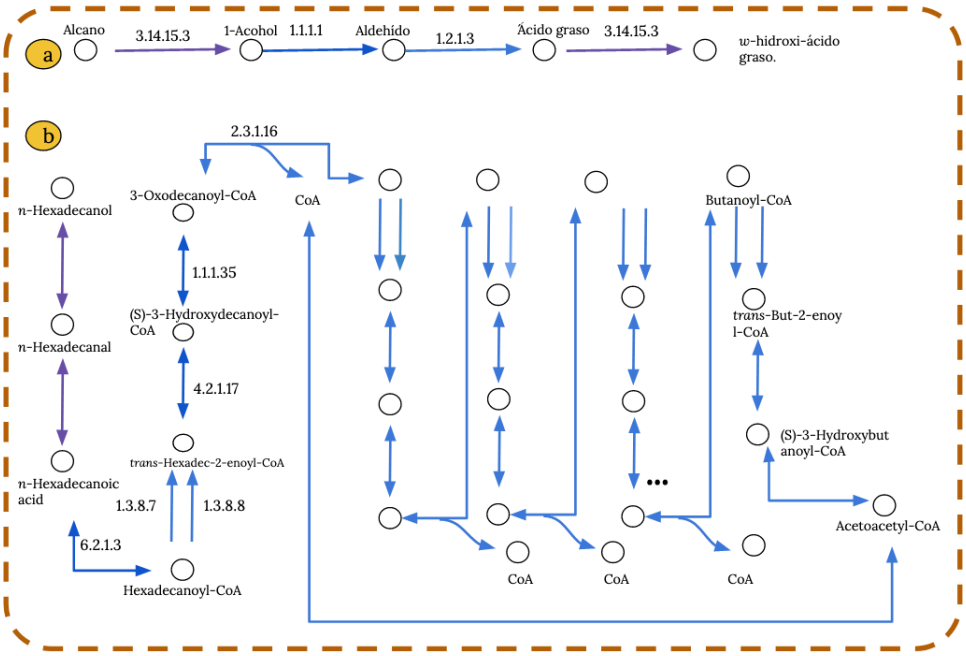

Figura 33: Metabolismo de degradación de alcanos e incorporación de hidrocarburos a la beta-oxidación en la muestra SRR11308316. a) degradación de alcanos, b) beta-oxidación. Código de color de flechas: rosa=z-score debajo de los valores de la referencia, azul=z-score con valores iguales que la referencia, rojo= z-score con valores mayores a los de la referencia, morado= enzima no encontrada, verde= enzima encontrada en la muestra perturbada o de análisis pero no en las muestras de la referencia, negro= conexión con otras vías dentro de metabolismo.

Por otro lado, es posible que la degradación de estos compuestos se encuentre asociada a la metanogénesis acetogénica, generando metano a partir del acetato, como se muestra en la Figura 32 y 36, ya que los valores de z-score para estas enzimas son compartidos con los valores de la referencia en ambas muestras.

Finalmente, la fijación de carbono en la muestra SRR11308316 (ver Figura 32 y 34), se da por el ciclo de Acetil-CoA reductivo o Wood-Ljungdahl. Está vía se divide en dos ramas, la rama del metil y la rama carbonil y tiene lugar principalmente en bacterias y archaeas reductoras de sulfato, además de archaeas metanogénicas. La enzima clave de esta vía es la CO deshidrogenasa/acetil-CoA sintasa-1.2.7.4. Como se puede observar en la imagen de la Figura 34, la monóxido de carbono deshidrogenasa tiene un z-score de 2.6615 y la CO-methiladora-acetyl-CoA sintasa- 2.3.1.169, un z-score de 1.9985, esto refuerza la vía de Wood-Ljungdahl, como principal ruta de fijación de carbono [25] para la muestra analizada, la cuál se puede estar llevando a cabo por bacterias reductoras de sulfato (Figura 31, sección c) y metanogénesis (Figura 34).

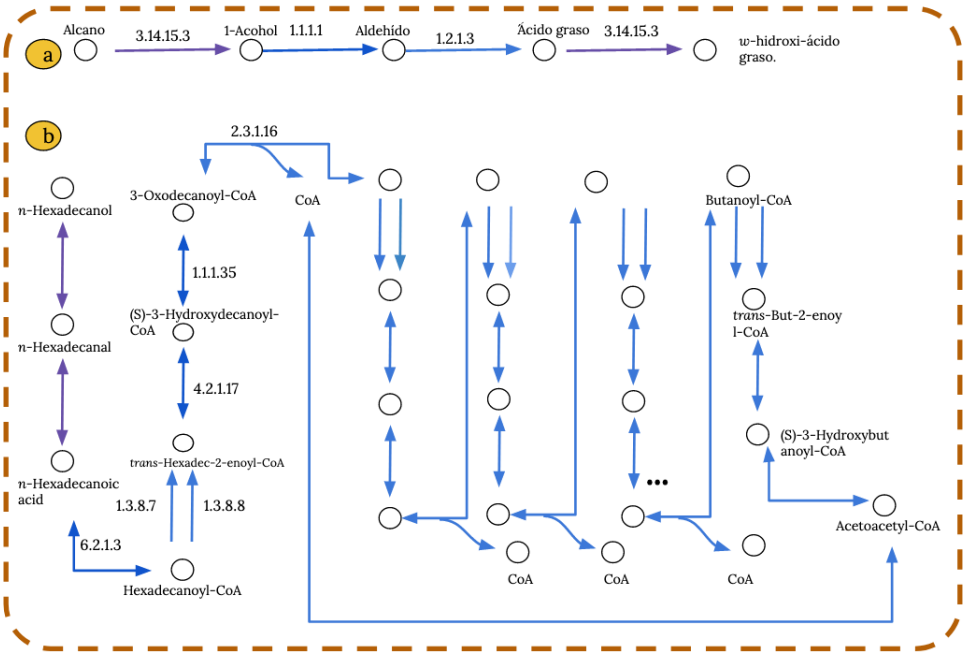

Figura 34: Metabolismo de degradación de alcanos e incorporación de hidrocarburos a la beta-oxidación en la muestra SRR11308316. a) degradación de alcanos, b) beta-oxidación. Código de color de flechas: rosa=z-score debajo de los valores de la referencia, azul=z-score con valores iguales que la referencia, rojo= z-score con valores mayores a los de la referencia, morado= enzima no encontrada, verde= enzima encontrada en la muestra perturbada o de análisis pero no en las muestras de la referencia, negro= conexión con otras vías dentro de metabolismo.

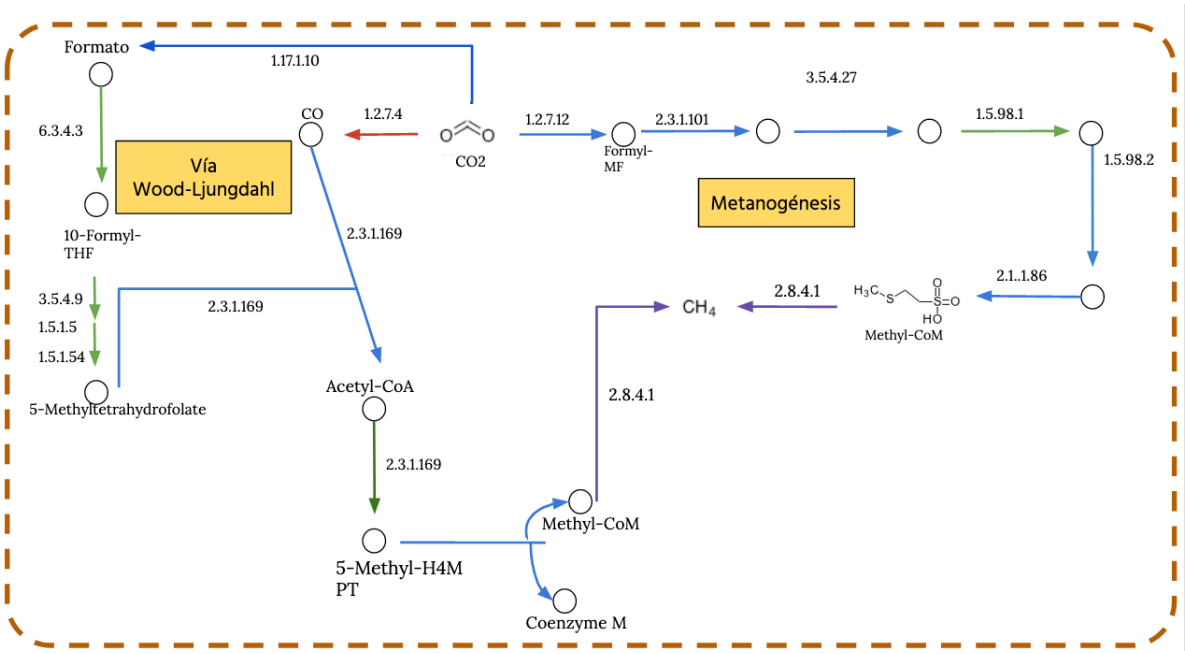

Figura 35: Metanogénesis y fijación de carbono en la muestra SRR11308316. Código de color de flechas: rosa=z-score debajo de los valores de la referencia, azul=z-score con valores iguales que la referencia, rojo= z-score con valores mayores a los de la referencia, morado= enzima no encontrada, verde= enzima encontrada en la muestra perturbada o de análisis pero no en las muestras de la referencia, negro= conexión con otras vías dentro de metabolismo.

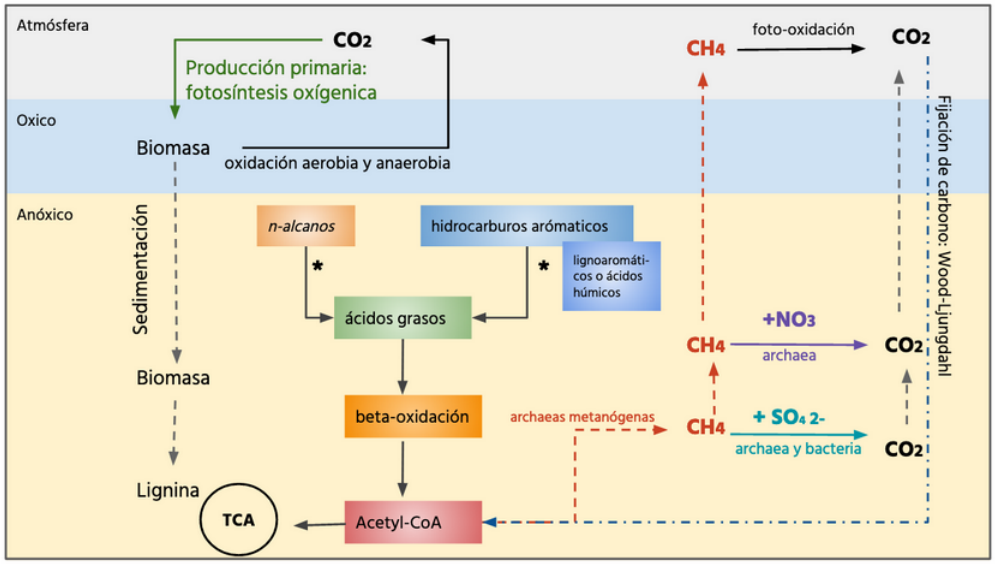

Figura 36: Esquema global del metabolismo en la muestra SRR11308316.

Por otro lado, en la muestra SRR11308317 la fijación de carbono, mostrada en la Figura 37, es llevada a cabo mayoritariamente por el ciclo de ácidos tricarborxilicos oxidativo o rTCA, oxidando dos moléculas de dióxido de carbono a Acetil-CoA. La mayoría de las enzimas del ciclo del ácido cítrico se utilizan en el rTCA, ya que son reacciones reversibles, sin embargo, las enzimas fumarato-reductasa, 2-oxoglutarato sintasa y las enzimas que escinden el citrato son las únicas enzimas del rTCA. La enzima clave del ciclo es la encargada de romper la molécula de citrato a Acetil-CoA y oxalacetato, llamada ATP citrato liasa (CCL), o bien por la acción combinada de la citril-CoA sintasa y la citril-CoA liasa. La enzima clave de la ruta CCL, refleja una sobrerrepresentación en ésta muestra, con un z-score de 2.2397. Las gran mayoría de las enzimas que completan el rTCA tienen valores que se encuentran dentro de los quantiles de referencia, por lo tanto se puede confirmar que la vía de fijación de dióxido de carbono se da por la vía de rTCA [25] para ésta muestra. Otra vía posible de fijación de carbono para la muestra SRR11308317, es la vía Wood-Ljungdahl, pues la mayoría de las enzimas de está vía tienen valores de *z-score*, similares a la referencia, como se muestra en la Figura 39.

