



**UNIVERSIDAD POLITÉCNICA SALESIANA  
SEDE EL GIRON-QUITO  
CARRERA DE BIOTECNOLOGÍA**

**ENSAMBLAJE DE GENOMA DE *Sphingobium yanoikuyae* A PARTIR DE DATOS DE  
SECUENCIACIÓN DE ILLUMINA Y PACBIO ARCHIVADOS EN LA BASE DE DATOS DE  
NCBI.**

**Trabajo de titulación previo a la obtención del título de:  
INGENIERAS BIOTECNÓLOGAS**

**AUTORES: ANDRADE GONZÁLEZ KAROL SARAI & PAZMIÑO CARRERA  
CAMILA MAYTE**

**TUTOR: VACA SUQUILLO IVONNE DE LOS ÁNGELES**

**Quito-Ecuador**

**2024**

## CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Nosotros, Karol Sarai Andrade González con documento de identificación N° 1751846120 y Camila Mayte Pazmiño Carrera con documento de identificación N° 1724949076; manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 13 de septiembre del año 2024

Atentamente,



-----  
Karol Sarai Andrade González  
1751846120



-----  
Camila Mayte Pazmiño Carrera  
1724949076

**CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE  
TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA**

Nosotros, Karol Sarai Andrade González con documento de identificación No.1751846120 y Camila Mayte Pazmiño Carrera con documento de identificación No.1724949076, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Trabajo experimental: “ Ensamblaje de genoma de *Sphingobium yanoikuyae* a partir de datos de secuenciación de Illumina y PACBIO archivados en la base de datos de NCBI”, el cual ha sido desarrollado para optar por el título de: Ingenieras Biotecnólogas en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribimos este documento en el momento que hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 13 de septiembre del año 2024

Atentamente,



-----  
Karol Sarai Andrade González  
1751846120



-----  
Camila Mayte Pazmiño Carrera  
1724949076

## CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Ivonne De Los Ángeles Vaca Suquillo con documento de identificación N° 1714726906 docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: ENSAMBLAJE DE GENOMA DE *Sphingobium yanoikuyae* A PARTIR DE DATOS DE SECUENCIACIÓN DE ILLUMINA Y PACBIO ARCHIVADOS EN LA BASE DE DATOS DE NCBI, realizado por Karol Sarai Andrade González con documento de identificación N° 1751846120 y por Camila Mayte Pazmiño Carrera con documento de identificación N° 1724949076, obteniendo como resultado final el trabajo de titulación bajo la opción Trabajo experimental que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 13 de septiembre del año 2024

Atentamente,

A handwritten signature in blue ink, reading "Ivonne Vaca Suquillo", written over a horizontal dashed line.

Ivonne Vaca Suquillo MSc.

C.I. 1714726906

## **Resumen**

*Sphingobium yanoikuyae* es una bacteria perteneciente a la familia Sphingomonadaceae, se caracteriza porque degrada varios compuestos aromáticos, ya que posee enzimas que metabolizan estos compuestos contaminantes, por ello tiene un enorme potencial para ser usada en biorremediación. Se realizó una investigación para establecer un *pipeline* para el ensamblaje del genoma de *S. yanoikuyae*, a partir de secuencias de Illumina y PACBIO, almacenadas en NCBI. Las secuencias crudas se preprocesaron mediante filtrado y recorte; a través de la implementación de herramientas bioinformáticas en la plataforma de Galaxy y KBase; se hizo el control de calidad en *FASTQC* obteniendo parámetros como el N50, el tamaño total de ensamblaje y el número de *contigs*. Además, programas para ensamblaje como *SPAdes*, *Megahit*, *Skesa* o *Velvet* permitieron realizar un ensamblaje, sin embargo, sus resultados fueron poco precisos y confiables; mientras que, con *Unicycler* se obtuvo un mejor ensamblaje del genoma completo, destacando el ensamblaje fusionado o híbrido con 89 *contigs* y una longitud de 5402608 pb. Finalmente, para su anotación se usó *Prokka* identificando 5045 elementos genéticos.

**Palabras clave:** Flujo de trabajo, programas bioinformáticos, ensamblaje híbrido, *contigs*.

## **Abstract**

*Sphingobium yanoikuyae* is a bacterium belonging to the Sphingomonadaceae family, it is characterized by the fact that it degrades several aromatic compounds, since it has enzymes that metabolize these contaminating compounds, therefore it has enormous potential to be used in bioremediation. Research was carried out to establish a pipeline for *S. yanoikuyae* genome assembly from Illumina and PACBIO sequences, stored at NCBI. Raw sequences were preprocessed by filtering and trimming; through the implementation of bioinformatics tools on the Galaxy and KBase platform; Quality control was done in *FASTQC* obtaining parameters such as the N50, the total assembly size and the number of contigs. Furthermore, assembly programs such as *SPAdes*, *Megahit*, *Skesa* or *Velvet* allowed an assembly to be carried out, however their results were not very precise or reliable; while, with *Unicycler*, a better assembly of the complete genome was obtained, highlighting the fused or hybrid assembly with 89 contigs and a length of 5402608 bp. Finally, *Prokka* was used for annotation, identifying 5045 genetic elements.

**Keywords:** Workflow, bioinformatics programs, hybrid assembly, contigs.

## Índice de contenidos

1	Introducción .....	1
2	Fundamentación teórica .....	3
2.1	<i>Sphingobium yanoikuyae</i> .....	3
2.1.1	Taxonomía .....	3
2.1.2	Genoma de la especie .....	3
2.1.2.1	Genoma de referencia .....	4
2.1.2.2	Genes de interés relacionados a la degradación de hidrocarburos .....	5
2.2	Ensamblaje y anotación del genoma .....	6
2.2.1	Aislamiento de ADN .....	6
2.2.2	Secuenciación .....	6
2.2.3	Ensamblado de secuencias .....	8
2.2.4	Anotación .....	9
2.3	Análisis bioinformático .....	10
2.3.1	Control de calidad .....	10
2.3.2	Preprocesamiento .....	11
2.3.3	Ensamblaje .....	12
2.3.4	Evaluación de calidad del ensamblaje .....	14
2.3.5	Pulido .....	15
2.3.6	Anotación .....	16

3	Materiales y métodos .....	17
3.1	Preprocesamiento de las secuencias .....	17
3.2	Ensamblaje de secuencias Illumina (R1 y R2) .....	19
3.2.1	Control de calidad del ensamblaje de secuencias de Illumina .....	20
3.3	Fusión de ensamblajes con <i>Unicycler</i> (secuencias Illumina y PACBIO) .....	20
3.3.1	Plataforma KBase .....	20
3.3.1.1	Control de calidad en la plataforma KBase .....	21
3.3.2	Plataforma Galaxy .....	21
3.3.2.1	Control de calidad en la plataforma Galaxy .....	21
3.4	Anotación .....	22
3.4.1	Plataforma KBase .....	22
3.4.1.1	<i>Prokka</i> .....	22
3.4.1.2	<i>RAST</i> .....	22
3.4.2	Plataforma Galaxy .....	22
3.4.2.1	<i>Prokka</i> .....	22
4	Resultados y discusión .....	24
4.1	Calidad inicial de las secuencias crudas .....	25
4.1.1	Illumina SRR27033680 (R1 y R2) .....	25
4.1.2	PACBIO SRR27033679 .....	26
4.2	Calidad de las secuencias preprocesadas con <i>Trimmomatic</i> .....	26
4.2.1	Illumina SRR27033680 R1 y R2 .....	26
4.2.2	PACBIO SRR27033679 .....	27

4.3	Ensamblaje .....	28
4.3.1	Ensamblaje de las secuencias de Illumina.....	28
4.3.2	Fusión de ensamblajes Illumina y PACBIO.....	30
4.4	Anotación .....	32
4.4.1	Plataforma KBase.....	32
4.4.1.1	<i>Prokka</i> .....	32
4.4.1.2	<i>RAST</i> .....	34
4.4.2	Plataforma Galaxy.....	37
4.4.2.1	<i>Prokka</i> .....	37
5	Conclusiones.....	39
6	Bibliografía .....	40
7	Anexos .....	49

## **Índice de figuras**

Figura 1. Métodos y programas utilizados en el ensamblaje y anotación de *S. yanoikuyae*. .. 24

Figura 2. Longitud acumulada de ensamblajes de Illumina mediante Shovill..... 29

## Índice de tablas

Tabla 1. Taxonomía de <i>Sphingobium yanoikuyae</i> .....	3
Tabla 2. Características de las secuencias de <i>S. yanoikuyae</i> . ....	17
Tabla 3. Estadísticas básicas de las secuencias crudas de Illumina y PACBIO.....	25
Tabla 4. Estadísticas básicas de las secuencias recortadas en Trimmomatic de Illumina (R1 y R2) y PACBIO. ....	27
Tabla 5. Resultados de ensamblaje mediante Shovill de la secuencia de Illumina.....	28
Tabla 6. Resultados de la fusión de ensamblajes con Unicycler.....	31
Tabla 7. Resultados de CheckM de la fusión de ensamblajes.....	32
Tabla 8. Anotación con Prokka del genoma ensamblado de <i>S. yanoikuyae</i> .....	34
Tabla 9. Anotación con RAST del genoma ensamblado de <i>S. yanoikuyae</i> . ....	36
Tabla 10. Resultados de Prokka para la secuencia fusionada en Galaxy. ....	38

## Índice de anexos

Anexo 1. Resultados de <i>FASTQC</i> de las secuencias crudas de Illumina R1 y R2.....	49
Anexo 2. Resultados de <i>FASTQC</i> de la secuencia cruda de PACBIO.....	51
Anexo 3. Resultados de <i>FASTQC</i> de las secuencias de Illumina R1 y R2 antes y después de <i>Trimmomatic</i> .....	52
Anexo 4. Resultados de <i>FASTQC</i> de la secuencia de PACBIO antes y después de correr <i>Trimmomatic</i> . .....	52
Anexo 5. Resultado de las secuencias de Illumina recortadas. ....	53
Anexo 6. Resultados de la secuencia de PACBIO recortada. ....	53

## 1 Introducción

*Sphingobium yanoikuyae* es una bacteria perteneciente a la familia Sphingomonadaceae, se encuentra en varios hábitats terrestres, de agua dulce o salada y se caracteriza por degradar varios contaminantes, ya que es capaz de oxidar un gran abanico de compuestos aromáticos, como: hidrocarburos aromáticos policíclicos (PAH), y polihidroxicarboxilatos (PHA); resultando de mucho interés para combatir el incremento de la contaminación por hidrocarburos, se pronosticó que la contaminación por hidrocarburos irá en aumento, incrementando hasta en un 60% para el 2021 (INABIO, 2019; Ní Chadhain et al., 2007). Su capacidad degradadora se debe a la presencia de enzimas, como las oxigenasas, las cuales se involucran en la metabolización de los compuestos contaminantes (Mitra et al., 2020), adicionalmente la bacteria puede usarlos como fuentes de carbono, oxidándolos para convertirlos en compuestos solubles (Ministerio para la Transición Ecológica y el Reto Demográfico, 2022). *S. yanoikuyae* tiene un enorme potencial para ser usada en campos ambientales y agrícolas debido a su amplia versatilidad, y a sus características biorremediadoras y degradadoras de sustancias contaminantes (Sánchez et al., 2022).