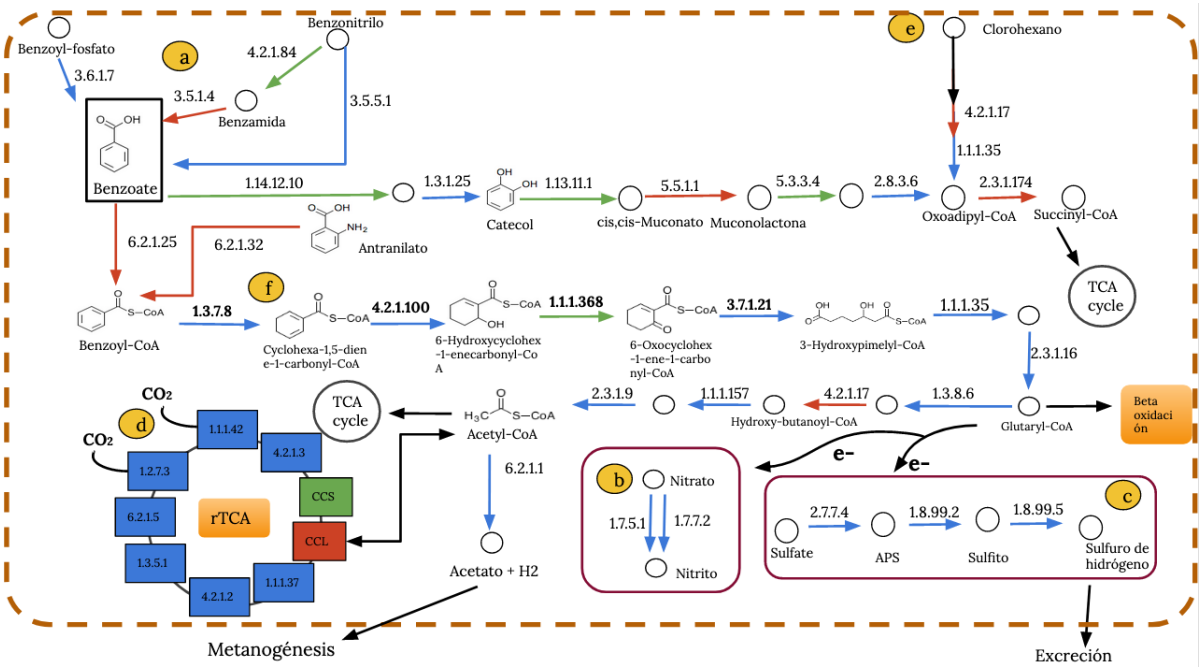

Figura 37: Metabolismo de degradación de hidrocarburos arómaticos en la muestra SRR11308317. a)Incorporación de compuestos aromáticos a la vía del Bezoil-CoA, b) Reducción de nitrato, c) Reducción desasimiladora de sulfato,d) Fijación de carbono por medio del rTCA, f)Vía del Benzoil-CoA para degradación de hidrocarburos aróma-ticos, e) incorporación de clorohexano para su degradación. Código de color de flechas: rosa=z-score debajo de los valores de la referencia, azul=z-score con valores iguales que la referencia, rojo= z-score con valores mayores a los de la referencia, morado= enzima no encontrada, verde= enzima encontrada en la muestra perturbada o de análisis pero no en las muestras de la referencia, negro= conexión con otras vías dentro de metabo-lismo.

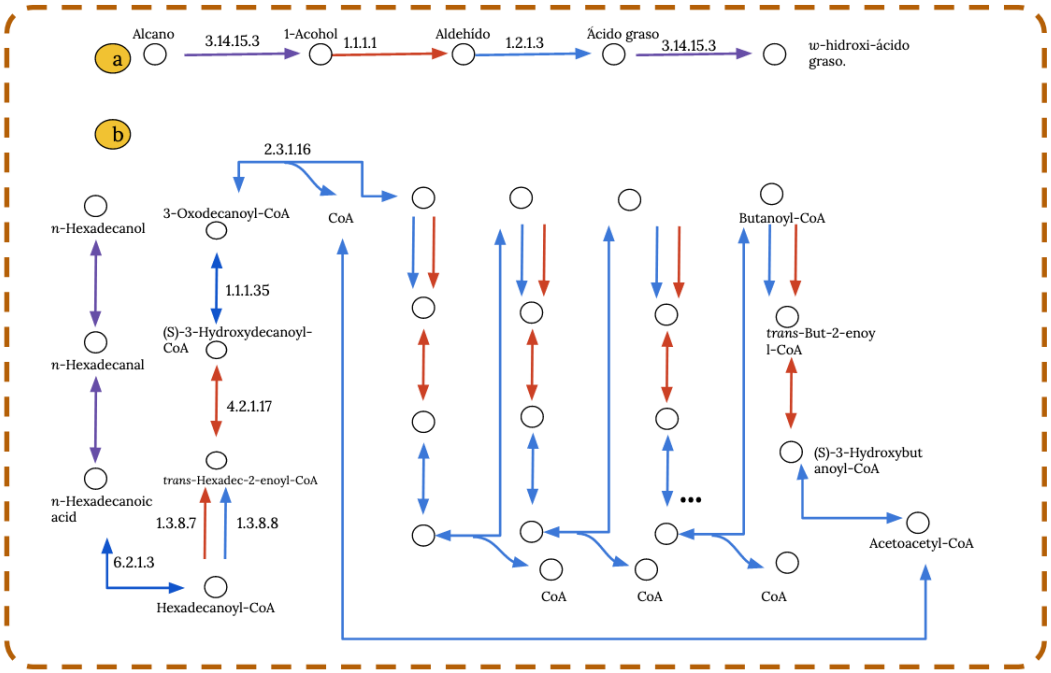

Figura 38: Metabolismo de degradación de alcanos e incorporación de hidrocarburos a la beta-oxidación en la muestra SRR11308317. a) degradación de alcanos, b) beta-oxidación. Código de color de flechas: rosa=z-score debajo de los valores de la referencia, azul=z-score con valores iguales que la referencia, rojo= z-score con valores mayores a los de la referencia, morado= enzima no encontrada, verde= enzima encontrada en la muestra perturbada o de análisis pero no en las muestras de la referencia, negro= conexión con otras vías dentro de metabolismo.

Las Figuras 36 y 40 muestran el esquema global de la reconstrucción metabólica para ambas muestras analizadas.

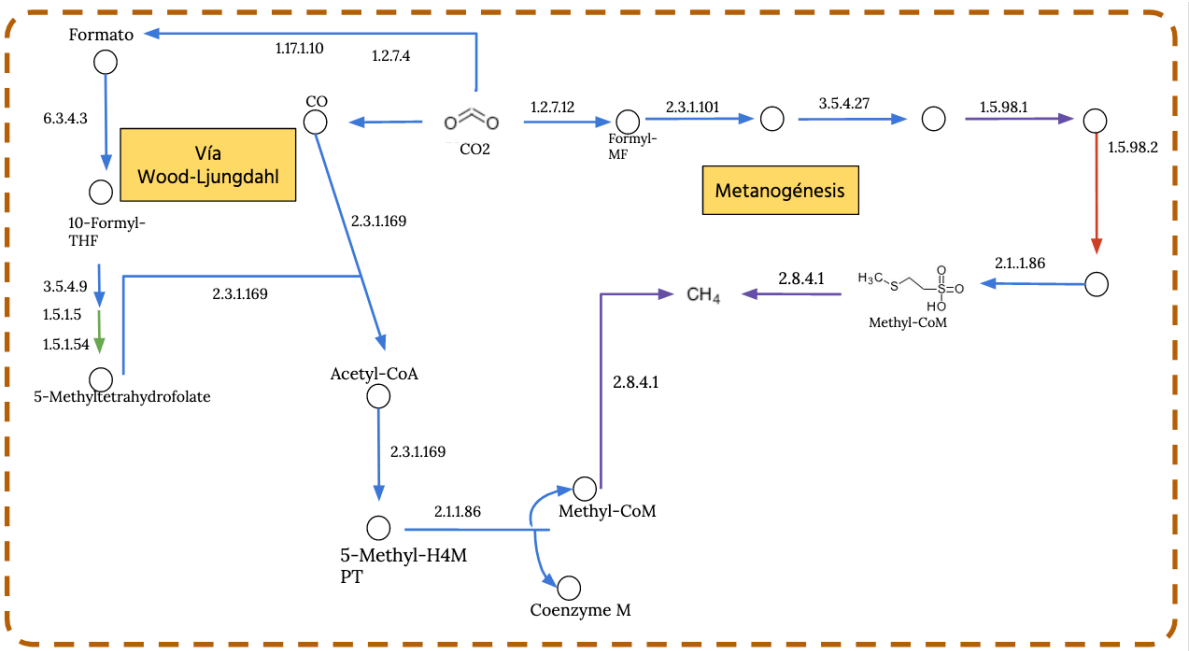

Figura 39: Metanogénesis y fijación de carbono en la muestra SRR11308317. Código de color de flechas: rosa=z-score debajo de los valores de la referencia, azul=z-score con valores iguales que la referencia, rojo= z-score con valores mayores a los de la referencia, morado= enzima no encontrada, verde= enzima encontrada en la muestra perturbada o de análisis pero no en las muestras de la referencia, negro= conexión con otras vías dentro de metabolismo

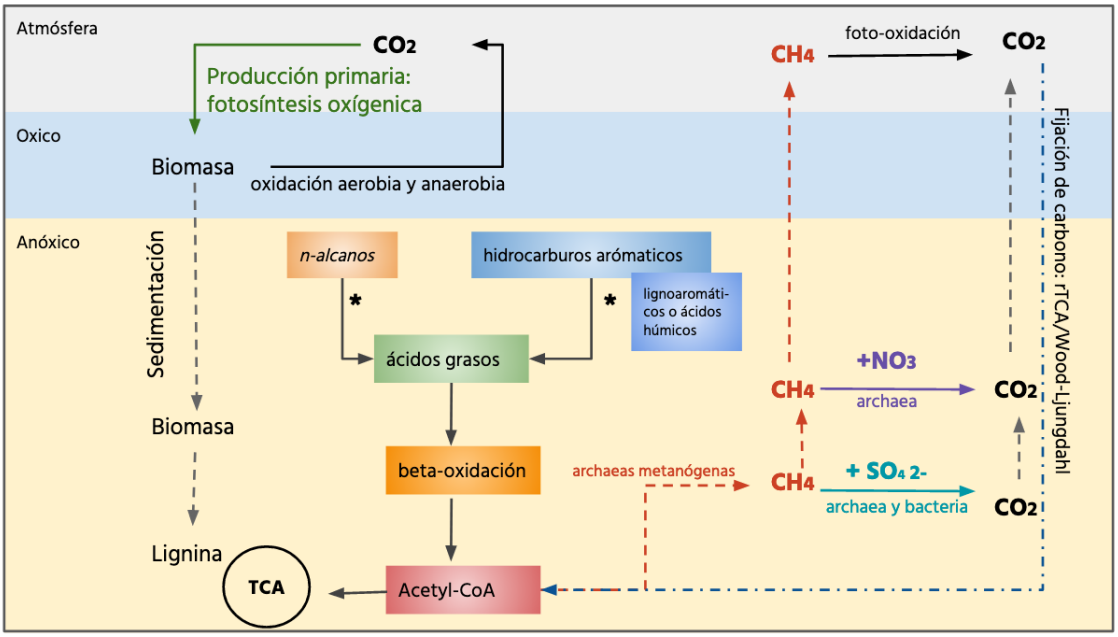

Figura 40: Esquema global del metabolismo en la muestra SRR11308317.

Finalmente, una de las muestras analizadas, fue la SRR8457023. Está muestra es una incubación *in situ* en sedimento marino del Golfo de México. El principal objetivo en la recolección de está muestra (es importante mencionar que no aparece ninguna publicación relacionada a ésta) es la búsqueda de la degradación de hidrocarburos.

En primer lugar, la anotación a nivel taxonómico se realizó para esta muestra con la herramienta Metaphlan v.3.0 y permitió asignar un enriquecimiento de ANME-2cluster-archaea confirmado con la anotación a nivel funcional. La reconstrucción metabólica para ésta muestra se puede observar en la Figura 41, en donde la vía predominante es la oxidación anaeróbica del metano (AOM, por sus siglas en inglés). Para que la AOM sea favorable en términos energéticos, organismos del dominio Archaea, que la llevan a cabo (ANME2c) necesitan de una asociación sintótrofa con bacterias reductoras de sulfato (SRB), las cuáles sirven como pozas de electrones que impulsan las reacciones del AOM [51].

La fijación de carbono en estos organismos se lleva a cabo por la vía de Wood-Ljungdhal,

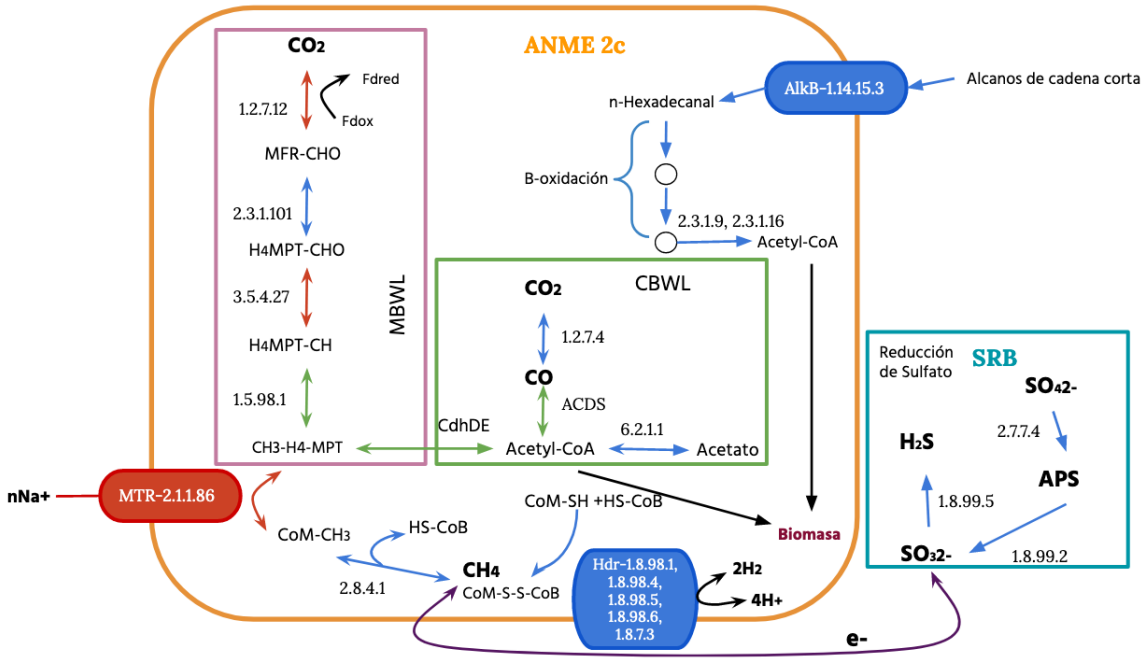

Figura 41: Descripción del metabolismo predominante en la muestra SRR8457023. MBWL- Rama metil de Wood-Ljungdahl, CBML- Rama carbonil de Wood-Ljungdahl. Código de color de flechas: rosa=z-score debajo de los valores de la referencia, azul=z-score con valores iguales que la referencia, rojo= z-score con valores mayores a los de la referencia, morado= enzima no encontrada, verde= enzima encontrada en la muestra perturbada o de análisis pero no en las muestras de la referencia, negro= conexión con otras vías dentro de metabolismo.

en donde tanto la rama carbonil como la metil se encuentran presentes en la muestra (ver Figura 41). Es importante resaltar que, la vía Wood-Ljungdhal no sólo es importante en la fijación de carbono, sino también en la generación de energía, así la enzima 2.1.1.86 o MTR (coenzima metanopterina), al llevar a cabo su función transloca iones de sodio fuera de la membrana creando una fuerza protonmotriz que da lugar a la síntesis de ATP. Como se observa en la Figura 38, está enzima se encuentra sobrerrepresentada con un z-score igual a 3.096.

En paralelo a esto, fue posible confirmar la degradación de alcanos de cadena corta al encontrar la enzima AlkB, con un EC number de 1.14.15.3, la cual se encuentra representada con valores que se encuentran dentro del rango de la referencia con un z-score de -0.019. Aunado a esto la vía de degradación de ácidos grasos o beta-oxidación, también se encuentra representada con valores dentro de los rangos de la referencia y da lugar a Acetil-CoA, la cual se integra a vías biosintetizadoras en el metabolismo microbiano.

Por otro lado, es importante mencionar que está muestra no presentaba metadatos de información con respecto a parámetros fisicoquímicos o de profundidad a lo largo de la columna del sedimento marino, sin embargo, es posible, dado las vías metabólicas representadas en la muestra y la taxonomía encontrada, que el hábitat en el sedimento marino sea la zona de transición de sulfato-metano, debido a que la mayoría del metano producido en el fondo marino es consumido por AOM en dicha región [27].

11. Discusión

Independientemente de la hipótesis biológica a probar o refutar, una de las mayores preocupaciones es el nivel de reproducibilidad en el análisis de datos metagenómicos encontrados en bases de datos. Esto debido a múltiples factores, como son el manejo de las muestras, el tamaño de las mismas, los métodos de extracción, recolección de DNA, así como las metodologías computacionales para el análisis de datos. Inconsistencias o errores en cada uno de estos pasos, así como la falta de estandarización tanto en el laboratorio como para los análisis computacionales, crean sesgos en los resultados y hacen difícil su comparación.

Recientemente, Barthi R. y Grimm D. [6], enlistan las múltiples fuentes de sesgo en el análisis de datos metagenómicos, además de posibles recomendaciones con la finalidad de estandarizar los protocolos de análisis para este tipo de datos.

Tomando esto en consideración, como primer paso en el diseño de éste estudio, la creación de una base de datos (con la herramienta: MySQL) que permitiera el acceso eficiente a través de la estandarización y homogeneización de los datos para cada una de las muestras de sedimento marino, permitió realizar consultas a lo largo del desarrollo del proyecto que agilizaron los análisis subsecuentes.

Aunado a lo anterior, las muestras elegidas para el análisis derivan de la misma plataforma de secuenciación y son bibliotecas pareadas, esto por un lado, intenta eliminar el sesgo ocasionado por los distintos protocolos de secuenciación entre plataformas de secuenciación, *e.g* Illumina vs 454, en donde el tamaño de las secuencias y las tasas de error son muy diferentes (ver Tabla 1). Por otro lado, las bibliotecas pareadas mejoran la resolución de los ensambles y detección de duplicados ópticos. Sin embargo, las variantes (ver Tabla no. 2 y 3) en las metodologías de secuenciación Illumina, generan sesgos que afectan la resolución de los ensambles y la interpretación biológica, además de factores experimentales como la extracción, tratamiento, almacenamiento de DNA y factores propios de las comunidades microbianas en cuestión y sus genomas como secuencias repetidas y/o secuencias no conocidas.

De tal forma que, la realización de dos flujos distintos de trabajo (Figura 7) para el análisis de las muestras permitió en primer lugar y en lo referente a las herramientas de limpieza, notar que los resultados de eliminación de adaptadores, eliminación de nucleótidos de baja calidad y de secuencias de baja complejidad, es similar para ambas herramientas. En la Figura no.11 se muestran ciertas muestras en las que los parámetros utilizados para Trim-Galore no mejoraron significativamente sus calidades y se optó por mantener sus características sin el filtrado de calidades para no afectarlas.

Una característica favorable de la herramienta Fastp en contraste con Trim-Galore es el tiempo computacional utilizado, el cual es menor para el primer software, además este tiene como ventaja el mostrar el porcentaje de secuencias duplicadas o redundantes de las muestras analizadas, lo que permite planear el siguiente paso en el análisis y tomar la decisión de si eliminar secuencias duplicadas o no.

El siguiente paso evaluado es la eliminación de redundancia antes del ensamble. Es importante mencionar que este paso es costoso en términos computacionales, pues requiere de grandes cantidades de memoria para hacer las comparaciones pareadas entre secuencias de ADN. Así, de manera global el flujo de trabajo no. 2 (Figura 7) no sólo necesito de más memoria computacional sino mayor cantidad de tiempo por la eliminación de redundancia [34].

La eliminación de duplicados ópticos y producto de PCR se realizó con la herramienta CD-HIT-DUP. Para este paso no se encontró ninguna asociación entre la plataforma de secuenciación y la cantidad de duplicados de la muestra, por lo tanto, es posible que sea más bien la cantidad total de DNA extraído así como las rondas de PCR realizadas antes de la secuenciación, las que tienen un impacto en el número de secuencias duplicadas encontradas en las muestras.

Por otro lado, se sabe que la diferenciación entre duplicados naturales y producto de PCR es un proceso muy difícil de detectar para los software actuales, en cambio, los duplicados ópticos son más fáciles de distinguirse cuando las bibliotecas de DNA son pareadas, ya que la probabilidad de tener dos secuencias exactamente iguales es muy baja [66].

Uno de los principales objetivos de la comparación de protocolos fue mejorar la resolución de los ensambles de las muestras. Para esto tanto la eliminación de redundancia antes como después del ensamble tuvo efectos favorables.

La manera de validar los ensambles consistió en evaluar tres métricas distintas: la integridad, continuidad y precisión de los ensambles.

La integridad es una métrica que mide la fracción de genomas recuperados en los ensambles, es decir, necesita de una base de datos taxonómica para encontrar y calcular la fracción taxonómica encontrada en las muestras ensambladas. A este respecto, los ensambles de ambos flujos de trabajo mostraron fracciones de genomas similares, por lo tanto, la eliminación de redundancia o duplicados, no tiene un efecto directo en como se ensamblan los genomas y por lo tanto en la búsqueda de marcadores utilizados para la asignación taxonómica.

En términos de continuidad de los ensambles, que se mide por el número total de *contigs*,longitud total del ensamble, así como longitudes de ensambles alineados a la referencia o la longitud del alineamiento más largo alineado a la referencia (Figuras 16-20), a lo cuales se les eliminó la redundancia mejoraron estas métricas significativamente, en especial MetaSpades seguido de Megahit.

Se ha reportado, que la continuidad de los ensambles realizados con MetaSpades tienen mejores resultados con respecto a otros ensambladores, esto debido a la implementación en la construcción de gráficas de *de Bruijn*. MetaSpades, resuelve mejor las secuencias repetidas y busca mantener la mayor cantidad de información durante la construcción de gráficos de *de Bruijn*, para preservar la información de cepas minoritarias y/o variación intraespecie [60]. Esta estrategia, por un lado, genera mayor información a nivel cepa de los ensambles analizados, y por el otro puede generar ensambles más desconectados(sólo cuando no deteriora la información de la gráfica consenso).

Es posible que debido a está implementación del algoritmo los ensambles tengan mayor redundancia en comparación con los ensambles realizados con Megahit, el cual es un software que mantiene sólo *k-meros* de alta cobertura (aparecen más de una vez), para la

construcción de gráficos de *de Brujin*. Las conexiones de las gráficas con menor cobertura (*'mercy k-mers'*), las utiliza para aumentar la continuidad de las regiones con baja cobertura. También elimina aquellos nodos si su cobertura es muy distinta a la de los nodos vecinos, esto resulta en un menor número de errores estructurales.

Como se muestra en los gráficos de barras del Anexo 1, sección Ensamble de muestras (Figura 5 y 6), la redundancia de los ensambles producto de Megahit es bajo en comparación con los ensambles realizados con MetaSpades, la figura 50 del Anexo 1 muestra para un grupo de muestras ensambladas con Megahit que el porcentaje de redundancia eliminada con CD-HIT-EST cae en su mayoría entre 0 y menos de 10.

Otra de característica evaluada es el número total de ORF, los cuáles aumentan en el protocolo no.2 para los ensambles realizados con MetaSpades seguido de Megahit. El hecho de que existan mayor cantidad de ORF, en los ensambles realizados con MetaSpades es resultado de la mayor redundancia observada (ver Anexo 1. Sección: Ensamble de las muestras).

En paralelo a esto, para las muestras de referencia y perturbadas analizadas (Tabla 2 y 3), la profundidad de secuenciación no tienen ningún efecto en el número de ORF detectados en el paso de anotación de las muestras, de hecho la composición taxonómica y en específico la completitud de sus genomas, juega un papel fundamental en el número total de secuencias codificantes encontradas. Por lo tanto, las bases de datos utilizadas para la anotación de genomas o metagenomas tienen una importante influencia en el número real de secuencias anotadas, pues la cobertura de anotación obedece a los sesgos de composición taxonómica encontrada en las bases de datos [32] (ver Figura 38).

Ahora bien, la eliminación de redundancia con CD-HI-EST de los ensambles realizados con MetaSpades y aquellos realizados con Megahit con un número mayor de contigs de 10^5 requirieron de una memoria computacional de más de 80G, lo cual impidió completarlos con los recursos computacionales con los que se contaba.

En consecuencia, se realizó un análisis de correlación para poder contrastar las abundancias de los genes anotados para ambos ensambladores (MetaSpades y Megahit) con

o sin la eliminación de redundancia, según la muestra. Así, es posible concluir que la redundancia encontrada en los ensambles realizados con MetaSpades, no afecta la abundancia con la que se encuentran los genes anotados en comparación con los genes anotados para los ensambles realizados con Megahit.

Otra de las métricas evaluadas, consistió en la precisión del ensamble en donde el número de inserciones y/o deleciones se ve aumentado en las muestras tratadas con el flujo de trabajo no.2, esto debido principalmente a que la continuidad de los ensambles también aumentó. Sin embargo, se ha reportado que MetaSpades en general tiende a tener mayor número de inserciones o deleciones por cada 100 kpb [60].