Conocer sobre los mecanismos de degradación y las aplicaciones en biorremediación, es posible mediante la localización de genes y la reconstrucción del genoma de *S. yanoikuyae*, Zhao et al. (2015), en su estudio obtuvieron un ensamblaje de 5,2 Mb, mediante el cual determinaron 35 dioxigenasas y encontraron 71 genes, algunos pueden estar relacionados con la diseminación de operones de degradación de PAH. El analizar el genoma de *S. yanoikuyae* por metodologías de ensamblaje y anotación, permite comprender el funcionamiento de su capacidad degradadora y su utilidad en diferentes campos. Además, ayudará en futuros análisis de expresión génica de ARN (Y. Wang et al., 2018); estudios y detección de mecanismos, identificación de genes de interés, análisis filogenéticos, entre otros (Sánchez et al., 2022). Por esto, el desarrollo de distintos *pipelines* bioinformáticos para ensamblar, alinear, pulir y anotar el genoma de la

especie es importante para la industria, sobre todo al describir los genes fundamentales en su capacidad de biodegradación de contaminantes.

La presente investigación fue realizada en la Universidad Politécnica Salesiana bajo el auspicio del grupo de investigación BIOARN, para establecer el proceso de ensamblaje para el genoma de *Sphingobium yanoikuyae* a partir de secuencias almacenadas en NCBI con número de accesión SRR27033680 (*Sphingobium yanoikuyae* HAMBI\_1842; short read Illumina) y SRR27033679 (*Sphingobium yanoikuyae* HAMBI\_1842; long read PACBIO), publicadas en diciembre de 2023, de las cuales no se reportan investigaciones previas hasta la fecha.

El objetivo general es generar un *pipeline* que permita el ensamblaje de lecturas crudas provenientes del SRA de NCBI, de la especie *Sphingobium yanoikuyae*, obtenidas a través de secuenciación PACBIO e Illumina, mediante la evaluación de la calidad las lecturas crudas y su preprocesamiento, el ensamblaje de las lecturas procesadas y la evaluación de la calidad de los ensamblajes finales, usando diferentes herramientas bioinformáticas para procesar las secuencias hasta obtener un ensamblaje adecuado y su posterior anotación.

## 2 Fundamentación teórica

### 2.1 *Sphingobium yanoikuyae*

#### 2.1.1 Taxonomía

*Sphingobium yanoikuyae* es una bacteria que posee una gran capacidad para crecer y sobrevivir en condiciones de deficiencia de nutrientes, mediante el uso de varios compuestos orgánicos y siendo capaz de degradar varios de los mismos (Duhan et al., 2023; Sánchez et al., 2022). A continuación, en la Tabla 1 se presenta la taxonomía de esta especie.

Tabla 1. Taxonomía de *Sphingobium yanoikuyae*

<b>Reino</b>	Bacteria
<b>Filo</b>	Proteobacteria
<b>Clase</b>	Alphaproteobacteria
<b>Orden</b>	Sphingomonadales
<b>Familia</b>	Sphingomonadaceae
<b>Género</b>	<i>Sphingobium</i>
<b>Especie</b>	<i>Sphingobium yanoikuyae</i>

Fuente: (Global Biodiversity Information Facility, 2023).

#### 2.1.2 Genoma de la especie

El genoma completo de *Sphingobium yanoikuyae* SHJ contiene 1 cromosoma circular y 2 plásmidos circulares, esta cepa degrada ésteres de ftalato. Los tamaños del cromosoma 1, el plásmido pSES220 y el plásmido pSES189 son 5260163 pb, 220037 pb y 189183 pb, respectivamente, y sumados dan un total de 5669383 pb. El contenido de GC del genoma de la cepa SHJ es del 64,23%. Se han identificado un total de 5402 genes, de los cuales 5183 son genes codificadores de proteínas y 143 pseudogenes. Los 76 genes restantes son genes de ARN, incluidos 12 genes de ARNr, 61 genes de ARNt y otros 3 genes de ARN no codificantes (Feng et al., 2018).

Por otro lado, en el artículo de Sánchez et al. (2022), se investigó el genoma de *Sphingobium yanoikuyae* S72, con el fin de determinar su base genética para la degradación de hidrocarburos, donde el genoma fue reducido a 1 *contig* y se analizaron varios genes únicos de los cuales varios han sido adquiridos por transferencia horizontal, que están comúnmente en las islas genómicas de bacterias o como elementos genéticos móviles; que estarían involucrados en degradación de contaminantes como xenobióticos, hidrocarburos, bifenilo o naftaleno; siendo útil para la biorremediación de suelos contaminados. Adicionalmente, con la herramienta *Island Viewer* 4.0 se visualizaron islas genómicas, siendo las regiones II y III las más grandes, con 317308 kb y 579060 kb de longitud, respectivamente que indica la presencia de secuencias genómicas significativas, que presentan alguna función o adaptación específica relevante para *S. yanoikuyae* S72.

La predicción genética del genoma de la cepa *Sphingobium yanoikuyae* YC-XJ2 mostró una longitud total del genoma de 5272134 pb, con un contenido del 64,86% de GC en la región genética y del 58,07% en la región intergenética; contiene 12 genes de ARNr, 64 genes de ARNt, 3 profagos (89345 pb), 22 islas de genes (560628 pb) y 5782 CDS (X. Li et al., 2020).

#### **2.1.2.1 Genoma de referencia**

Los genomas de referencia son un conjunto de datos que tiene como objetivo modelar y representar la secuencia de ADN del genoma de una especie, son ensamblados por científicos para servir como secuencias de ADN representativas para distintas especies (Alquicira, 2017). Los investigadores utilizan este mapeo para identificar nuevos genes, variantes de genes conocidos y otros elementos funcionales, y para compartir y comparar sus hallazgos con otros científicos (T. Wang et al., 2022).

El genoma de referencia de *S. yanoikuyae* cepa S72 (ASM250408v1), fue obtenido por el Instituto Politécnico Nacional de México en el año 2017. La bacteria fue aislada de la rizosfera

de una planta de sorgo en la ciudad de Río Bravo, Tamaulipas, México y secuenciada mediante la tecnología de Illumina (NCBI, 2017).

### **2.1.2.2 Genes de interés relacionados a la degradación de hidrocarburos**

En la cepa XLDN2-5 de *Sphingobium yanoikuyae*, se encontraron el grupo de genes car (carRAaBaBbCAc) y el gen *fdr*, que están acompañados en ambos lados por dos copias de elementos IS 6100 y organizados como IS 6100:: ISSspI- ORF1-carRAaBaBbCAc-ORF8 - IS 6100 - *fdr* -IS 6100. Estos genes se unieron al grupo de genes *ant* (*antRAcAdAbAa*), que participa en la conversión de antranilato en catecol, también está intercalado entre dos elementos IS6100 como IS6100-*antRAcAdAbAa*-IS6100. Juntos, los genes estructurales y los elementos IS6100 forman dos transposones catabólicos, responsables de la degradación del carbazol. Además, el gen *fdr* en *S. yanoikuyae* es un componente clave en el metabolismo microbiano, específicamente en la vía de degradación de hidrocarburos y una amplia gama de contaminantes ambientales (Gai et al., 2011).

Los genes *bphC* y *xylE*, codifican 2,3 -dihidroxi-bifenil 1,2-dioxigenasa y catecol 2,3-dioxigenasa, respectivamente, en la vía catabólica de PAH de *S. yanoikuyae* B1 (Cunliffe & Kertesz, 2006). Los genes *bphC* han demostrado que sus productos genéticos codificados por el operón *bph*, están implicados en la degradación del dibenzofurano y del bifenilo, mediante la fisión del anillo 1,2-dihidroxi-dibenzofurano catalizada por *bphC* (Wesche et al., 2005). Mientras los genes *xylE* se encargan de codificar la enzima catecol 2,3 dioxigenasa (C23O), la cual es capaz de degradar toluenos, benzoatos y sus derivados metílicos (Saunders et al., 1996).

Muchos genes están asociados con la degradación de xenobióticos e hidrocarburos, como la oxidorreductasa dependiente de NAD(P) de la familia SDR (código A6768\_07840), la aldo/ceto reductasa (código A6768\_07825), el citocromo P450 (código A6768\_11830), la 4-

hidroxibenzoato 3-monooxigenasa (código A6768\_12975) y el alcohol reductasa aromática (código A6768\_RS13080) (Sánchez et al., 2022).

## **2.2 Ensamblaje y anotación del genoma**

### **2.2.1 Aislamiento de ADN**

Para el cultivo de *S. yanoikuyae* se toma una muestra de la bacteria de ambientes terrestres o acuáticos, estas son transportadas al laboratorio según metodologías estándar ya establecidas y se siembran. Luego, se recogen los sedimentos celulares mediante centrifugación y se lava en solución salina comprobando un pH de 7. El ADN de la bacteria se extrae con un kit de purificación de ADN genómico siguiendo las indicaciones propuestas; por último, se cuantifica usando un fluorómetro, previo a la secuenciación (Mitra et al., 2020; Sánchez et al., 2022).

### **2.2.2 Secuenciación**

La secuenciación de genomas completos es un método que permite identificar los genes en un organismo, es una herramienta básica para posteriores análisis funcionales de los nuevos genes descubiertos. La secuencia genómica provee de un conjunto virtual de todas las proteínas que el organismo puede expresar (Aguilar-Bultet & Falquet, 2015). En el proceso de secuenciar un genoma, el ADN de un organismo es aislado, leído por un secuenciador y convertido en información digital que puede ser procesada por una computadora (Vera, 2014).

El método de secuenciación de Sanger permitió innumerables logros en este campo, como la secuenciación del primer genoma bacteriano (*Haemophilus influenzae*) y la primera secuencia completa del genoma humano. Esta tecnología, por sus limitaciones, trajo consigo la necesidad de desarrollar nuevas y mejores alternativas para secuenciar muchos genomas en corto tiempo (Aguilar-Bultet & Falquet, 2015). Debido a esto surgen las tecnologías de secuenciación de nueva generación (NGS, next-generation sequencing), como:

- Illumina, secuencia genomas pequeños y funciona mediante el proceso químico de secuenciación por síntesis (SBS) (Illumina, 2021). La secuenciación por síntesis consiste en la incorporación de nucleótidos utilizando una variedad de enzimas y esquemas de detección que permiten que la plataforma del instrumento correspondiente recopile datos al mismo tiempo que la síntesis enzimática en una plantilla (McCombie et al., 2019). Se prepara la librería y el ADN uno de sus extremos a adaptadores mediante complementariedad, atraviesan una celda de flujo, mientras los adaptadores se cohesionan a la placa y con polimerasas se amplifican en clústeres. Para la secuenciación por síntesis se elimina una de las cadenas, se adiciona un nucleótido con fluorescencia característica, con lo que se obtiene una secuencia exacta de nucleótidos del fragmento de ADN. La secuenciación puede ser *single read*, si se realiza desde un solo extremo de la doble cadena; o puede ser *paired end read*, incluyendo ambos extremos de la cadena (Vaca, 2024).
- PACBIO, este sistema brinda acceso a una secuenciación de lectura larga de alto rendimiento con secuenciación a tiempo real de una única molécula (SMRT) (PacBio, 2024). Las moléculas de ADN polimerasa, unidas a una plantilla de ADN, se unen al fondo de pocillos de 50 nm de ancho, a cada polimerasa se le permite llevar a cabo la síntesis de ADN de la segunda cadena, en presencia de nucleótidos marcados con fluorescencia de  $\gamma$ -fosfato, dando información sobre la secuencia en la replicación de la molécula de ADN objetivo (Quail et al., 2012).
- Ion Torrent, utiliza la tecnología de semiconductores para proporcionar la secuenciación de sobremesa de última generación más rápida. Su librería está unida a un adaptador, las secuencias se hibridan a perlas, y se realiza una amplificación mediante polimerasas (Vaca, 2024).

Ésta usa chips con sensores de voltaje en cada pocillo capaces de detectar los Hidrógenos (H<sup>+</sup>) liberados cada vez que un nucleótido se une a la nueva cadena en formación debido a cambios en el pH (Thermo Fisher Scientific, 2024).

- Oxford Nanopore, utiliza celdas de flujo que contienen una serie de pequeños agujeros (nanoporos) incrustados en una membrana electro resistente. Cada nanoporo corresponde a su propio electrodo conectado a un canal y un chip sensor, que mide la corriente eléctrica que fluye a través del nanoporo; cuando una molécula pasa a través de él, la corriente se interrumpe para producir un "garabato" característico. Luego, el garabato se decodifica utilizando algoritmos de llamada base para determinar la secuencia de ADN o ARN en tiempo real (Oxford Nanopore Technologies, 2024a).

### 2.2.3 Ensamblado de secuencias

Es el proceso para descifrar una secuencia genómica mediante alineamiento y mezcla de fragmentos de secuencias de ADN para reconstruir la secuencia original (Vera, 2014).

Los genomas pueden ser ensamblados de dos maneras: ensamblaje por comparación, en el que se utiliza un genoma como referencia; y ensamblaje *de novo*, en el cual se aplica solamente la información obtenida de la secuenciación para reconstruir el genoma, sin conocimiento anterior de la organización de este. Dependiendo de la información previa que se tenga respecto a la secuencia a ensamblar se definirá la mejor estrategia a seguir (Aguilar-Bultet & Falquet, 2015).

Los ensamblados se obtendrán usando herramientas bioinformáticas y se compararán usando métricas comunes para evaluar su calidad (Vera, 2014), como:

- Cantidad de *contigs* ensamblados: Un *contig* relacionado a estudios genómicos, se deriva de la palabra "contiguo", es un conjunto de segmentos o secuencias de ADN que se superponen parcialmente de forma tal que colectivamente dan una representación

continua de una región genómica. Un buen ensamblaje debe minimizar el número de fragmentos obtenidos del proceso de ensamblado a partir de un conjunto inicial de fragmentos.

- Suma de las longitudes de los *contigs*: un buen ensamblaje debe aproximar la suma de las longitudes de los *contigs* obtenidos, a la longitud total de la secuencia que se desea reconstruir.
- Longitud del *contig* más largo: es la longitud de la secuencia ensamblada más larga.
- N50: Es la longitud del *contig* más pequeño del conjunto de *contigs* más largos cuya suma de longitudes es al menos el 50% de la suma de las longitudes de todos los *contigs*.

#### 2.2.4 Anotación

La anotación de genoma es el proceso de unir información biológica a las secuencias. Se basa en dos pasos principales: la identificación de elementos en el genoma o predicción génica, y agregar información biológica a estos elementos. Actualmente, se usan distintos programas bioinformáticos como *BLAST*, para encontrar similitudes entre genomas (BIOREN-UFRO, 2024). Entre los tipos de anotación están:

- Anotación estructural: Busca predecir y localizar todas las secuencias codificantes (genes) y determinar e identificar la estructura de estas. Se usan programas bioinformáticos específicos para análisis del genoma estudiado (Cruz Cubas & Rolland Burger, 2013).
- Anotación funcional: Se encarga de completar y aprovechar el análisis estructural del transcriptoma y del proteoma. Predice funciones de la búsqueda de similitudes con otras secuencias de organismos de la misma especie (secuencias parálogas) o de otras especies (secuencias ortólogas) (Cruz Cubas & Rolland Burger, 2013).

- Anotación relacional: Identifica las redes de interacciones génicas y vías metabólicas en las que participan los productos de los genes (las proteínas) (Cruz Cubas & Rolland Burger, 2013).

## **2.3 Análisis bioinformático**

### **2.3.1 Control de calidad**

El control de calidad de secuencias crudas es un proceso que se encarga de evaluar y garantizar la calidad de las lecturas (*reads*) obtenidos directamente desde los secuenciadores. Usando programas bioinformáticos, que eliminan las secuencias de baja calidad, los adaptadores y varios contaminantes, garantizando una secuencia óptima para análisis posteriores. El paso inicial para el preprocesamiento de datos crudos es el control de calidad para cualquier experimento o investigación que pretenda llevarse a cabo, para ejecutarlo existen diversas herramientas de acceso libre en línea, muchos de ellos permitiendo procesar datos crudos FASTQ sin ningún formato en especial. De esta forma, se comprueba la calidad de la secuencia, la base y distribución de los nucleótidos, por medio de varios parámetros o métricas de control para proporcionar una visión concisa de la calidad de la muestra, permitiendo identificar cualquier defecto y arreglarlo para análisis posteriores (Guo et al., 2014).

Es un proceso crítico para cualquier tecnología de secuenciación porque puede comprometer los resultados y conclusiones de un experimento. Por esto, la detección e interpretación temprana de sesgos de secuenciación o alineación de las lecturas, es fundamental y es necesario contar con sistemas de control de calidad que sean rápidos, livianos y basados en los valores atípicos que podrían llegar a presentarse en la lectura por medio de varios métodos estadísticos (Kumar et al., 2020).

- *FASTQC* es un programa que hace una serie de análisis básicos y estándar de calidad, realiza un análisis muy completo, que abarca varios parámetros como estadísticas

básicas, calidad de secuencia por base, distribución de longitud de secuencia y entre otros, que permiten identificar problemas que hubiera durante la secuenciación; genera un reporte para cada uno de los parámetros indicando si la muestra analizada es correcta, presenta advertencias o errores. El programa permite importar archivos de entrada y salida en los formatos BAM, SAM o FASTQ (Babraham Institute, 2024).

### **2.3.2 Preprocesamiento**

El preprocesamiento realiza un filtrado de secuencias con errores de lectura exigiendo una cantidad mínima de cotejos exactos entre las demás cadenas (Vera, 2014). Es común que los datos obtenidos por tecnologías de secuenciación de nueva generación tengan información de mala calidad, por lo cual es necesario el preprocesamiento de la posible contaminación generada por la presencia de adaptadores en la lectura para los posteriores procesos de análisis, además debido a que los secuenciadores actuales tienen cada vez más capacidad de realizar secuenciación de lecturas de mayor longitud; el preprocesamiento es un filtro que permite ir descartando la mayor parte de los sesgos de información inválida (Sun, 2020).

Entre los softwares utilizados para preprocesar secuencias se encuentran:

- *Trimmomatic* es una herramienta que hace varias tareas de recorte útiles para datos con extremo único y emparejado de Illumina. Los pasos para recorte y los parámetros asociados se proporcionan por línea de comando o en plataformas bioinformáticas como KBase y Galaxy. Funciona con archivos en formato FASTQ (Bolger, 2014).
- *Porechop* es un programa para buscar y eliminar adaptadores de lecturas de Oxford Nanopore, elimina los adaptadores en los extremos de las lecturas y, cuando el adaptador se encuentra en el medio corta en lecturas separadas. Además, hace alineamientos complejos para buscar adaptadores, aun cuando la identidad de secuencia es baja. Permite la manipulación de archivos FASTA y FASTQ (Galaxy, 2024).

- *PRINSEQ* es una herramienta para filtrar, reformatear o recortar los datos de una secuencia genómica y metagenómica. El programa crea estadísticas resumidas de las secuencias como gráficos y tablas de la longitud de lectura, el contenido de GC, puntuaciones de calidad, entre otros parámetros. Usa los formatos FASTA o FASTQ como entrada (PRINSEQ, 2024).

### 2.3.3 Ensamblaje

Los programas utilizados para el ensamblaje de genoma son:

- *Shovill* es un ensamblador de genomas aislados de bacterias partiendo de lecturas *paired* de Illumina. *Shovill* canaliza y utiliza el programa de *SPAdes* en su núcleo alterando los pasos previos y posteriores del ensamblaje principal, con el fin de obtener el ensamblaje en menos tiempo. De igual forma esta herramienta admite el uso de otros ensambladores como son *Skesa*, *Velvet* y *Megahit*. Se debe considerar que *Shovill* es únicamente para datos aislados en FASTQ y no funciona con genomas muy grandes como el metagenoma, al final se obtiene el ensamblaje en formato FASTA (Seeman, 2020).
- *SPAdes* es una herramienta desarrollada para el ensamblaje *de novo* de datos de secuenciación para genomas microbianos aislados y cultivados para secuenciación de ADN genómico unicelular, actualmente se ha ampliado sus funciones permitiendo hacer un ensamblaje híbrido usando lecturas cortas y largas, llamándolo un *biread* (Prjibelski et al., 2020). *SPAdes* se basa en los gráficos de De Bruijn con la información obtenida de los k-mers en el ensamblaje híbrido (k-bimers); acepta archivos de entrada en formato FASTQ, *interleaved* o archivos *contigs*; generando archivos de salida de *contigs* o andamios y una tabla de estadísticas acerca de las métricas de calidad del ensamblaje (Vera, 2014).
- *Skesa* es un ensamblador *de novo* que está basado en los gráficos de De Bruijn, acepta archivos de entrada en los formatos SRA, FASTA o FASTQ. *Skesa* se encarga de

ensamblar los genomas microbianos sobre todo de secuencias de Illumina, la herramienta usa distintas longitudes de k-mers, y heurísticas conservadoras creando rupturas en regiones repetitivas del genoma, cabe recalcar que el programa también hace un previo recorte de calidad de los datos; generando un genoma ensamblado en formato FASTA (Souvorov et al., 2018; Souvorov & Agarwala, 2021).