Ya que no existe un valor específico que permita determinar el nivel ya sea bueno o malo, referente a la precisión del ensamble es recomendable comparar distintos herramientas de ensamble con la finalidad de elegir el ensamblador que mejora la precisión. Al mismo tiempo considerar como un factor de ruido la complejidad de la muestra, es decir, el ambiente de donde proviene, la diversidad propia de la muestra, secuencias repetidas etc, pues afectará directamente está métrica.

Finalmente, un aspecto importante a resaltar en lo referente al rendimiento y velocidad computacional, es que Megahit tiene el mejor rendimiento y tiempos en comparación con MetaSpades [60].

En consecuencia, de acuerdo a los dos flujos de trabajo seguidos durante este trabajo, la elección de herramientas para el análisis de secuencias metagenómicas depende de la naturaleza de los datos, por ejemplo, la cantidad de secuencias duplicadas determina la eliminación de redundancia.

11.1. Optimización del método estadístico para la definición de perfiles funcionales

Como primer control en la optimización del método estadístico propuesto en el presente trabajo se analizaron las abundancias a través de las muestras de referencia y

perturbadas, de genes esenciales de copia única que han sido reportados como fundamentales para el mantenimiento de la integridad celular en organismos modelo [37] [18]. Las abundancias de estos genes a través de las muestras de referencia se mantienen constantes entre muestras (Figura 26), lo cual no sólo demuestra que los genes esenciales para la supervivencia de cualquier bacteria o archaea se encuentran presentes en las muestras metagenómicas analizadas, sino también que existe una constancia en su abundancia esperando así que sus distribuciones de probabilidad sean de tipo normal, esto último se ve confirmado, pues de acuerdo a la Figura 28-Distribuciones de probabilidad de enzimas esenciales, se confirma que el 60 % de las enzimas esenciales siguen una distribución normal, el 33 % se ajustan a una distribución gamma y el 6 % a una distribución weibull.

Este análisis, por otro lado puede estar dando cuenta de la calidad de la secuenciación o extracción y tratamiento del DNA, pues en el caso de las muestras de la categoría de Perturbados (figura 27), existen muestras como la ERR3365596 y ERR1730306, en donde, la mayoría de las enzimas esenciales analizadas no están presentes. Pueden existir múltiples razones que expliquen la falta de enzimas esenciales en estas muestras, una de ellas es la poca profundidad de secuenciación de las mismas o bien, problemas en la extracción de DNA, construcción de bibliotecas de e incluso en la secuenciación, cualquiera de los casos, esto refleja la calidad de las muestras y para fines del proyecto, este tipo de muestras con poca profundidad en la anotación de enzimas esenciales fueron eliminadas.

Los métodos de análisis estadísticos de datos metagenómicos ya sea a nivel taxonómico (OTU's) o funcional (genes, proteínas o enzimas), asumen que la información ecológica es equivalente a la información derivada de las plataformas de secuenciación, representada por el número de conteos de taxa o genes normalizada por una variable constante. Sin embargo, es importante resaltar la existencia de una dependencia entre la capacidad de secuenciación y el número de conteos finales obtenidos por muestra [19]. Así el número total de conteos por muestra están determinados por el número total de secuencias que el secuenciador pueda generar y la abundancia del gene o taxa encontrado, esto hace de los datos metagenómicos datos tipo composicionales [8].

La gran variabilidad existente en los datos composicionales no permite utilizar las abundancias crudas o conteos crudos para hacer comparaciones entre muestras, por lo tanto, se han desarrollado una serie de metodologías que intentan normalizar los datos y así permitir la realización de análisis estadísticos *e.g* trimmed mean method (TMM, por sus siglas en inglés), cumulative sum scaling (CSS), entre otros [8], [42].

Por otro lado, se han desarrollado métodos para el tratamiento específico de datos composicionales que incluyen el *centered-log-ratio* (clr), *additive-log-transformatio*n (alr) e *isometric-log-transformation*(ilr), estas transformaciones permiten trabajar con proporciones entre los componentes, lo que cancela el efecto de los conteos totales por muestra y hace que no se requiera de una normalización como las anteriormente señaladas [2].

Independientemente del método utilizado, ya sea normalización o escalamiento de datos, el objetivo general es eliminar la variación derivada de las diferencias en la profundidad de secuenciación. Sin embargo, es importante tomar en cuenta que cada método asume distintas propiedades de los datos, por lo tanto, su elección debe tomar en consideración la naturaleza de los datos y el objetivo final del investigador, por ejemplo: los métodos de escalamiento asumen que los datos tienen una media o mediana similar entre muestras, la normalización por cuantiles asume que las muestras tienen distribuciones iguales de sus datos y la normalización por TMM no es apropiada para datos con muchos ceros y datasets muy asimétricos[5].

Como se puede observar, este tipo de transformaciones permite hacer comparaciones entre muestras (uno de los principales objetivos en análisis de datos tipo RNA-seq), sin embargo el objetivo del método planteado en este estudio busca generar un dataset de referencia, que permita inferir la probabilidad de observar un dato en un grupo problema (muestras perturbadas). Para este fin se calcularon las funciones de densidad de probabilidad de las enzimas a través de las muestras de referencia con el objetivo de modelar los datos. Este procedimiento cae dentro de la estadística inferencial y permitió acceder a la incertidumbre propia de los datos al estimar parámetros que describan la función de probabilidad permitiendo finalmente hacer predicciones para las observaciones problema [43].

Esto, a diferencia de muchas metodologías permite capturar la distribución de un grupo de datos y al mismo tiempo explicar su variación a través de los parámetros (*e.g* media y desviación estándar) , calculados para cada función de probabilidad de cada una de las enzimas.

Como se ha mencionado anteriormente, la diferencia en la profundidad de secuenciación es un aspecto propio de las muestras, el cálculo de frecuencias relativas permitió eliminar dicho sesgo en los datos tanto de las muestras de referencia como en las muestras perturbadas.

A este respecto, el cálculo de las frecuencias relativas de las enzimas anotadas. tiene como desventaja en el presente método que no toma en cuenta la cobertura de los reads para su asignación, teniendo posibles efectos en la interpretación debido a que no se toma en cuenta la contribución específica de los miembros de la comunidad microbiana analizada.

La metodología planteada en el presente estudio dio lugar a grupos de enzimas con valores representados por z-scores o cuantiles, que se encuentran por encima del 95 % de los datos de la referencia para cada una de las enzimas, y que agrupan enzimas con mayor representatividad en las muestras analizadas como perturbadas. De igual forma, permite encontrar grupos de enzimas con menor representatividad o con una representatividad que cae dentro del 97.25 % de los datos según la referencia y que no sugieren ningún aumento en su abundancia o representación en las muestras analizadas.

Esto significa que el método permite estimar una linea base para cada enzima, que sirve para contrastar a la misma enzima con otros metagenomas problema e identificar enzimas con comportamiento atípico.

Derivado de lo anterior, la elección de las muestras que constituyen la referencia es de fundamental importancia pues las firmas o grupos de enzimas definidos por su *z-score* dependen completamente de la distribución de estos datos, por lo tanto, la recolección de metadatos de cada muestra, así como su análisis debe estar estandarizado para poder realizar conclusiones a nivel biológico.

Sin embargo, las muestras utilizadas como referencia en el presente estudio permitieron diferenciar entre muestras con mayor capacidad para la degradación de compuestos aromáticos y permitieron describir las distintas reacciones metabólicas que tienen lugar en las distintas zonas del sedimento marino.

De acuerdo con Raggi. L, *et.al.*,2020, en las muestras SRR11308317 y SRR11308316 se encontró evidencia de reducción desasimiladora de sulfito y sulfato, metanogénesis y reducción desasimiladora de nitrato, además de una fijación de carbono llevada a cabo por la vía de Wood-Ljungdahl y la degradación de hidrocarburos aramáticos. Estos resultados concuerdan con las vías reconstruidas con el presente método estadístico y además sugieren la vía de reductiva del ciclo de ácidos carboxílicos como vía alternativa en la fijación de carbono [48]

En el caso específico de la muestra SRR8457023, donde no hay evidencia de publicación y tampoco existen metadatos de datos fisicoquímicos disponibles, fue posible reconstruir las vías predominantes que caracterizan a este consorcio de sedimento marino *in situ*. Resultando en una relación sintotrófica entre una bacteria sulfato reductora y una archaea ANME 2c (confirmado por asignación taxonómica) metanotrófica que lleva a cabo la oxidación anóxica del metano.

A pesar que el grupo de muestras de referencia permitió definir el perfil metabólico de las muestras perturbadas anteriormente mencionadas, es importante resaltar que los valores de z-score de las enzimas de las vías descritas para éstas muestras son relativas a la referencia, por lo tanto, una posible desventaja del método propuesto es que la comparación con una línea base de referencia distinta podrá cambiar la red o perfil metabólico obtenido. Adicionalmente, si las muestras a analizar o perturbadas son muy similares a las muestras utilizadas para construir la referencia, muy pocas enzimas se saldrían de la norma o tendrían un comportamiento atípico, lo cual sería poco útil para definir el perfil metabólico característico de la muestra, aunque este resultado nos permitiría encontrar muestras con perfiles similares

Esto confirma los alcances del método en permitir reconstruir reacciones bioquímicas en cualquier tipo de muestra que tenga evidencia de haber sido contaminada o perturbada

y no cuente con información referente a las variables ambientales, además de que permite realizar una inferencia metabólica sin la identificación *a priori* de marcadores de ruta.

Finalmente cabe resaltar que este método puede utilizarse para el análisis de cualquier bioma, y más aún podría utilizarse para establecer semejanzas y diferencias entre perfiles de biomas distintos.

12. Conclusiones

- La correcta y detallada documentación de los metadatos aumentó la reproducibilidad y las interpretaciones de los datos considerablemente, al hacer más fácil el acceso a determinada información referente a las muestras.
- En lo referente a la comparación de flujos de análisis de datos metagenómicos, es importante tomar en consideración el poder computacional con el que se cuenta y las distintas implementaciones de los software a utilizar porque eso determina el tiempo requerido. Así, el flujo de trabajo no.2 permitió mejorar la calidad de los ensambles de las muestras analizadas, pues el eliminar la redundancia de las secuencias antes del ensamble permite tener mejores resultados en los mismos, en especial los obtenidos de Megahit pues no presentaron redundancia. Sin embargo, el tiempo ganado en la limpieza de las bibliotecas para el flujo de trabajo no.2 fue utilizado para la eliminación de redundancia con CD-Hit y esto requirió una gran cantidad de tiempo computacional, por lo tanto la eliminación de redundancia puede representar una desventaja en este sentido.
- El método estadístico planteado, permite modelar vías metabólicas sin ningún sesgo que interfiera en la elección de vías mayor o menormente representadas, pues en primer lugar no utiliza un grupo de marcadores para la reconstrucción de vías y tampoco toma en cuenta la presencia o ausencia de genes, sino que el valor final asignado a cada enzima anotada de las muestras problema (*z-score*) deriva de la distribución de probabilidad de los datos observados y clasificados como referencia, así la inferencia depende directamente de este grupo de datos.

13. Perspectivas

- Desarrollar un método de clusterización de metadatos asociados a las muestras a analizar, con la finalidad de separar las categorías sin ninguna suposición hecha *a priori* para después contrastar esos grupos y su correspondencia con los perfiles metabólicos encontrados.
- Un aspecto importante a evaluar, ya que a nivel funcional no existen diferencia significativas entre ensambladores, es la asignación a nivel taxonómico, con la finalidad principal de encontrar una mejor captura a nivel de cepa o variación intraespecie en los ensambles derivados de MetaSpades en comparación con los ensambles derivados de Megahit.
- Realizar un análisis de redes global de las vías metabólicas representadas en las muestras contaminadas, permitirá capturar la información de aquellas vías con mayor representatividad y la continuidad existente entre las enzimas descritas.
- Automatizar el proceso y publicar el flujo de trabajo para uso público

14. Referencias

Referencias: Artículos

[1] John Aitchison. "The statistical analysis of compositional data". En: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2 (1982), págs. 139-160.

[2] John Aitchison y Juan José Egozcue. "Compositional data analysis: where are we and where should we be heading?" En: *Mathematical Geology* 37.7 (2005), págs. 829-850.

[3] Brett J Baker y Gregory J Dick. "Omic approaches in microbial ecology: charting the unknown". En: *Microbe* 8.9 (2013), págs. 353-359.

[4] Brett J Baker y col. "Diversity, ecology and evolution of Archaea". En: *Nature Microbiology* (2020), págs. 1-14.

[6] Richa Bharti y Dominik G Grimm. "Current challenges and best-practice protocols for microbiome analysis". En: *Briefings in bioinformatics* 22.1 (2021), págs. 178-193.

[8] M Luz Calle. "Statistical analysis of metagenomics data". En: *Genomics & informatics* 17.1 (2019).

[10] Valerie De Anda y col. "MEBS, a software platform to evaluate large (meta) genomic collections according to their metabolic machinery: unraveling the sulfur cycle". En: *GigaScience* 6.11 (2017), gix096.