- *Megahit* es un ensamblador de un solo nodo de nueva generación que se encarga de ensamblar distintos datos metagenómicos y complejos de una manera sencilla, pero a su vez eficiente, con una buena calidad, usando múltiples k-mers que se recomienda que sean impares para un mejor ensamblaje. La herramienta ensambla los datos obtenidos como un todo, es decir; que en algunas ocasiones para ciertos datos no es necesario un preprocesamiento ya que el mismo programa por lo general limpia las lecturas erróneas (D. Li et al., 2015).
- *Velvet* es un software que se encarga de ensamblar lecturas de ADN previamente recortadas en *contigs* o andamios; a partir de lecturas cortas por manipulación de los grafos de De Bruijn, donde se puede agregar dos bibliotecas separadas; una para lecturas emparejadas y otra para una sola lectura; con las cuales a través de distintos parámetros buscará el mejor ensamblaje automáticamente. Además, *Velvet* puede leer archivos FASTA, FASTQ, SAM O BAM (Gladman, 2019). *Velvet* tiene dos fases H y G que se describen a continuación:
  - *Velveth*: Es un subprograma que ayuda a formar la estructura de datos, produce una tabla hash, luego genera dos archivos en un directorio de salida, secuencias y hojas de ruta que se usarán para ensamblar el genoma completo, mediante el programa *Velvetg* (Blanco, 2013; Zerbino, 2008).

- *Velvetg*: Es el núcleo del ensamblador. Se encarga de construir el Grafo de De Bruijn y manipularlo para conseguir ensamblar las lecturas (Blanco, 2013; Zerbino, 2008).
- *Unicycler* es un software para procesos de ensamblaje de genomas bacterianos, forjando un ensamblaje híbrido, que utiliza una combinación de lecturas cortas y largas. Puede ensamblar conjuntos de lectura de solo Illumina donde funciona como optimizador de *SPAdes*. También puede ensamblar conjuntos de solo lectura larga (PACBIO o Nanopore), donde ejecuta un *pipeline* de *Miniasm* + *Racon* o *Pilon*. Al incluir estas herramientas, este software adicionalmente efectúa el pulido del ensamblaje de la secuencia de interés, por ello, no es necesario realizar un pulido adicional. Los mejores ensamblajes posibles con *Unicycler*, se realizan mediante lecturas de Illumina y lecturas largas, con las que se generará un ensamblaje híbrido de lectura corta primero. Después de obtener el ensamblaje fusionado inicial, el software integrará las lecturas largas para mejorar la continuidad y corregir errores (Wick et al., 2017).

#### 2.3.4 Evaluación de calidad del ensamblaje

La evaluación de la calidad del ensamblaje se realiza mediante diversas herramientas bioinformáticas (Galaxy, 2024; PRINSEQ, 2024):

- *QUAST* (Quality ASsessment Tool) es una herramienta para evaluar el ensamblaje del genoma que funciona con y sin genomas de referencia (Gualdrón, 2022). Sin embargo, es mucho más informativo si se proporciona al menos un genoma de referencia cercano junto con los ensamblajes. La herramienta acepta múltiples ensamblajes, por lo que es adecuada para comparar. Aplica métricas basadas en *contigs* como: número de *contigs* y su longitud total, longitud del *contig* más grande, N50, entre otras métricas basadas en la longitud de la secuencia y alineación de esta, que se muestran cuando se proporciona un genoma de referencia (Gurevich, 2021).

- *CheckM* es un programa que de manera automatizada evalúa la calidad del genoma obtenido, mediante el uso de un amplio repertorio de genes marcadores específicos dentro del genoma, basándose en algún genoma o árbol genómico de referencia. La herramienta proporciona una estimación bastante acertada de la integridad, la contaminación del genoma analizado y la presencia de errores al momento del ensamblaje. Utiliza como formato de entrada FASTA, y genera archivos de salida como gráficos e informes de calidad (Parks et al., 2015).

### 2.3.5 Pulido

Es el procesamiento de un ensamblaje de secuencia en borrador, típicamente un genoma. Por lo general, para eliminar artefactos del proceso de ensamblaje y mejorar la precisión local y del consenso general de la secuencia de borrador ensamblado (Oxford Nanopore Technologies, 2024b). Este proceso también se aplica a regiones de interés específicas, al combinar múltiples copias exactas de un único fragmento o molécula original en una única secuencia de alta calidad. Entre las herramientas que deben usarse para pulir el ensamblaje respectivo (Oxford Nanopore Technologies, 2024b), están:

- *Racon* está pensado como un módulo de consenso independiente para corregir *contigs* sin procesar, generados por métodos de ensamblaje rápido que no incluyen un paso de consenso. El objetivo de *Racon* es generar consenso genómico que sea de calidad similar o mejor en comparación con el resultado generado por métodos de ensamblaje que emplean pasos de corrección de errores y de consenso, además es mucho más rápido en comparación con esos métodos. *Racon* se puede utilizar como herramienta de pulido después del ensamblaje con datos de Illumina o con datos producidos por secuenciación de tercera generación. *Racon* toma como entrada solo tres archivos: *contigs* en formatos FASTA/FASTQ, lecturas en formatos FASTA/FASTQ y superposiciones/alineaciones

entre las lecturas y los *contigs* en formatos MHAP/PAF/SAM. La salida es un conjunto de *contigs* pulidos en formato FASTA (Kundu et al., 2019).

- *Pilon* es un software para mejorar automáticamente los borradores de ensamblajes y descubrir alguna variación entre cepas. El software trabaja con archivos de entrada FASTA o BAM, analizando su alineación y descubriendo inconsistencias. Además, arregla el genoma de entrada y revisa la calidad del genoma (Gualdrón, 2022).

### 2.3.6 Anotación

En cuanto a los programas que se pueden aplicar para la anotación de la secuencia ensamblada tenemos:

- *Prokka* es una herramienta de software para anotar rápidamente los borradores de genomas ensamblados de bacterias, arqueas y virus, y producir archivos de salida que cumplan con los estándares. Coordina un conjunto de herramientas de software existentes para lograr una anotación rica y confiable de secuencias bacterianas genómicas. Es muy adecuado para modelos iterativos de análisis de secuencias e integración en procesos de software genómico. *Prokka* utiliza una variedad de bases de datos como: ISfinder y BLAST, cuando intenta asignar funciones a las características CDS previstas (Seemann, 2014).
- *RAST* (Anotación rápida mediante tecnología de subsistema): realiza anotación de genomas bacterianos y arqueales completos o casi completos. Genera anotaciones genómicas de alta calidad, permitiendo el análisis de borradores de genomas; una vez lista toda la anotación, los genomas se descargan en varios formatos o se ven en línea. Adicionalmente, genera un mapeo de genes en subsistemas y una reconstrucción metabólica. Cabe resaltar que usa un formato de archivo especial llamado Genome Typed Object (GTO) (Brettin et al., 2015).

### 3 Materiales y métodos

Las dos secuencias de *Sphingobium yanoikuyae* fueron descargadas desde la base de datos públicos de NCBI, sus números de acceso son SRR27033680 (*Sphingobium yanoikuyae* HAMBI\_1842; *short read* Illumina; *paired*) y SRR27033679 (*Sphingobium yanoikuyae* HAMBI\_1842; *long read* PACBIO; *single*), ambas en formato FASTQ, publicadas en diciembre de 2023 (Tabla 2).

Tabla 2. Características de las secuencias de *S. yanoikuyae*.

Nombre	<i>Sphingobium yanoikuyae</i> HAMBI_1842; lectura corta (a)	<i>Sphingobium yanoikuyae</i> HAMBI_1842; lectura larga (b)
Código de acceso	SRX22725517	SRX22725518
Instrumento	ILLUMINA MiSeq	PACBIO_SMRT
Estrategia de secuenciación	WGS	WGS
Fuente	Genómica	Genómica
Selección	RANDOM	RANDOM
Disposición	PAIRED	SINGLE
Lectura	SRR27033680	SRR27033679
Número de bases	238,8M	666,7M
Tamaño	126,6Mb	546,5Mb
ID	30760073	30760074

Nota. Secuencias del repositorio de NCBI. a. PACBIO (NCBI, 2023a), b. Illumina (NCBI, 2023b).

Para las secuencias de *Sphingobium yanoikuyae* se formuló un *pipeline* que detalla las herramientas bioinformáticas para el preprocesamiento de las secuencias, el ensamblaje con diferentes herramientas, y también el posterior pulido y anotación de este.

#### 3.1 Preprocesamiento de las secuencias

Inicialmente la secuencia de Illumina SRR27033680 tenía un modo de lectura desde ambos extremos (*paired*), fueron separados en R1 (*forward*) y R2 (*reverse*), mediante línea de comando ‘grep -A 1 ".1 M03602" SRR27033680.fastq’ y ‘grep -A 1 ".2 M03602" SRR27033680.fastq’ en Linux; que es un sistema operativo de código abierto basado en el

sistema operativo Unix, que permite trabajar de manera más segura y gratuita que otros sistemas (Carrillo, 2023).

Posteriormente se revisó la calidad de las secuencias con los parámetros preestablecidos, en la plataforma Galaxy (<https://usegalaxy.eu/>) con la herramienta *FASTQC* versión 0.74+galaxy0, la cual proporciona una descripción general de las métricas básicas de control de calidad en las secuencias sin procesar y procesadas, provenientes de canales de secuenciación de alto rendimiento (Babraham Institute, 2024). Para el control de calidad se usaron como archivos de entrada las secuencias de Illumina SRR27033680 y de PACBIO SRR27033679 en formato FASTQ, el programa generó como archivos de salida un reporte en formato html, el cual contiene las métricas que se evaluaron; y un archivo txt, que da un resumen de las evaluaciones de calidad.

Más adelante se recortaron y filtraron las secuencias de Illumina SRR27033680 R1 y R2 (formato FASTQ), con la herramienta *Trimmomatic* versión 0.39+galaxy2. Los parámetros usados fueron:

- *Headcrop*, esta operación corta un número especificado de bases desde el inicio de la lectura (Bolger, 2014). Se eliminaron las bases desde la 1 hasta la 19, debido al error que presentaron los resultados obtenidos en *FASTQC* en la métrica contenido de secuencia por base (Anexo 1c).
- *Quality score encoding* de Phred 33, este es un sistema de codificación que se utiliza en secuenciaciones de Illumina de forma predeterminada, así como también *FASTQC* en los archivos que analiza para representar la calidad de las bases de secuenciación y permite que la información obtenida sea más concisa y fácil de manejar para los posteriores análisis (Al-Maeni & Al-Khazraji, 2021; Katta et al., 2015).
- El resto de los parámetros se mantuvieron por defecto.

Mientras que para la secuencia obtenida por PACBIO SRR27033679 (formato FASTQ) se usó *Trimmomatic* versión 0.36 en la plataforma de KBase, se corrió con los siguientes parámetros:

- *Quality score encoding* en Phred33, debido a su codificación en la secuenciación.
- *Post tail crop length* hace referencia al número de bases que se busca conservar desde el inicio de la lectura (Bolger, 2014), en este caso se usó un valor de 400.
- *Headcrop length* con un valor de 25, para cortar un número específico de nucleótidos al inicio de cada lectura, ya que se requiere eliminar las lecturas de baja calidad de esa zona de la secuencia (Bolger, 2014).
- *Minimum read length* de 30, especifica la longitud mínima de lecturas que queremos conservar, con el fin de evitar lecturas que sean demasiado pequeñas (Bolger, 2014).

Tras correr *Trimmomatic* se generaron 2 archivos de salida de lecturas emparejadas R1 y R2 en formato fastqsanger. Posterior a esto, se realizó un nuevo control de calidad de todas las secuencias preprocesadas con *FASTQC*, con la finalidad de comprobar si los recortes corrigieron los parámetros erróneos en las secuencias.

### **3.2 Ensamblaje de secuencias Illumina (R1 y R2)**

A través de la herramienta *Shovill* versión 1.1.0+galaxy2 se realizó el ensamblaje de las secuencias R1 y R2 de Illumina previamente recortadas. *Shovill* permite el uso de cuatro ensambladores: *Skesa*, *Megahit*, *Velvet* y *SPAdes*; por lo que, se realizaron ensamblajes con cada uno de ellos, utilizando los siguientes parámetros:

- *Input reads type*: se usó *paired*, este parámetro indica que el tipo de lecturas de secuenciación que se están usando para correr la herramienta (Tabla 2) (Atxaerandio-Landa et al., 2022).

- *Collection or single library* como *paired end*, ya que se refiere a la configuración original de las secuencias, que permite procesar las secuencias como lecturas cortas emparejadas (Atxaerandio-Landa et al., 2022).
- Los parámetros restantes se mantuvieron por defecto.

Como archivos de salida *Shovill*, generó un archivo de *contigs* en formato FASTA y un archivo *contig graph* en formato txt para cada uno de los ensambladores.

### 3.2.1 Control de calidad del ensamblaje de secuencias de Illumina

Se ejecutó la herramienta *QUAST* en su versión 5.2.0+galaxy0 con el archivo de salida de *Shovill* para el ensamblaje de Illumina. Con los parámetros predeterminados, añadiendo el genoma de referencia de la bacteria (ASM250408v1). La herramienta generó como archivo de salida una tabla reporte en formato html.

### 3.3 Fusión de ensamblajes con *Unicycler* (secuencias Illumina y PACBIO)

Se utilizó la herramienta *Unicycler* con la finalidad de realizar un ensamblaje de genoma fusionado y pulido, utilizando las secuencias preprocesadas de Illumina (R1 y R2) y PACBIO (formato FASTQ), la herramienta se evaluó en dos plataformas: KBase y Galaxy.

#### 3.3.1 Plataforma KBase

Se usó *Unicycler* en la versión 0.4.8, se establecieron los siguientes parámetros para la ejecución (Wick et al., 2017):

- *Minimum contig length* con un valor de 500, es decir, el *contig* más corto para aceptar en el ensamblaje resultante con la finalidad de tener un número de *contigs* que no sea demasiado pequeño o grande.
- *Minimum long read length* de 5000, que se refiere a la lectura "larga" más corta para aceptar. Porque, con un valor 5000 se eliminarían lecturas de mala calidad enfocándose en aquellas de buena calidad para el ensamblaje.

- *Expected number of linear contig* con un valor de 50, el número esperado de secuencias lineales en la secuencia subyacente para obtener un ensamblaje óptimo.
- *Threshold for bridging contigs* se procesó como *conservative*, la herramienta realiza una construcción de *contigs* puente para conectar pares de *contigs* de una sola copia, en el modo *conservative* el límite de calidad es alto, existe un riesgo muy bajo de mal ensamblaje, pero es menos probable producir un ensamblaje completo.
- Los demás parámetros se manejaron por defecto.

Al correr la herramienta se generó como archivo de salida una tabla reporte en formato html.

#### 3.3.1.1 Control de calidad en la plataforma KBase

El ensamblaje fusionado se analizó usando las herramientas *QUAST* versión 4.4 y *CheckM* versión 1.0.18 con todos sus parámetros por defecto; generando para *QUAST* una tabla reporte html y para *CheckM* un archivo *CheckM\_Plot.html* y una tabla resumen de estadísticas de calidad del ensamblaje en formato qa.

### 3.3.2 Plataforma Galaxy

Se aplicó *Unicycler* en la versión 0.5.0+galaxy1, con los mismos parámetros que se usaron en KBase; excepto por *Minimum long read length*, pues Galaxy no permite modificar este parámetro. Se generaron como archivos de salida, un archivo en formato FASTA del ensamblaje y un archivo de *SPAdes graph* en formato gfa1.

#### 3.3.2.1 Control de calidad en la plataforma Galaxy

La evaluación del ensamblaje fusionado fue efectuada mediante *QUAST* versión 5.2.0+galaxy0 con los parámetros predeterminados y el genoma de referencia de la bacteria (ASM250408v1) dando como resultado una tabla reporte en formato html.

## 3.4 Anotación

### 3.4.1 Plataforma KBase

#### 3.4.1.1 Prokka

La anotación del ensamblaje fusionado generado en KBase se realizó con la herramienta *Prokka*, en la versión 1.14.5, utilizando el ensamblaje fusionado. Se establecieron los siguientes parámetros:

- *Scientific name* de la especie NCBI Tax ID 13690: *Sphingobium yanoikuyae*;
- *Kingdom*: Bacteria;
- El resto de los parámetros se usaron con sus valores predeterminados.

El archivo de reporte de anotación de genoma generado por la herramienta se encuentra en formato GenBank y en formato gff.

#### 3.4.1.2 RAST

Al mismo tiempo, se realizó la anotación del mismo ensamblaje con la herramienta *RAST*, en la versión 1.073, empleando el ensamblaje de *Unicycler*. Se establecieron los siguientes parámetros:

- *Genetic Code* seleccionado como 11 (*Archaea, most Bacteria, most Virri, and some Mitochondria*);
- *Scientific name* de la especie NCBI Tax ID 13690: *Sphingobium yanoikuyae*;
- *Domain* como B (Bacteria);

Se produjo un archivo de reporte de anotación de genoma en formato GenBank y en formato gff.

### 3.4.2 Plataforma Galaxy

#### 3.4.2.1 Prokka

Para el ensamblaje fusionado y pulido ejecutado en Galaxy se realizó la anotación a través de *Prokka*, en la versión 1.14.6+galaxy1. Se establecieron los siguientes parámetros:

- *Genus name: Sphingobium*
- *Species name: Sphingobium yanoikuyae*
- *Kingdom: Bacteria*
- El resto de los parámetros se usaron con sus valores iniciales.

Se generaron doce archivos de reporte de anotación de genoma en formato fna, faa, ffn, fsa, tbl, gff, log, txt, err, tsv, sqn y gbk.

## 4 Resultados y discusión

Se realizó el *pipeline* para el preprocesamiento, ensamblaje y anotación del genoma de *Sphingobium yanoikuyae*, que se describe a continuación (Figura 1):

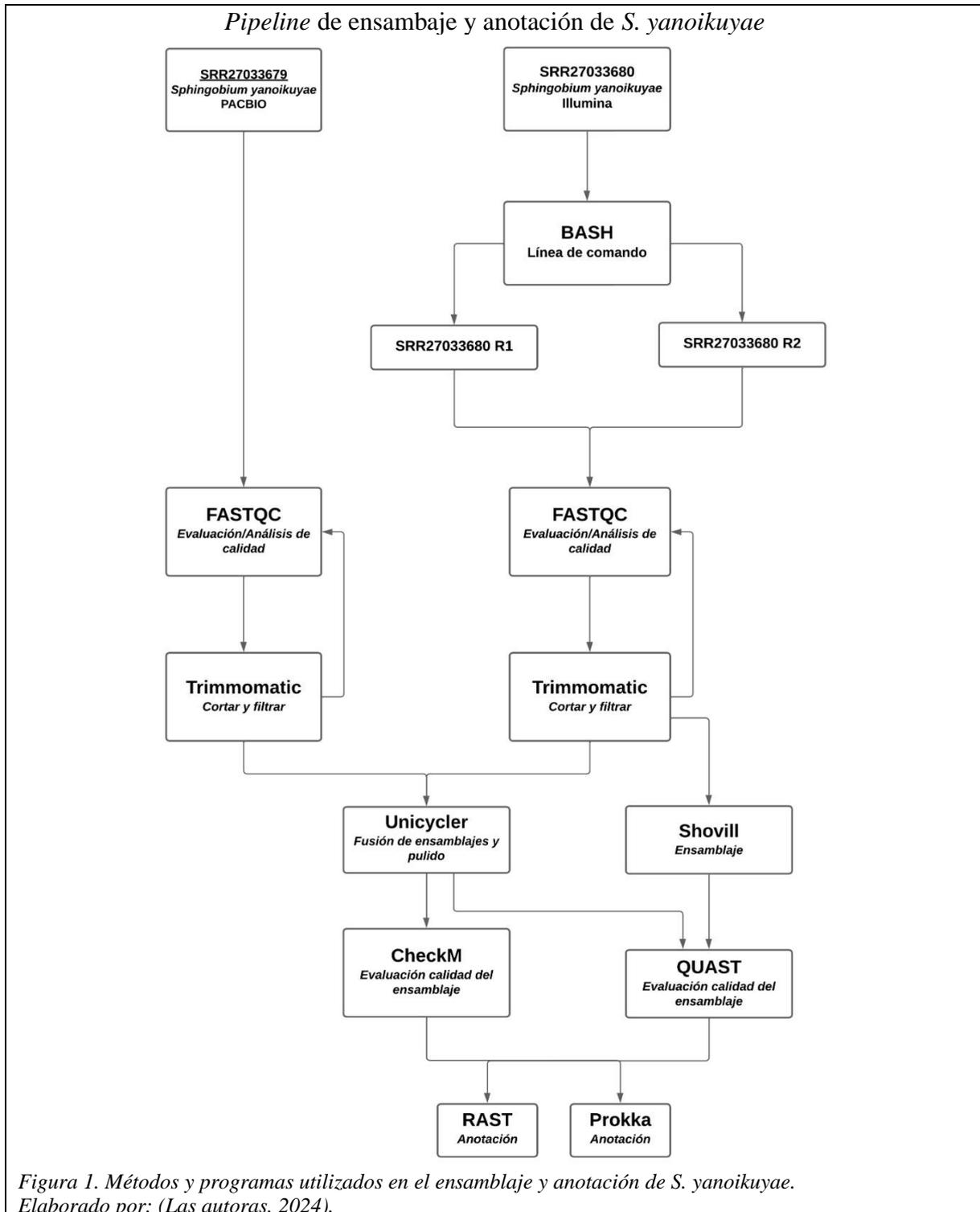


Figura 1. Métodos y programas utilizados en el ensamblaje y anotación de *S. yanoikuyae*. Elaborado por: (Las autoras, 2024).

## 4.1 Calidad inicial de las secuencias crudas

### 4.1.1 Illumina SRR27033680 (R1 y R2)

Los resultados de *FASTQC* para las secuencias de Illumina previamente separadas en R1 y R2 (Tabla 3; Anexo 1), fueron los mismos. Su calidad de secuencia por base fue de 30 en escala de calidad Phred, marcado por una línea azul en la zona verde del gráfico, representando una calidad óptima de sus bases (Anexo 1a; **Error! No se encuentra el origen de la referencia.**) (Babraham Institute, 2024). También, se obtuvo una puntuación de calidad adecuada, porque no hubo presencia de ningún subconjunto con una calidad deficiente (Anexo 1b).

Tabla 3. Estadísticas básicas de las secuencias crudas de Illumina y PACBIO.