[11] Valerie De Anda y col. "Understanding the mechanisms behind the response to environmental perturbation in microbial mats: a metagenomic-network based approach". En: *Frontiers in microbiology* 9 (2018), pág. 2606.

[13] Hui Dong y col. "Artificial duplicate reads in sequencing data of 454 Genome Sequencer FLX System". En: *Acta Biochim Biophys Sin* 43.6 (2011), págs. 496-500.

[14] Deviram Garlapati y col. "A review on the applications and recent advances in environmental DNA (eDNA) metagenomics". En: *Reviews in Environmental Science and Bio/Technology* 18.3 (2019), págs. 389-411.

[15] Jay S GHURYE, Victoria CEPEDA-ESPINOZA y Mihai POP. “Metagenomic assembly: Overview, challenges and applications. 2016”. En: *Yale Journal of Biology and Medicine. ISBN* (), págs. 1551-4056.

[16] KATHARINE J Gibson y JANE Gibson. “Potential early intermediates in anaerobic benzoate degradation by Rhodopseudomonas palustris”. En: *Applied and environmental microbiology* 58.2 (1992), págs. 696-698.

[17] Lisa M Gieg, S Jane Fowler y Carolina Berdugo-Clavijo. “Syntrophic biodegradation of hydrocarbon contaminants”. En: *Current opinion in biotechnology* 27 (2014), págs. 21-29.

[18] Rosario Gil y col. “Determination of the core of a minimal bacterial gene set”. En: *Microbiology and Molecular Biology Reviews* 68.3 (2004), págs. 518-537.

[19] Gregory B Gloor y col. “Microbiome datasets are compositional: and this is not optional”. En: *Frontiers in microbiology* 8 (2017), pág. 2224.

[20] Vicente Gomez-Alvarez, Tracy K Teal y Thomas M Schmidt. “Systematic artifacts in metagenomes from complex microbial communities”. En: *The ISME journal* 3.11 (2009), págs. 1314-1317.

[21] Caroline S Harwood y col. “Anaerobic metabolism of aromatic compounds via the benzoyl-CoA pathway”. En: *FEMS Microbiology reviews* 22.5 (1998), págs. 439-458.

[22] Roland Hatzenpichler y col. “Next-generation physiology approaches to study microbiome function at single cell level”. En: *Nature Reviews Microbiology* (2020), págs. 1-16.

[23] Jaime Huerta-Cepas y col. “eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences”. En: *Nucleic acids research* 44.D1 (2016), págs. D286-D293.

[24] Jaime Huerta-Cepas y col. “Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper”. En: *Molecular biology and evolution* 34.8 (2017), págs. 2115-2122.

[25] Michael Hügler y Stefan M Sievert. “Beyond the Calvin cycle: autotrophic carbon fixation in the ocean”. En: *Annual review of marine science* 3 (2011), págs. 261-289.

[26] Ramana M Idury y Michael S Waterman. “A new algorithm for DNA sequence assembly”. En: *Journal of computational biology* 2.2 (1995), págs. 291-306.

[27] Katrin Knittel y Antje Boetius. “Anaerobic oxidation of methane: progress with an unknown process”. En: *Annual review of microbiology* 63 (2009), págs. 311-334.

[28] Eugene V Koonin. “Comparative genomics, minimal gene-sets and the last universal common ancestor”. En: *Nature Reviews Microbiology* 1.2 (2003), págs. 127-136.

[29] Diana Laempe, Martina Jahn y Georg Fuchs. “6-Hydroxycyclohex-1-ene-1-carbonyl-CoA dehydrogenase and 6-oxocyclohex-1-ene-1-carbonyl-CoA hydrolase, enzymes of the benzoyl-CoA pathway of anaerobic aromatic metabolism in the denitrifying bacterium Thauera aromatica”. En: *European journal of biochemistry* 263.2 (1999), págs. 420-429.

[30] Ilaria Laudadio y col. “Next-generation metagenomics: Methodological challenges and opportunities”. En: *Omics: a journal of integrative biology* 23.7 (2019), págs. 327-333.

[31] Dinghua Li y col. “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph”. En: *Bioinformatics* 31.10 (2015), págs. 1674-1676.

[32] Briallen Lobb y col. “An assessment of genome annotation coverage across the bacterial tree of life”. En: *Microbial genomics* 6.3 (2020).

[34] Andrea Manconi y col. “Removing duplicate reads using graphics processing units”. En: *BMC bioinformatics* 17.12 (2016), págs. 59-71.

[35] Margaret McFall-Ngai y col. “Animals in a bacterial world, a new imperative for the life sciences”. En: *Proceedings of the National Academy of Sciences* 110.9 (2013), págs. 3229-3236.

[36] Paul J McMurdie y Susan Holmes. “Waste not, want not: why rarefying microbiome data is inadmissible”. En: *PLoS computational biology* 10.4 (2014), e1003531.

[37] Seong-In Na y col. “UBCG: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction”. En: *The Journal of Microbiology* 56.4 (2018), págs. 281-285.

[38] Stephen Nayfach y col. “An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography”. En: *Genome research* 26.11 (2016), págs. 1612-1625.

[39] Sergey Nurk y col. “metaSPAdes: a new versatile metagenomic assembler”. En: *Genome research* 27.5 (2017), págs. 824-834.

[40] Nathan D Olson y col. “Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes”. En: *Briefings in bioinformatics* 20.4 (2019), págs. 1140-1150.

[41] R John Parkes y col. “A review of prokaryotic populations and processes in subseafloor sediments, including biosphere: geosphere interactions”. En: *Marine Geology* 352 (2014), págs. 409-425.

[42] Ana Elena Pérez-Cobas, Laura Gomez-Valero y Carmen Buchrieser. “Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses”. En: *Microbial Genomics* 6.8 (2020).

[43] Alexander Petersen, Chao Zhang y Piotr Kokoszka. “Modeling Probability Density Functions as Data Objects”. En: *Econometrics and Statistics* (2021).

[44] Caitlin Petro y col. “Microbial community assembly in marine sediments”. En: *Aquatic Microbial Ecology* 79.3 (2017), págs. 177-195.

[45] Abigail W Porter y Lily Y Young. “Benzoyl-CoA, a universal biomarker for anaerobic degradation of aromatic compounds”. En: *Advances in applied microbiology* 88 (2014), págs. 167-203.

[47] Christopher Quince y col. “Shotgun metagenomics, from sampling to analysis”. En: *Nature biotechnology* 35.9 (2017), págs. 833-844.

[48] Luciana Raggi y col. “Metagenomic profiling and microbial metabolic potential of perdido fold belt (NW) and campeche knolls (SE) in the Gulf of Mexico”. En: *Frontiers in microbiology* 11 (2020), pág. 1825.

[49] Raffaella Rizzi y col. “Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era”. En: *Quantitative Biology* (2019), págs. 1-15.

[51] Silvan Scheller, Ulrich Ermler y Seigo Shima. “Catabolic pathways and enzymes involved in anaerobic methane oxidation”. En: *Anaerobic Utilization of Hydrocarbons, Oils, and Lipids* (2020), págs. 31-59.

[53] Nicola Segata y col. “Metagenomic biomarker discovery and explanation”. En: *Genome biology* 12.6 (2011), págs. 1-18.

[55] Piotr Starnawski y col. “Microbial community assembly and evolution in subseafloor sediment”. En: *Proceedings of the National Academy of Sciences* 114.11 (2017), págs. 2940-2945.

[56] Matthew CB Tsilimigras y Anthony A Fodor. “Compositional data analysis of the microbiome: fundamentals, tools, and challenges”. En: *Annals of epidemiology* 26.5 (2016), págs. 330-335.

[57] Sagar M Utturkar y col. “Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences”. En: *Bioinformatics* 30.19 (2014), págs. 2709-2716.

[58] Joseph J Vallino y Christopher K Algar. “The thermodynamics of marine biogeochemical cycles: Lotka revisited”. En: *Annual review of marine science* 8 (2016), págs. 333-356.

[59] John Vollmers, Sandra Wiegand y Anne-Kristin Kaster. “Comparing and evaluating metagenome assembly tools from a microbiologist’s perspective-not only size matters!” En: *PloS one* 12.1 (2017), e0169662.

[60] Ziye Wang y col. “Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences”. En: *Briefings in bioinformatics* 21.3 (2020), págs. 777-790.

[61] Aaron Weimann y col. “From genomes to phenotypes: Traitar, the microbial trait analyzer”. En: *MSystems* 1.6 (2016).

[62] James Robert White, Niranjan Nagarajan y Mihai Pop. “Statistical methods for detecting differentially abundant features in clinical metagenomic samples”. En: *PLoS Comput Biol* 5.4 (2009), e1000352.

[63] Nur Nadhirah Zakaria y col. “Oil bioremediation in the marine environment of Antarctica: A review and bibliometric keyword cluster analysis”. En: *Microorganisms* 9.2 (2021), pág. 419.

[64] Peng Zhai y col. “MetaComp: comprehensive analysis software for comparative metaomics including comparative metagenomics”. En: *BMC bioinformatics* 18.1 (2017), pág. 434.

[65] Chuwen Zhang y col. “Marine sediments harbor diverse archaea and bacteria with the potential for anaerobic hydrocarbon degradation via fumarate addition”. En: *FEMS Microbiology Ecology* 97.5 (2021), fiab045.

[66] Xiaofan Zhou y Antonis Rokas. “Prevention, diagnosis and treatment of high-throughput sequencing data pathologies”. En: *Molecular ecology* 23.7 (2014), págs. 1679-1700.

Referencias: Libros

[5] David J Balding, Martin Bishop y Chris Cannings. *Handbook of statistical genetics*. John Wiley & Sons, 2008.

[7] Christoph Bleidorn. *Phylogenomics*. Springer, 2017, págs. 173-193.

[9] Trevor C Charles, Mark R Liles y Angela Sessitsch. *Functional metagenomics: Tools and applications*. Springer, 2017.

[12] Gregory Dick. *Genomic approaches in earth and environmental sciences*. Wiley Online Library, 2019.

[33] Michael T Madigan, John M Martinko, Jack Parker y col. *Brock biology of microorganisms*. Vol. 11. Prentice hall Upper Saddle River, NJ, 1997.

[46] Jerry P Queen, Gerry P Quinn y Michael J Keough. *Experimental design and data analysis for biologists*. Cambridge university press, 2002.

[50] Fernando Rojo. *Aerobic utilization of hydrocarbons, oils, and lipids*. Springer, 2019.

[52] Horst D Schulz y Matthias Zabel. *Marine geochemistry*. Vol. 2. Springer, 2006.

[54] Joao C Setubal, Jens Stoye y Peter F Stadler. *Comparative Genomics*. Springer, 2018.

Anexo 1: Procesamiento de muestras metagenómicas de sedimentos marinos

María del Carmen Sánchez Olmos

Febrero 2020 - Diciembre 2022

Índice de figuras

Índice de cuadros

15. Anexo 1:Limpieza de *reads*

Todas las muestras de metagenomas shotgun de sedimento marino, son producto de secuenciación con plataformas de Illumina y sus múltiples variantes (Illumina Hiseq, Illumina Miseq, etc.). Las bibliotecas de DNA son pareadas y tanto la información referente a los metadatos de las muestras como las secuencias de DNA se obtuvieron de la base de datos SRA(Sequence Read-Archive).

Para la etapa de recolección de secuencias de cada una de las muestras, se utilizó la herramienta sra-toolkit, específicamente fastq-dump, con el siguiente comando:

```
$ fastq-dump –split-files SRA_ID
```

Para conocer la calidad de cada uno de los *reads* de las muestras, se utilizó el programa FASTQC v.3. La visualización de la calidad en de las secuencias tiene como finalidad remover secuencias de baja complejidad (*e.g* segmentos no resueltos por el secuenciador NNNNN, remover secuencias de adaptadores, secuencias con baja calidad (puntuación Phred), secuencias sobrerrepresentadas, etc.

Las herramientas utilizadas para la limpieza de las secuencias fueron, para el primer flujo de trabajo: Trim Galore v.3 y para el segundo flujo de procesamiento: Fastp 0.21.0

El comando general utilizado para Trim Galore, fue el siguiente:

```
$ trim_galore fastqc paired retain_unpaired clip_R1 INT clip_R2 INT
three_prime_clip_R1 INT three_prime_clip_R2 INT SRA_ID_1.fastq
SRA_ID_2.fastq
```

En la siguiente tabla se muestran los parámetros establecidos para cada una de las muestras analizadas, así como la longitud, el tamaño del archivo y el número de secuencias totales, después de realizada la limpieza de calidades.