	Illumina (R1 y R2) <i>forward y reverse</i>	PACBIO
<b>Medida</b>	<b>Valor</b>	<b>Valor</b>
ID de la secuencia	SRR27033680	SRR27033679
Secuencias totales	824568	63512
Bases totales	238,7 Mbp	666,6 Mbps
Secuencias de mala calidad	0	0
Longitud de la secuencia	26-300	550-30451
%GC	64	64

Fuente: (Las autoras, 2024).

El reporte generado reveló problemas en la uniformidad de las bases, desde el inicio de la secuencia hasta aproximadamente la posición número 19 (Anexo 1c), lo que indica un contenido muy disperso entre las bases de ADN (Babraham Institute, 2024). En cuanto al contenido de GC, hubo una diferencia insignificante entre el valor teórico comparado con la lectura analizada, por lo que el nivel de GC estaba equilibrado (Anexo 1d). Por otro lado, no se reportó la presencia de proporciones N en ambas secuencias (Anexo 1e). Adicionalmente, se reflejó una variabilidad significativa en la distribución de la longitud de las lecturas (Anexo 1f), el contenido de secuencias duplicadas fue del 21,17 % (Anexo 1; **Error! No se encuentra el origen de la referencia.g**) y no se encontró adaptadores en las secuencias (Anexo 1; **Error! No se encuentra el origen de la referencia.h**).

#### **4.1.2 PACBIO SRR27033679**

El reporte indicó una buena calidad de bases por secuencia (Tabla 3, Anexo 2a), ya que se encontraba en la región verde de la gráfica, tampoco hubo presencia de proporciones significativas de una calidad general baja (Anexo 2b). Se reflejó desigualdad en la uniformidad de las bases al inicio de la secuencia y al final de esta (Anexo 2c). El gráfico del contenido de CG evidencia diferencia entre el valor teórico versus la lectura analizada (Anexo 2d). No presenta proporciones N (Anexo 2e), presentó variabilidad en la distribución de la longitud de las lecturas (Anexo 2f). Por otro lado, hubo un 0,35 % de duplicación en la parte inicial de la secuencia (Anexo 2g) y no muestra la presencia de adaptadores (Anexo 2h).

### **4.2 Calidad de las secuencias preprocesadas con *Trimmomatic***

#### **4.2.1 Illumina SRR27033680 R1 y R2**

El parámetro que producía problemas de calidad en la secuencia era el contenido de la secuencia por bases. Tras el recorte realizado mediante *Trimmomatic* (Tabla 4), la secuencia mejoró en la uniformidad de las trazas que representan a las bases, presentándose líneas casi paralelas entre sí. Por lo que, el porcentaje de cada una de las bases de ADN se tornó cercano. Como sugiere Martínez (2023), en su investigación durante el proceso de *trimming* se eliminan las lecturas de mala calidad y los adaptadores para evitar análisis erróneos de la secuencia analizada. Por otro lado, la distribución de la longitud de las secuencias y el contenido de GC por secuencia se mantienen en un rango aceptable para la buena calidad de la secuencia, ya que el contenido de GC, que es del 64% (Tabla 4), se aproxima al reportado para el genoma de la cepa SHJ, que es del 64,23% (Feng et al., 2018).

Tabla 4. Estadísticas básicas de las secuencias recortadas en Trimmomatic de Illumina (R1 y R2) y PACBIO.

	Illumina R1 y R2	PACBIO
<b>Medida</b>	<b>Valor</b>	<b>Valor</b>
ID de la secuencia	SRR27033680	SRR27033679
Secuencias totales	412284	31756
Bases totales	113,8 Mbp	11,9 Mbp
Secuencias de mala calidad	0	0
Longitud de la secuencia	7-281	375
%GC	64	64

Fuente: (Las autoras, 2024).

#### 4.2.2 PACBIO SRR27033679

Existían varias irregularidades en los siguientes parámetros analizados: el contenido de la secuencia por base y la distribución de longitud por secuencia presentaban una advertencia; mientras que el contenido de GC por secuencia, presentaba error. Estos fueron mejorados a una calidad muy buena después del recorte realizado mediante *Trimmomatic* (Anexo 4), ya que ahora todas las métricas no presentan advertencia, ni error.

Después del proceso de *trimming*, mejoró la uniformidad de las bases (Anexo 4a); además se reportó una mejora significativa del contenido de GC por secuencia, porque la curva del conteo de GC de la secuencia analizada es muy similar a la curva teórica (Anexo 6b), corrigiendo adecuadamente este parámetro. Finalmente, en la distribución de la longitud de la secuencia recortada mejoró su uniformidad y consistencia, mejorando la calidad de esta (Anexo 6c).

Los resultados obtenidos son similares a los presentados por MacManes (2014), ya que se detectaron nucleótidos de baja calidad con una alta probabilidad de error en su secuencia, los cuales se eliminaron previamente al ensamblaje para tener una mayor precisión. Siendo

relevante para el estudio porque, al mejorar la calidad se tiene un gran efecto en la calidad del ensamblaje, el mismo que al contar con muy pocos errores de nucleótidos en relación con alguna referencia establecida pueden indicar alta calidad.

### 4.3 Ensamblaje

#### 4.3.1 Ensamblaje de las secuencias de Illumina

En la Tabla 5 se presentan los resultados de calidad, generados por la herramienta *QUAST*, para el ensamblaje realizado con las secuencias de Illumina mediante *Shovill*:

Tabla 5. Resultados de ensamblaje mediante *Shovill* de la secuencia de Illumina.

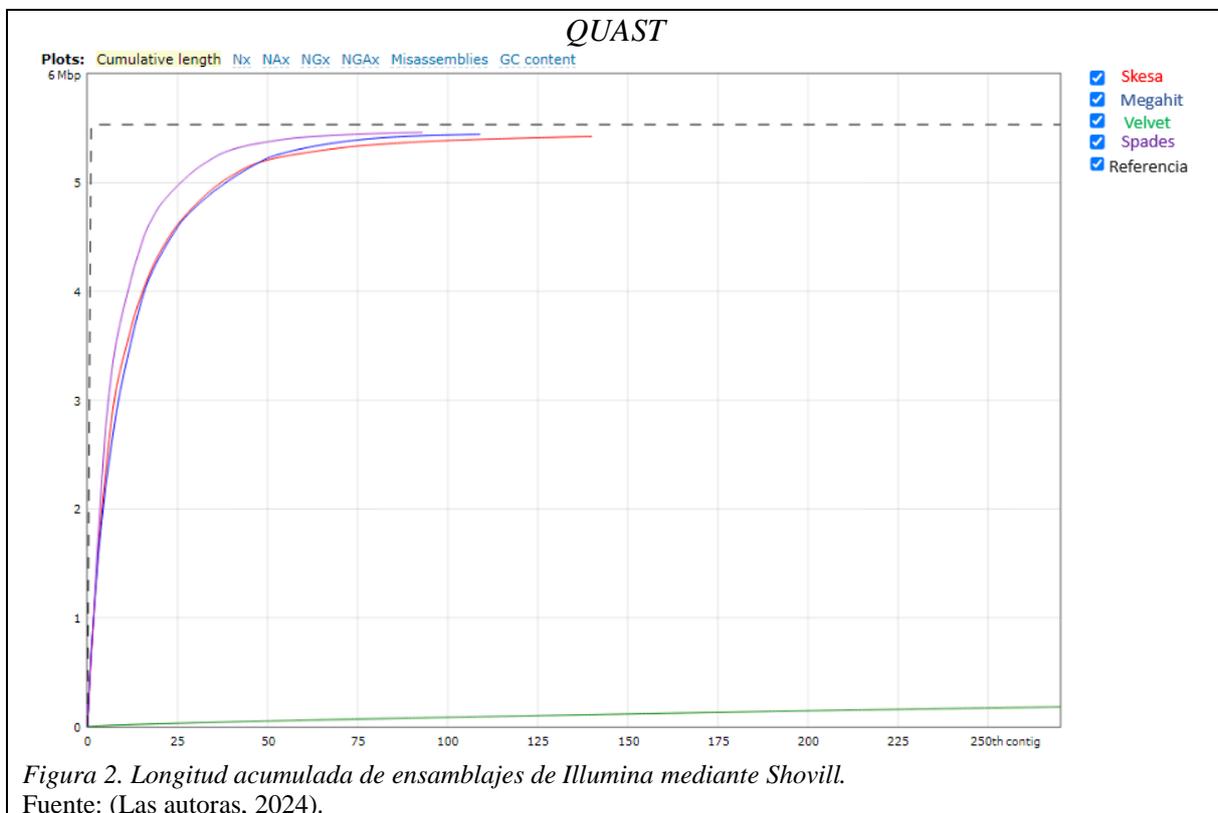
<b>Estadísticas del genoma</b>	<i>Skesa</i>	<i>Megahit</i>	<i>Velvet</i>	<i>SPAdes</i>
Fracción del genoma (%)	77,285	77,191	1,116	77,306
Tasa de duplicación	1,001	1,002	1	1,002
Alineación más grande	269792	232195	1112	269792
Longitud total alineada	4261834	4267973	61738	4269229
N50	268574	223940	630	477357
L50	7	8	109	5
<b>Errores de montaje</b>				
# desensamblajes	44	50	0	50
Longitud de los <i>contigs</i> mal ensamblados	4168017	4124554	0	4393477
<b>Desajustes</b>				
# discrepancias por 100 kbp	3183,14	3183,83	2428	3185,1
# indeles por 100 kbp	85,39	85,9	43,73	86,78
# N por 100 kbp	0	0	0	0
<b>Estadísticas sin referencia</b>				
# <i>contigs</i>	140	109	270	93
<i>Contig</i> más grande	629882	682954	2254	630141
Largo total	5423683	5443003	180457	5460460
Longitud total (>= 1000 pb)	5401189	5434903	20180	5452823

Fuente: (Las autoras, 2024).

El ensamblaje realizado con la herramienta *SPAdes* logró analizar el 77.306% del genoma, este fue el mejor, debido a la obtención del menor número de *contigs* y una longitud total de la secuencia de 5460460 pb (Tabla 5; Figura 2), que se asemeja significativamente a la longitud de la secuencia de referencia de 5532633 pb reportada por NCBI (2017); también se obtuvo el mejor N50 de 477357 nucleótidos, este valor representa el *contig* más pequeño, de modo que la mitad del genoma está representado por *contigs* de tamaño N50 o mayor, es decir, se

generaron *contigs* más grandes que en los demás ensamblajes, lo que se evidencia con un menor número de *contigs* (Thrash et al., 2020). Sin embargo, la longitud obtenida de *contigs* mal ensamblados fue de 4393477 bp, estos *contigs* mal ensamblados pueden deberse a una combinación de varios factores durante la secuenciación de la muestra de ADN, la misma calidad de la lectura, la contaminación, algoritmos empleados hasta la cobertura y la naturaleza misma del genoma que se está analizando (Alonso, 2021).

El ensamblador *Velvet* concluyó con poca eficiencia (Figura 2), siendo el peor ensamblaje ya que la fracción del genoma que analizó fue tan solo del 1.116%; obtuvo 270 *contigs* y una longitud total del genoma de 180457, que es alrededor de la treintava parte del genoma de referencia (NCBI, 2017). La longitud de *contigs* mal ensamblados fue de 0 y su N50 de 630.