Cuadro 4: Características de las muestras después de la limpieza con Trim-Galore

Características de las muestras después de la limpieza con Trim Galore					
SRA_ID	Parámetros en Trim Galore	% GC	Longitud de *reads*	Total de secuencias	Tamaño del archivo
SRR10276792	10,10,5,5	54	20-136	1918301	569M
SRR8452062	15,15,2,2	44	20-233	2670986	1.3G
SRR2133847	15,15,5,5	43	20-80	65316491	14G
SRR5242450	15,15,10,10	49	20-276	247962	115M
SRR6201607	15,15,10,10	59	20-125	3139806	707M
SRR8799002	15,15,5,5	55	20-134	18052620	5.1G
SRR5716311	10,10,2,2	47	20-89	9858812	2.0G
SRR4069407	10,10,5,5	50	20-98	1952934	438M
SRR6664618	10,10,2,2	45	20-238	1207890	413M
SRR6344986	15,15,2,2	37	20-134	7891137	2.2G
SRR5881625	15,15,2,2	50	20-134	25655580	7.3G
ERR1811646	15,15,5,5	62	18-131	7346534	2.3G
SRR3620787	15,15,2,2	55	20-84	57725243	11G
SRR1737221	15,15,2,2	51	20-83	5020687	972M
SRR7051260	10,10,5,5	46	20-136	25385828	7.4G
SRR1971625	NT	37	108	15402223	5.2G
SRR5189787	NT	52	113	16518093	4.7G
SRR11343834	10,10,2,2	52	20-140	100134291	30G
ERR1730306	15,15,2,2	53	20-59	2254707	399M
SRR8457023	15,15,8,8	44	20-274	1440151	355M
SRR9694893	8,8,1,1	62	20-92	24124661	5.1G
DRR163688	15,15,2,2	51	20-184	4887588	1.3G
ERR906879	10,10,5,5	53	20-286	6461504	3.7G

ERR1992809	NT	50	100	12500000	3.8G
SRR11400156	15,15,1,1	46	20-109	25924150	7.0G
ERR1664711	15,15,5,5	43	20-93	3942219	971M
SRR7614706	NT	44	100	89509974	27G
ERR3771514	15,15,5,5	46	23-208	112441	48M
DRR165667	15,20,5,5	62	20-276	1056594	573M
SRR9650801	10,10,5,5	54	20-141	95672063	29G
SRR6193154	10,10,5,5	52	20-135	13970129	3.4G
ERR3365596	10,10,1,1	43	20-140	8983006	2.9G
SRR9649755	NT	53	100-251	21308190	9.0G
SRR1628697	5,5,2,2	51	10-101	20838411	5.3G
SRR8581483	15,15,5,5	55	20-131	6338879	1.8G
SRR8707055	10,10,5,5	56	20-136	13433703	3.7G
SRR4026060	10,10,1,1	49	20-103	10566474	2.3G
SRR2657579	5,5,1,1	51	20-95	64702119	16G
SRR5242455	10,10,5,5	50	20-286	954176	489M
SRR6193124	10,10,2,2	58	20-138	954176	489M
SRR8709623	10,10,2,2	54	20-138	23589982	7.7G
SRR6660647	5,5,1,1	49	20-95	27710507	6.8G
ERR1995205	15,15,5,5	52	18-131	13354940	4.0G
SRR3095933	15,15,2,2	45	20-233	25597048	11G
SRR6208352	10,10,2,2	48	20-138	174151156	46G
SRR5229884	10,10,5,5	37	20-234	18841269	8.9G
SRR10072859	8,8,1,1	53	20-242	76354800	37G
ERR2431949	8,8,1,1	54	20-142	47353259	14G
SRR11049210	NT	45	151	103759302	37G
SRR500735	10,10,3,3	43	20-112	27424601	5.4G

NT:No trimming- No limpieza. Los parámetros corresponden al comando mostrado

anteriormente.

Por otro lado, para el flujo de trabajo no. 2, el análisis de calidades se realizó con el software Fastp, con los mismos parámetros establecidos para cada una de las muestras. En el siguiente listado se muestran con mayor detalle cuáles fueron los parámetros elegidos para la limpieza de las muestras:

- -w, number of threads.
- -q –qualified_quality_phred = 15
- -x –poly_x_min_len = 40
- -g –poly_g_min_len (detect poly G in read tail)= 10
- -l (reads shorter that the length establish will be discarded)= 40
- -n –n_base_limit (if one read has more N's number will be discarded) = 15
- -P (one in overrepresentation reads will be computed for overrepresentation analysis) = 20
- -p (enable overrepresented sequence analysis)
- -y (low complexity filter)
- -Y (threshold for low complexity) = 30
- -t (trim trail 1) = 5
- -T (trim trail 2) = 5

15.1. Calidades de las secuencias

En la siguiente sección se muestran imágenes que representan las calidades por nucleótido de las secuencias, contenido de GC, la representatividad de bases al inicio y final de las secuencias o *reads*, así como la presencia de secuencias adaptadoras en un grupo seleccionado de muestras procesadas con Trim Galore.

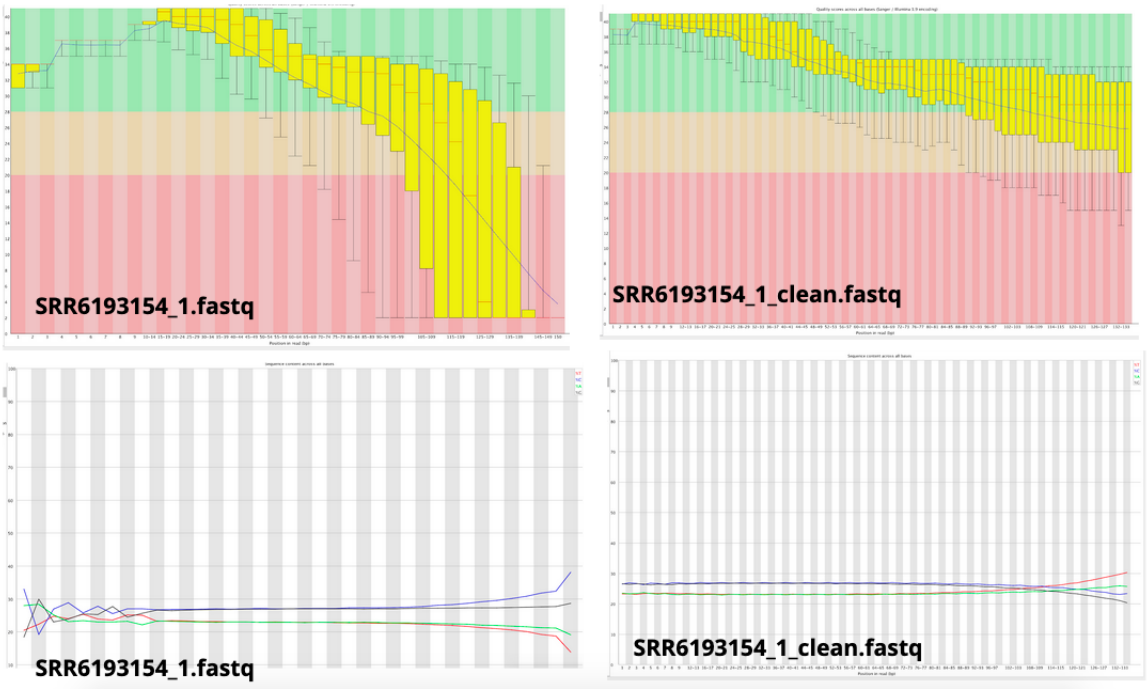

Figura 42: Calidad de la muestra SRR6119354. En la sección superior de la figura se muestran las calidades por nucleótido a través de la secuencia de DNA. En la parte inferior se observa la representatividad por nucleótido a lo largo de la secuencia antes y después de la limpieza con (Trim-Galore).

Las siguientes gráficas son ejemplos representativos de la calidad antes y después del análisis y filtrado por calidades, de las muestras analizadas con Fastp.

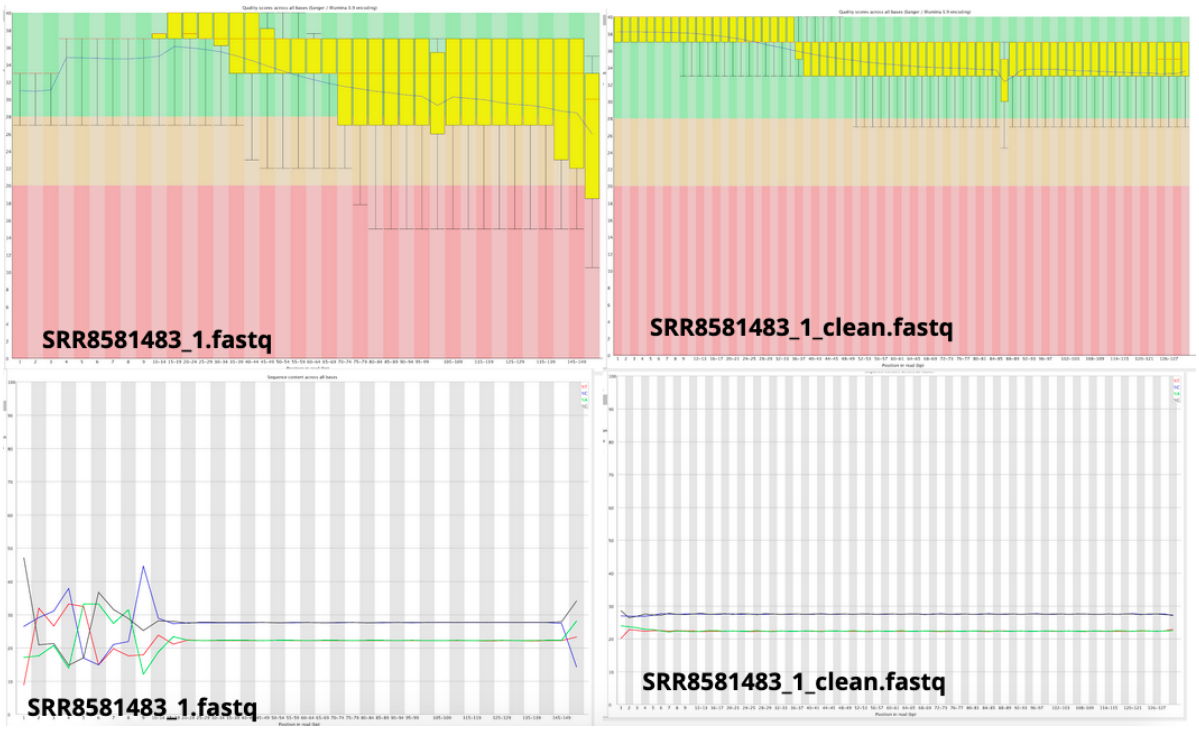

Figura 43: Calidad de la muestra SRR8581483. En la sección superior se muestra la calidad por nucleótido a través de la secuencia de DNA antes y después de la limpieza. En la sección inferior se puede observar la representatividad por nucleótido a través de la secuencia.

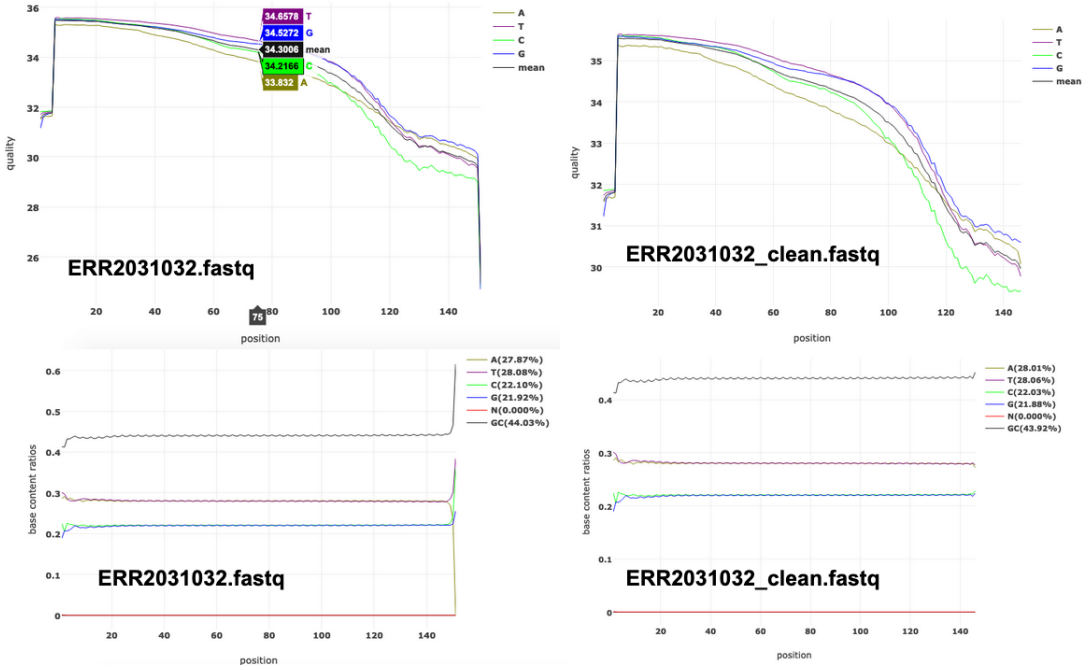

Figura 44: Calidad de la muestra ERR2031032. En la sección superior se muestra la calidad por nucleótido a través de la secuencia de DNA antes y después de la limpieza, exclusivamente para el read forward. En la sección inferior se puede observar la representatividad por nucleótido a través del *read*.

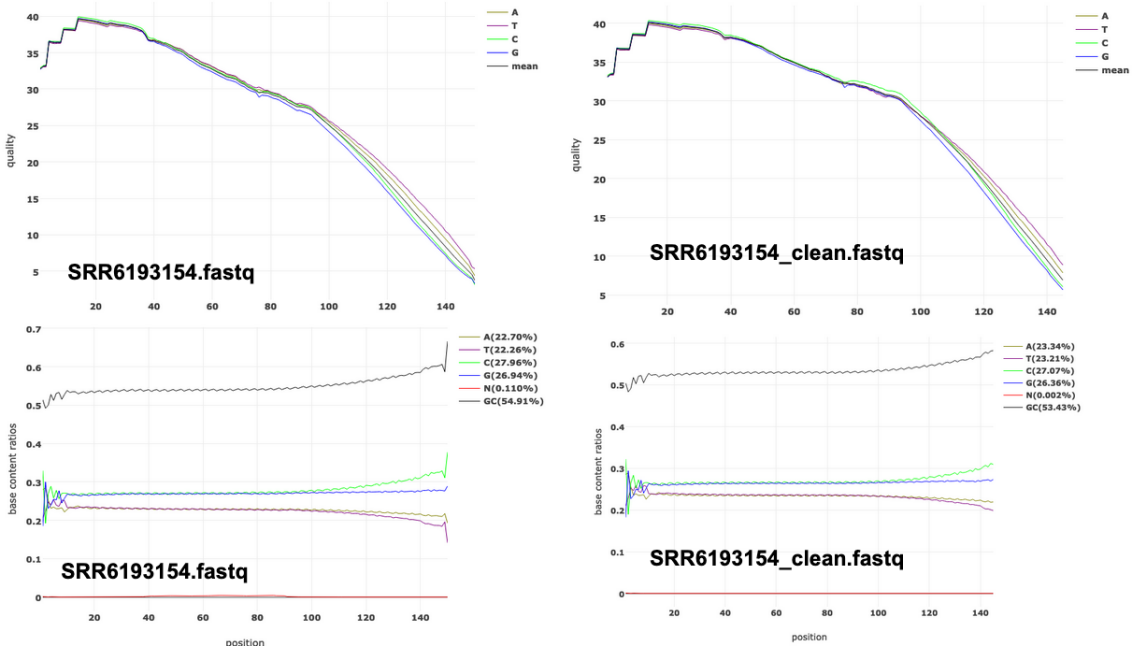

Figura 45: Calidad de la muestra SRR6193154. En la sección superior se muestra la calidad por nucleótido a través de la secuencia de DNA antes y después de la limpieza, exclusivamente para el read forward. En la sección inferior se puede observar la representatividad por nucleótido a través del *read*.

Las siguientes gráficas muestran el número de secuencias y la longitud de secuencias, respectivamente después de hacer la limpieza con Trim-Galore y Fastp.

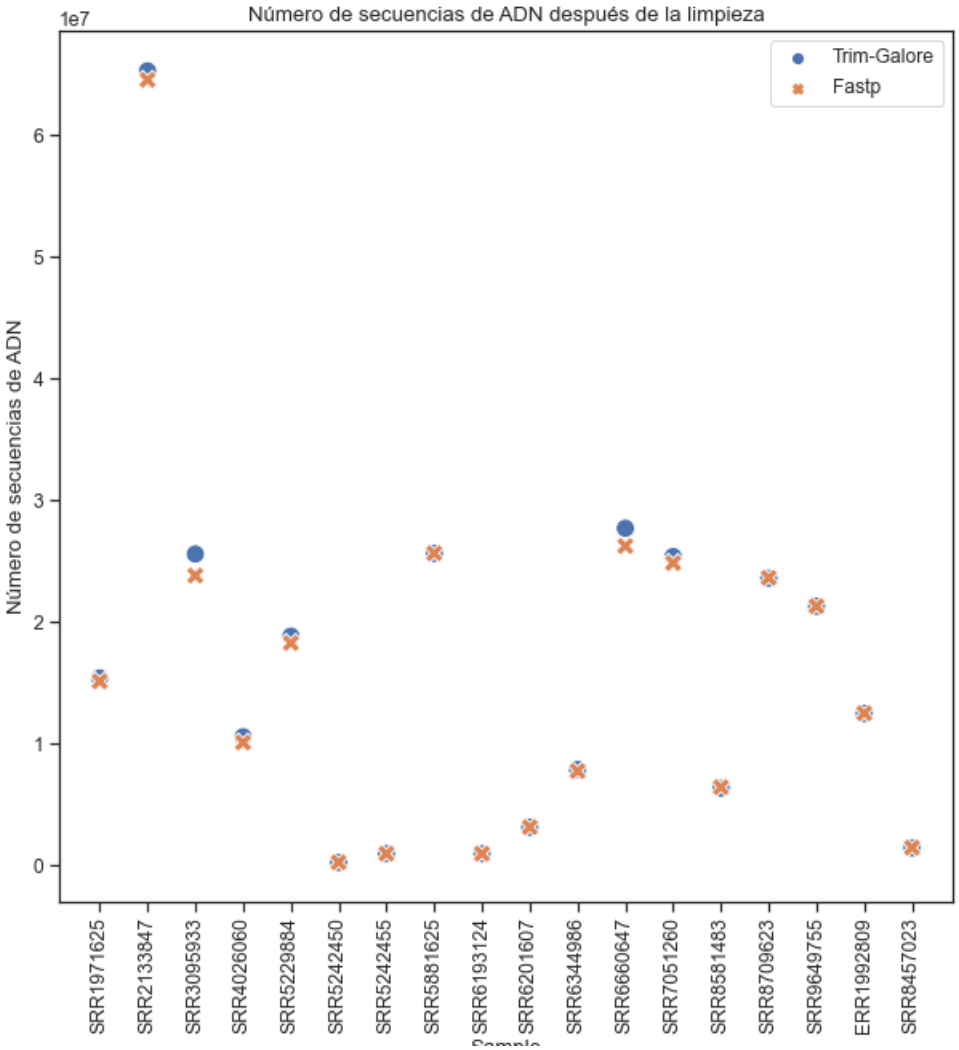

Figura 46: Número de secuencias después de la limpieza con Trim-Galore y Fastp

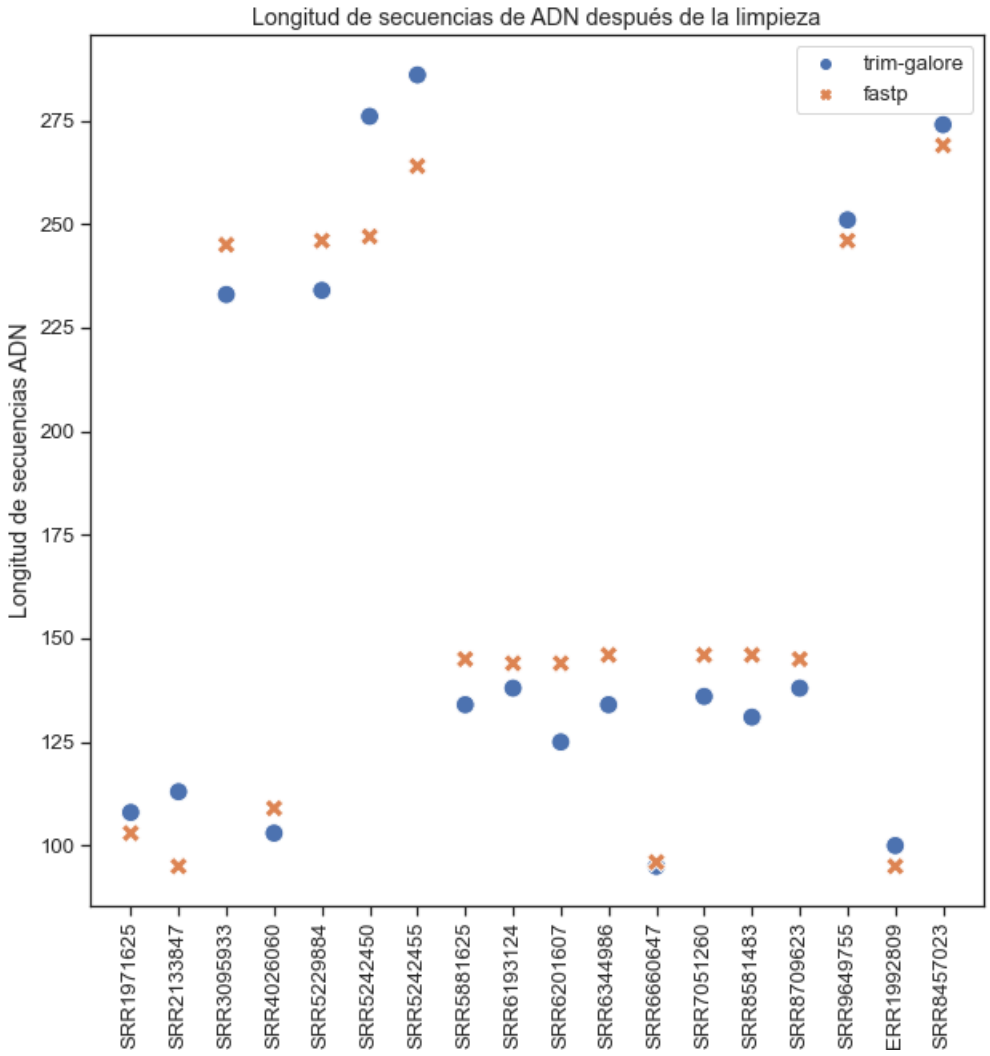

Figura 47: Longitud de secuencias después de la limpieza con Trim-Galore y Fastp

En lo referente al flujo de trabajo no.2, después de la realización de la limpieza por calidades con el software fastp, el proceso de filtrado siguiente, consistió en la eliminación de secuencias duplicadas, producto de PCR o duplicados ópticos resultado del software de detección de imágenes usado por la plataforma de secuenciación. Para la eliminación de dichas secuencias repetidas, se utilizó la herramienta CD-HIT-DUP, con el siguiente comando:

```
$ cd-hit-dup -i SRA_ID_QF_1.fastq -i2 SRA_ID_QF_2.fastq -o
SRA_ID_QF_Dp_1.fastq -o2 SRA_ID_QF_Dp_2.fastq -e 0 -m false
```

16. Anexo 1:Ensamble de las muestras

Una vez que las secuencias se han limpiado correctamente, el siguiente paso consiste en el ensamble de los *reads* en secuencias de mayor longitud nombradas contigs. En general, existen dos paradigmas que ayudan a la resolución de un ensamble, el primero se conoce como OLC (Overlap Layout Consensus, por sus siglas en inglés), el cual identifica sobrelapamientos entre los *reads* para resolver un contig o secuencia de mayor tamaño. Por otro lado, los algoritmos que emplean gráficas de de Bruijn, en donde las lecturas de DNA se fragmentan en secuencias de tamaño n conocidas como k-meros, los cuales representan los nodos de la gráfica, las conexiones entre nodos resultan del sobrelapamiento sin mismatches de k-1 bases entre los k-meros. De tal forma, que el algoritmo busca el camino que explique mejor las uniones entre k-meros.

Existen una gran cantidad de programas que permiten resolver un ensamble. Para el presente análisis se utilizaron los siguientes:

- **Megahit:** Este programa crea gráficas de Bruijn para resolver ensambles de novo de una manera muy eficiente, al usar gráficas sucintas de Bruijn reduciendo la necesidad de memoria y disminuye el sesgo derivado de errores de secuenciación. Elimina contigs que presenten una baja cobertura.

  Git-hub: Megahit

- **MetaSpades:** Ensamblador basado en la construcción de gráficas de de Bruijn con múltiples tamaños de k-mero, al mismo tiempo detecta y elimina las quimeras que surgen en los *reads* y detecta información de variantes intraespecie (cepas). Spades v3.13.1

Los comando específicos que se utilizaron para los ensambles de las muestras son los siguientes:

Megahit v1.2.9:

```
$ megahit -1 sample_1.fq -2 sample_2.fq -t [2-5] –k-list
21,33,55,77,99,111,127 -o sample_megahit
```

Spades v3.13.1:

```
$ spades.py –meta -1 sample_1.fq -2 sample_2.fq -k 21,33,55,77,99,111,127
-o sample_metaspades
```

En el caso específico de los ensambles realizados con MetaSpades, para el flujo de trabajo no. 1, fue necesario verificar que exista el mismo número de reads en los dos archivos (forward y reverse), es decir, que las lecturas queden apareadas, para poder realizar el ensamble de las mismas correctamente.