En cuanto a las herramientas bioinformáticas *Skesa* y *Megahit*, consiguieron un número total de *contigs* de 140 para el primero y 109 para el segundo ensamblador, y también longitudes del

genoma similares de 5423683 y 5443003 pb, respectivamente (Tabla 5; Figura 2), estas longitudes alcanzadas son cercanas a la longitud del genoma de referencia de 5532633 pb (NCBI, 2017). Sin embargo, en cuanto a la longitud de *contigs* mal ensamblados se obtuvieron valores muy altos: 4168017 con *Skesa* y 4124554 con *Megahit*, y el N50 de 268574 y 223940 respectivamente, resultando ambos menos eficientes que el ensamblaje obtenido por *SPAdes*.

Los resultados obtenidos mediante *Shovill* y los ensambladores canalizados utilizados: *Skesa*, *Megahit*, *Velvet* y *SPAdes*, fueron poco precisos y confiables, principalmente los errores de montaje que presentan irregularidades muy altas. Según Seeman (2020), *Shovill* al modificar los pasos antes y después del proceso de ensamblaje principal para obtener resultados buenos en menos tiempo puede fallar, pues algunos de sus componentes pueden ser lentos y no manejar bien las lecturas superpuestas de pares. En adición, *Shovill* tiene muchas dependencias, si las dependencias entran en conflicto con otros programas o bibliotecas en ejecución en la plataforma, pueden producirse errores. Así mismo, las versiones del software de estas herramientas pueden presentar desaciertos o limitaciones que perjudican la calidad del ensamblaje. Por ello, no se recomiendan estas herramientas para el ensamblaje y anotación del genoma de la bacteria de interés.

#### **4.3.2 Fusión de ensamblajes Illumina y PACBIO**

A partir de las secuencias de Illumina y PACBIO, después del proceso de *trimming*, se usó el ensamblador *Unicycler* cuyos resultados del ensamblaje fusionado se exponen en la Tabla 6.

Después de ejecutar *QUAST* para evaluar la calidad de los ensamblajes realizados con *Unicycler*, en KBase se obtuvieron 107 *contigs* y en Galaxy 89 *contigs*, el N50 fue de 268581 pb y 310780 pb, respectivamente (Tabla 6), lo que significa que los *contigs* generados en Galaxy son mucho más grandes con respecto a KBase. Se obtuvo una longitud total del genoma de 5424101 pb y 5402608 pb, respectivamente; como indica Feng et al. (2018), la longitud del

genoma de *Sphingobium yanoikuyae* es de 5669383 pb, por lo que los genomas ensamblados, obtuvieron 245282 pb menos en el ensamblaje realizado en KBase y 266775 pb menos en Galaxy.

Tabla 6. Resultados de la fusión de ensamblajes con Unicycler.

<b>Estadísticas sin referencia</b>	<b>Unicycler en KBase</b>	<b>Unicycler en Galaxy</b>
# <i>contigs</i>	107	89
<i>Contig</i> más grande	920235	920234
Longitud total	5424101	5402608
Longitud total (>= 0 bp)	5424101	5402608
Longitud total (>= 1000 bp)	5413821	5397221
N50	268581	310780
L50	6	5
GC (%)	64,43	64,44
<b>Desajustes</b>		
# N's	0	0
# N's por 100 kbp	0	0

Fuente: (Las autoras, 2024).

Se estima que el mejor ensamblaje híbrido es el realizado en Galaxy, ya que, a pesar de que presenta una longitud de genoma ligeramente menor con respecto al ejecutado en KBase, y ambos son cercanos a la longitud de la secuencia de referencia de la bacteria (5669383 pb) (Feng et al., 2018); el ensamblaje obtenido en Galaxy tiene menos *contigs* generados (89), es decir, sugiere un ensamblaje menos fragmentado (Thrash et al., 2020). Además, tiene un N50 más alto (310780 pb) y un L50 más bajo (5 *contigs*), lo que muestra que los *contigs* son más largos y que se necesitan menos *contigs* para cubrir el 50% del ensamblaje (Tabla 6) (Thrash et al., 2020). Estos factores combinados permiten sugerir que el ensamblaje generado por *Unicycler* en Galaxy es de mayor calidad.

Los resultados de calidad obtenidos con la herramienta *CheckM* en la plataforma KBase, se presentan en la Tabla 7.

Tabla 7. Resultados de CheckM de la fusión de ensamblajes.

Linaje marcador	# Genomas	# Marcadores	# Conjuntos de marcadores	Completo	Contaminación
Esfingomonadales	26	569	293	99,59	0,34

Fuente: (Las autoras, 2024).

La información proporcionada por *CheckM* indica que el ensamblaje obtenido es de buena calidad, presenta una alta cobertura (99,59%) (Tabla 7). La herramienta estima la integridad y contaminación, que para el genoma evaluado fue de 0,34%; adicionalmente la herramienta con el uso de genes marcadores específicos determinó que el linaje pertenece al orden Esfingomonadales, al cual pertenece *S. yanoikuyae* (Alonso, 2021). Además, al evaluar el linaje de la bacteria los genes marcadores permiten tener una idea acerca de la calidad e integridad del genoma obtenido; se recomienda agrupar los genes marcadores, dando un total de 293 grupos, ya que la información proporcionada por un solo gen marcador en el genoma no es lo suficientemente minuciosa; además una evaluación de la calidad del genoma es más precisa con múltiples marcadores (Parks et al., 2015). Adicionalmente, los resultados de esta investigación sugieren que el ensamblaje es confiable y puede ser utilizado para análisis posteriores del genoma bacteriano.

## 4.4 Anotación

### 4.4.1 Plataforma KBase

#### 4.4.1.1 Prokka

A través de *Prokka* se obtuvieron 15011 predicciones de genes, codificantes de proteínas, regiones reguladoras y demás elementos genéticos dentro del genoma fusionado ensamblado en KBase. Se encontraron los siguientes resultados:

- Genes predichos: 5045, representa el tamaño del genoma ensamblado y la cantidad de genes que se han identificado en él. Orienta sobre la complejidad y el contenido genético de la bacteria analizada (Stein, 2001).
- Número de genes codificantes de proteínas: 4983, son genes del organismo capaces de sintetizar proteínas con la finalidad de realizar funciones biológicas esenciales (Ejigu & Jung, 2020).
- Número de genes con función no hipotética: 2498, es la cantidad de genes de los cuales ya se sabe su función específica, pues, ya han sido comprobados experimentalmente o se ha especulado con alta confianza (Seemann, 2014).
- Número de genes con número EC: 1118, es un dato que determina el número de genes asociados con enzimas específicas. Tiene alta relevancia para comprender la capacidad metabólica de la bacteria y el potencial para catalizar reacciones químicas específicas (Green, 2005).
- Ontología: 0, este dato representa información sobre ortólogos, estos son genes que se originaron a partir de un único gen ancestral en el último ancestro común de los genomas comparados (Ejigu & Jung, 2020).
- Longitud promedio de la proteína: 321 aa, este dato brinda información sobre la complejidad estructural y funcional de las proteínas producidas por el organismo (Ejigu & Jung, 2020).

De acuerdo con la investigación de Feng et al. (2018), al realizar el análisis genómico de *Sphingobium yanoikuyae* SHJ identificaron un total de 5402 genes. Este dato se puede comparar con el obtenido en este trabajo, que fue de 5045 genes, en donde se aprecia que el valor es bastante cercano, razón por la cual se puede deducir que la calidad del ensamblaje obtenido fue óptima.

La herramienta presenta como resultados una tabla con los elementos genómicos encontrados en los *contigs* de la secuencia. En la Tabla 8 se exponen las 10 primeras predicciones del primer *contig*.

Tabla 8. Anotación con Prokka del genoma ensamblado de *S. yanoikuyae*.

ID de característica	Tipo	Función	Ontología	Inicio	Hebra	Longitud	Contig
00001	gen	proteína hipotética		90	+	2334	1
00001_CDS	CDS	proteína hipotética		90	+	2334	1
00001_mRNA	mRNA			90	+	2334	1
00002	gen	proteína hipotética		2420	+	498	1
00002_CDS	CDS	proteína hipotética		2420	+	498	1
00002_mRNA	mRNA			2420	+	498	1
00003	gen	proteína hipotética		3977	-	1053	1
00003_CDS	CDS	proteína hipotética		3977	-	1053	1
00003_mRNA	mRNA			3977	-	1053	1
00004	gen	L-treonina aldolasa de baja especificidad	ec:4.1.2.48 - L-treonina aldolasa de baja especificidad.	5262	-	1047	1

Fuente: (Las autoras, 2024).

#### 4.4.1.2 RAST

Por medio de *RAST*, para el genoma fusionado ensamblado en KBase se obtuvo:

- Genes codificantes: 5208, es la cantidad de genes identificados que codifican para proteínas. Estos indican la capacidad de la bacteria de interés para producir proteínas específicas con numerosas funciones biológicas (Ejigu & Jung, 2020).
- Repeticiones no codificantes: 65, por el contrario, estas son las regiones identificadas que no codifican para proteínas.

- ARN no codificante: 50, genes de ARN no codificantes incluyen genes para ARN ribosómico (ARNr), ARN de transferencia (ARNt), microARN (miARN), ARN nuclear pequeño y ARN nucleolar (Ejigu & Jung, 2020).
- Recuento total de genes: 8821, es el número total de genes identificados en el genoma ensamblado, tanto codificantes como no codificantes, además, indica el tamaño del genoma y la complejidad genética del organismo (Ejigu & Jung, 2020).
- Recuento de *contigs*: 107, este dato indica la cantidad de *contigs* y refleja la calidad del ensamblaje del genoma (Gregory, 2005).
- Recuento de funciones: 5208
- Contenido de GC: 64,4%, es el porcentaje de guanina o citosina en el genoma. Repercute en la estructura y función de las proteínas codificadas y en la estabilidad del genoma (Babraham Institute, 2024).

La herramienta presenta como resultados una tabla con los elementos genómicos encontrados en el genoma anotado. En la Tabla 9 se muestran los 10 primeros productos codificantes de los *contigs* 1 y 2 del genoma.

En el software *RAST* utilizado para la anotación se obtuvo un número total de genes predichos mucho mayor que en *Prokka*, pues, en el primero se obtuvieron 8821 genes entre codificantes y no codificantes. Se identificaron 4983 genes con capacidad para sintetizar proteínas en *Prokka* mientras que en *RAST* fueron 5208.

Tabla 9. Anotación con RAST del genoma ensamblado de *S. yanoikuyae*.