Ahora bien, con la finalidad de eliminar la redundancia en los ensambles resultantes, para el marco de procesamiento de los datos no. 2, se realizó un paso más en el filtrado y eliminación de dicha redundancia, con el software CD-HIT-EST. En el siguiente recuadro se muestra el comando utilizado para las muestras:

```
$cd-hit-est -i SRA_ID.fasta -o SRA_ID_CDhit.fasta -c 0.9 -M 50000 -aS 0.9 -G
0 -l 100 -p 1 -g 1 -T NUM
```

En las gráficas de las figuras número 5 y 6 se muestran el número total de contigs antes y después de la eliminación de redundancia con CD-HIT-EST.

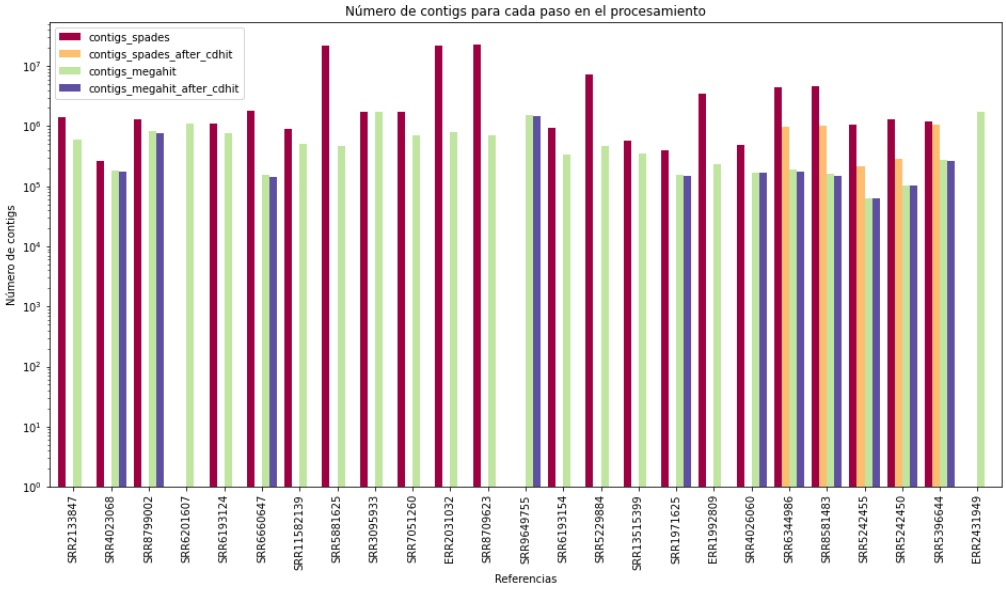

Figura 48: Número total de contigs, después de cada paso en su procesamiento para las muestras clasificadas como referencias.

La siguiente gráfica muestra el porcentaje de redundancia eliminado en los ensambles realizados con Megahit.

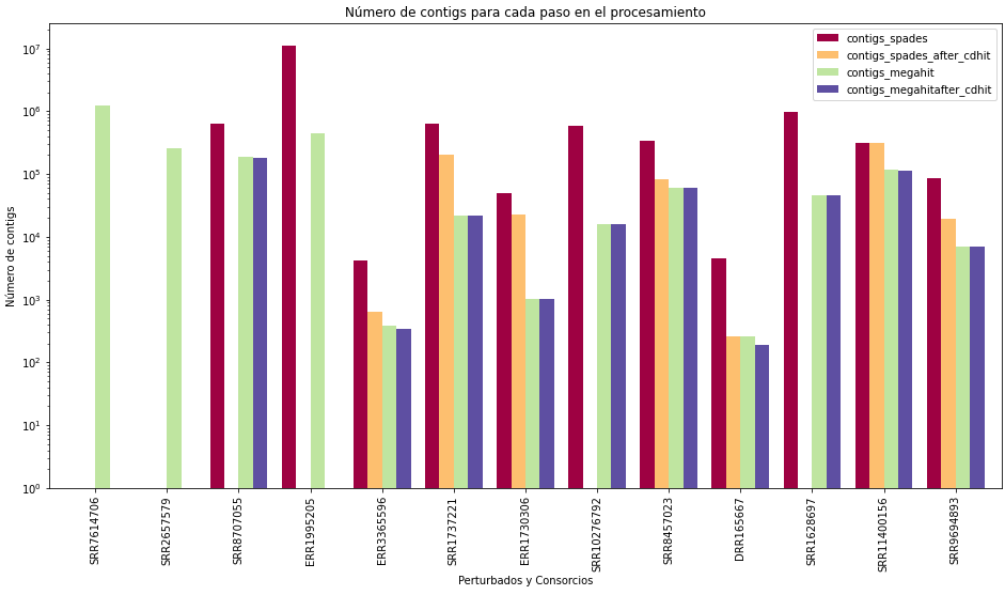

Figura 49: Número total de contigs, después de cada paso en su procesamiento para las muestras clasificadas como perturbados y consorcios

17. Anexo 1: Validación de ensambles

Debido a la heterogeneidad propia de las muestras, de la complejidad del ambiente del que provienen e incluyendo también la variabilidad entre muestras (es decir, las tres categorías establecidas), fue necesario realizar una comparación entre los métodos utilizados para ensamblar, con la finalidad, de obtener aquél ensamble que refleje con mayor veracidad el contenido funcional y taxonómico.

Para la validación de los ensambles realizados, se utilizó el programa MetaQuast, con la finalida de acceder a métricas que reflejen la continuidad, integridad y precisión del ensamble, con el siguiente comando:

```
$ metaquast.py SRA_ID_contigs_megahit.fa SRA_ID_contigs_spades.fa
–threads $NSLOTS
```

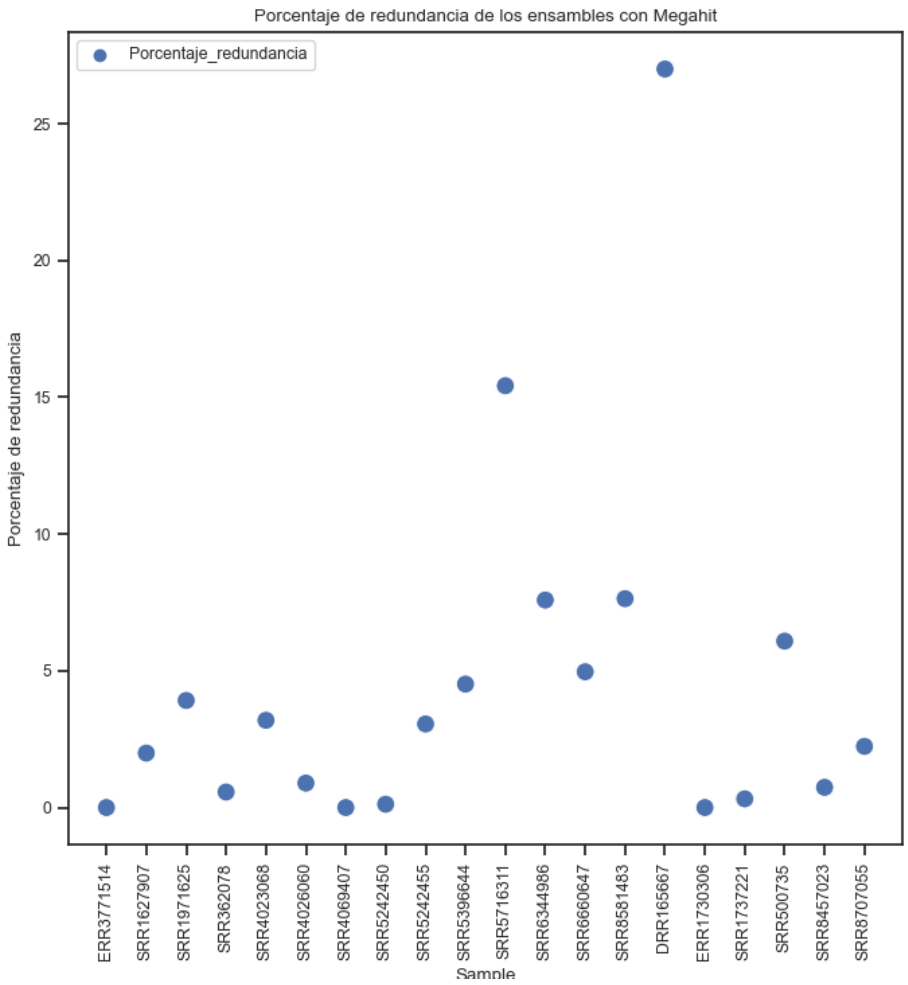

Figura 50: Porcentaje total de redundancia eliminado en los ensambles realizados con Megahit

Los resultados de las tres propiedades analizadas: precisión, integridad y continuidad se encuentran explicadas en el texto de la tesis.

18. Anexo 1: Anotación funcional

El paso de anotación a nivel funcional para los dos métodos de procesamiento de las mustras, se realizó con el software EggNOG-emmaper, con los siguientes parámetros:

- Búsqueda con Diamond: e-value = 0.001
- –itype metagenome
- Identifiación de secuencias codificantes con Prodigal – –genepred prodigal

en el siguiente recuadro se muestra el comando utilizado para cada una de las muestras:

$ emapper.py –cpu NUM –data_dir /data/eggnog_db -i SRA_ID_contigs.fasta –itype metagenome –genepred prodigal -o SRA_ID_genes

19. Anexo 1:Método Estadístico

En el campo de la estadística, al conjunto de observaciones que describen el comportamiento global de una población se le denomina muestra.

En este caso concreto, las anotaciones derivadas de los metagenomas analizados, representan la(s) muestra(s) que describe(n) de manera general el metabolismo que llevan a cabo grupos de procariontes habitantes de sedimentos marinos.

Con el objetivo de asignar a cada una de las enzimas anotadas, una distribución de probabilidad, que sirva como base para identificar una medida que califique la representatividad de dicha enzima en los metagenomas dentro de las categorías Perturbado y Consorcio, se realizaron los siguientes puntos:

- Contar el número de veces que existe cada enzima en cada uno de los metagenomas analizados.
- Calcular la frecuencia relativa de cada una de las enzimas encontradas, de la siguiente manera:

$$RF = n/N(10^6)^1 \tag{5}$$

  Donde, *RF* se refiere a la frecuencia relativa para cada una de las enzimas o proteínas anotadas en el metagenoma, n el número de veces que se encuentra la enzima o proteína en el metagenoma analizado, N el número total de proteínas o enzimas en todo el metagenoma.

- Dado que cada una de las enzimas anotadas representa una variable continua, para observar la tendencia central y la forma de la distribución de los datos se construye un histograma de frecuencias, con la finalidad de ajustar los datos a alguna de las siguientes distribuciones: Gamma, Normal, Lognormal o Weibull, estas distribuciones candidatas se ajustarán a los datos a través del método (Máxima verosimilitud)

[1] Introducción de datos faltantes de las frecuencias relativas al multiplicar por 10^6.

- Ajustar los valores de los parámetros de cada una de las distribuciones usando el método de Máxima Verosimilitud, además la prueba de Anderson-Darling se realiza para determinar que tan bien se ajustan estos datos a la distribución asignada.
- Cálculo de la probabilidad
- Cálculo del z-score (para las distribuciones no normales se asume una media = 0 y una desviación estándar = 1, de tal forma que son z-score(s) equivalentes a la distribución normal o z).

20. Glosario

COG: Clasificación filogenética de proteínas de genomas completos. Grupos de clusters de ortólogos.

Illumina: Técnica de secuenciación por síntesis de ADN.

Metagenoma shotgun: Metodología de la microbiología ambiental que permite evaluar la diversidad de microorganismos y detectar la abundancia de estos en un ambiente. También permite conocer o estudiar organismos que no se pueden cultivar en laboratorio.

Profundidad de secuenciación: El número de veces que nucleótido se encuentra representado en las secuencias de ADN generadas por la plataforma de secuenciación.

PacBio: Metodología de secuenciación de cuarta generación. Produce secuencias de ADN más largos en comparación con otras técnicas de secuenciación.

Secuencias de ADN: o *reads*, se refiere a los fragmentos de ADN producidos por plataformas de secuenciación.

Puntuación Phred: o calificación Phred, es una medida de calidad asignada a los nucleótidos detectados por plataformas de secuenciación de ADN. Esta mediada refleja la

probabilidad de error en la asignación de cada base o nucleótido por las plataformas de secuenciación.

Split-reads: En el proceso de re-mapeo de secuencias al genoma de referencia o secuencia consenso. Cuando una porción de las secuencia mapean en un lugar y la otra porción de la secuencia en otro distinto del genoma consenso o de referencia.

SEED: Jerarquía de clasificación de funciones del metabolismo global de organismos.

contig: Secuencia de ADN de mayor tamaño generada a partir del ensamble de secuencias cortas de ADN o reads.

Número EC: Esquema númerico utilizado para la clasificación de enzimas basado en las reacciones químicas que catalizan. La nomenclatura consiste en la siguiente: EC 1: Oxidoreductasas, EC 2:Transferasas, EC 3: Hidrolasas, EC 4: Liasas, EC 5: Isomerasas, EC 6: Ligasas y EC 7: Translocasas.