ID de característica	Tipo	Función	Ontología	Inicio	Hebra	Longitud	Contig
1	gen	FIG00636985: proteína hipotética	SSO:000014469 - FIG00636985: proteína hipotética	90	+	2334	1
10	gen	Permeasa de D-beta-hidroxibutirato	SSO:000012692 - Permeasa de D-beta-hidroxibutirato	8589	+	1461	1
100	gen	Repetición de TPR que contiene proteína exportada. Proteína periplásmica contiene un dominio de proteína prenilitransferasa.		100288	-	942	1
1000	gen	FIG00637072: proteína hipotética		128118	-	558	2
1000_CDS	CDS	FIG00637072: proteína hipotética		128118	-	558	2
1000_mRNA	ARNm			128118	-	558	2
1001	gen	2-poliprenil-3-metil-6-metoxi-1,4-benzoquinol hidroxilasa, tipo coq7	SSO:000034017 - 2-poliprenil-3-metil-6-metoxi-1,4-benzoquinol hidroxilasa, tipo coq7	128691	-	555	2
1001_CDS	CDS	2-poliprenil-3-metil-6-metoxi-1,4-benzoquinol hidroxilasa, tipo coq7	SSO:000034017 - 2-poliprenil-3-metil-6-metoxi-1,4-benzoquinol hidroxilasa, tipo coq7	128691	-	555	2
1001_ARNm	ARNm			128691	-	555	2
1002	gene	Tiol periplásmico: disulfuro oxidoreductasa DsbB, necesario para la reoxidación de DsbA	SSO:000005850 - Tiol periplásmico:disulfuro oxidoreductasa DsbB, necesario para la reoxidación de DsbA	129182	-	495	2

Fuente: (Las autoras, 2024).

## 4.4.2 Plataforma Galaxy

### 4.4.2.1 Prokka

A través de *Prokka* se obtuvo la identificación de genes, secuencias codificantes, regiones reguladoras y demás elementos genéticos dentro del genoma para el ensamblaje fusionado producido en Galaxy que se describen a continuación.

- *Contigs*: 89, este dato denota la calidad y la complejidad del ensamblaje del genoma. La formación de menos *contigs* indica un ensamblaje más completo y contiguo del genoma (Gregory, 2005).
- Bases: 5402608, representa la cantidad total de pares de bases nucleotídicas obtenidas en el ensamblaje (Ejigu & Jung, 2020).
- CDS: 4946, es el número total de genes codificantes de proteínas (Ejigu & Jung, 2020).
- rRNA: 3, ARN no codificante esencial para la función ribosomal en la síntesis de proteínas (Ejigu & Jung, 2020).
- tRNA: 61, ARN no codificante crucial para transportar aminoácidos durante la síntesis de proteínas (Ejigu & Jung, 2020).
- tmRNA: 1, ARN no codificante. Participa en la degradación de proteínas truncadas (Ejigu & Jung, 2020).

En este caso la herramienta presenta como resultados los locus, longitud, genes y el producto codificante. En la Tabla 10 se presentan los 20 primeros locus de la secuencia.

Tabla 10. Resultados de Prokka para la secuencia fusionada en Galaxy.

Etiqueta locus	Tipo	Longitud_pb	Gene	Número EC	COG	Producto
00001	CDS	432				Proteína hipotética
00002	CDS	450				Proteína hipotética
00003	CDS	279				Proteína hipotética
00004	CDS	540	sigF			Factor sigma de ARN polimerasa ECF SigF
00005	CDS	642	nrsf		COG4944	Supuesto factor anti-sigma-F NrsF
00006	CDS	270				Proteína hipotética
00007	CDS	831				Proteína hipotética
00008	CDS	777				Proteína hipotética
00009	CDS	888				Proteína hipotética
00010	CDS	348				Proteína hipotética
00011	CDS	2601	rscC_1	2.7.13.3		Sensor histidina quinasa RcsC
00012	CDS	375				Proteína hipotética
00013	CDS	282				Proteína hipotética
00014	CDS	243				Proteína hipotética
00015	CDS	192				Proteína hipotética
00016	CDS	768				Proteína hipotética
00017	CDS	447				Proteína hipotética
00018	CDS	477	ibpA_1		COG0071	Pequeña proteína de choque térmico IbpA
00019	CDS	567				Proteína hipotética
00020	CDS	732	rsmA_1	2.1.1.182		Subunidad pequeña del ARN ribosómico metiltransferasa A

Fuente: (Las autoras, 2024).

Mediante las anotaciones del genoma realizadas tanto en *Prokka* como en *RAST* (en KBase y Galaxy), se identificó el gen *xylE* y el gen *bphC* entre los elementos genómicos anotados en *Prokka* (en KBase y Galaxy). La relevancia de estos genes está en la competencia para codificar las enzimas catecol 2,3-dioxigenasa y 2,3 -dihidroxibifenil 1,2-dioxigenasa en la vía catabólica de PAH de *S. yanoikuyae* B1. Ambos genes están involucrados en la degradación de hidrocarburos y otros grandes contaminantes ambientales (Cunliffe & Kertesz, 2006; Saunders et al., 1996).

## 5 Conclusiones

Se generó un *pipeline* que permite el ensamblaje a partir de lecturas crudas provenientes del SRA de NCBI, de la especie *Sphingobium yanoikuyae*, obtenidas a través de secuenciación PACBIO e Illumina.

Se evaluó la calidad de las lecturas crudas mediante *FASTQC* con la finalidad de encontrar los errores que contenían la mismas y posterior a ello se preprocesaron las secuencias con la herramienta *Trimmomatic*, generando lecturas con mejor calidad para los posteriores análisis.

Se ensamblaron las lecturas de Illumina con el ensamblador *Shovill* que permitió utilizar cuatro herramientas: *Skesa*, *Megahit*, *Velvet* y *SPAdes*; este último logró un ensamblaje con una longitud del genoma de 5452823 pb y 93 *contigs*. Sin embargo, no se recomienda el uso del ensamblador *Shovill*, pues no se obtuvo un correcto y preciso ensamblaje, presentando significativos errores de montaje.

Se realizó una fusión de ensamblajes con secuencias cortas (Illumina R1 y R2) y largas (PACBIO), mediante el software *Unicycler*, obteniéndose el mejor ensamblaje en la plataforma Galaxy, con una longitud total de 5402608 pb y 89 *contigs*. El ensamblaje híbrido fue anotado con la herramienta *Prokka* con lo que se identificó y asignó funciones a los elementos genéticos dentro del genoma de *Sphingobium yanoikuyae*.

## 6 Bibliografía

- Aguilar-Bultet, L., & Falquet, L. (2015). Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos. *Rev. Salud Anim*, 37(2).
- Al-Maeni, M. A. R., & Al-Khazraji, S. F. R. (2021). Bioinformatics Analyses of the Next Generation Sequencing: A Review. *ACE Journal of Research Studies in Biosciences*, 1(1).  
[https://www.researchgate.net/publication/358742151\\_Bioinformatics\\_Analyses\\_of\\_the\\_Next\\_Generation\\_Sequencing\\_A\\_Review#fullTextFileContent](https://www.researchgate.net/publication/358742151_Bioinformatics_Analyses_of_the_Next_Generation_Sequencing_A_Review#fullTextFileContent)
- Alonso, M. (2021). *Caracterización genómica de una colección de plásmidos accesorios presentes en la bacteria simbiote de alfaalda, E. Meliloti* [Universidad Nacional de La Plata]. <http://sedici.unlp.edu.ar/handle/10915/163286>
- Alquicira, J. (2017). *Genoma de referencia*. Conogasi.  
<https://conogasi.org/diccionario/genoma-de-referencia/>
- Atxaerandio-Landa, A., Arrieta-Gisasola, A., Laorden, L., Bikandi, J., Garaizar, J., Martinez-Malaxetxebarria, I., & Martinez-Ballesteros, I. (2022). A Practical Bioinformatics Workflow for Routine Analysis of Bacterial WGS Data. *Microorganisms*, 10(12).  
<https://doi.org/10.3390/microorganisms10122364>
- Babraham Institute. (2024). *FastQC*.  
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- BIOREN-UFRO. (2024). *Preparación y Control de Calidad*.  
<https://bioren.ufro.cl/servicios/servicios-bioinformatica/>
- Blanco, A. (2013). *Implementación de algoritmos de ensamblaje de genomas en sistemas de memoria compartida y memoria distribuida* [Universidad Politécnica de Madrid].  
[https://oa.upm.es/21934/1/TESIS\\_MASTER\\_ADOLFO\\_BLANCO\\_DIEZ.pdf](https://oa.upm.es/21934/1/TESIS_MASTER_ADOLFO_BLANCO_DIEZ.pdf)

- Bolger, A. M. L. M. U. B. (2014). *Trimmomatic: A flexible read trimming tool for Illumina NGS data*. <http://www.usadellab.org/cms/?page=trimmomatic>
- Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., Stevens, R., Vonstein, V., Wattam, A. R., & Xia, F. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, 5. <https://doi.org/10.1038/srep08365>
- Carrillo, A. (2023). *Lo Esencial de la Línea de Comandos de GNU/Linux*. <https://sites.google.com/ciencias.unam.mx/acl/en-desarrollo>
- Cruz Cubas, A., & Rolland Burger, L. (2013). La ciencia del genoma. *Anales de La Facultad de Medicina*, 63(4). <https://doi.org/10.15381/anales.v63i4.1509>
- Cunliffe, M., & Kertesz, M. A. (2006). Effect of *Sphingobium yanoikuyae* B1 inoculation on bacterial community dynamics and polycyclic aromatic hydrocarbon degradation in aged and freshly PAH-contaminated soils. *Environmental Pollution*, 144(1), 228–237. <https://doi.org/10.1016/j.envpol.2005.12.026>
- Duhan, A., Bhatti, P., Pal, A., Parshad, J., Kumar Beniwal, R., Verma, D., & Bir Yadav, D. (2023). Potential role of *Pseudomonas fluorescens* c50 and *Sphingobium yanoikuyae* HAU in enhancing bioremediation of persistent herbicide atrazine and its toxic metabolites from contaminated soil. *Total Environment Research Themes*, 6. <https://doi.org/10.1016/j.totert.2023.100052>
- Ejigu, G. F., & Jung, J. (2020). Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. *Biology*, 9(9), 295. <https://doi.org/10.3390/biology9090295>

- Feng, L., Liu, H., Cheng, D., Mao, X., Wang, Y., Wu, Z., & Wu, Q. (2018). Characterization and Genome Analysis of a Phthalate Esters-Degrading Strain *Sphingobium yanoikuyae* SHJ. *BioMed Research International*, 2018, 1–8. <https://doi.org/10.1155/2018/3917054>
- Gai, Z., Wang, X., Tang, H., Tai, C., Tao, F., Wu, G., & Xu, P. (2011). Genome sequence of *Sphingobium yanoikuyae* XLDN2-5, an efficient carbazole-degrading strain. In *Journal of Bacteriology* (Vol. 193, Issue 22). <https://doi.org/10.1128/JB.06050-11>
- Galaxy. (2024, January 12). *porechop*. [https://usegalaxy.eu/?tool\\_id=toolshed.g2.bx.psu.edu%2Frepos%2Ffiuc%2Fporechop%2Fporechop%2F0.2.4%2Bgalaxy0&version=latest](https://usegalaxy.eu/?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Ffiuc%2Fporechop%2Fporechop%2F0.2.4%2Bgalaxy0&version=latest)
- Gladman, Simon. (2019). *De Novo Genome Assembly for Illumina Data*. Melbourne Bioinformatics. <https://www.melbournebioinformatics.org.au/tutorials/tutorials/assembly/assembly/>
- Global Biodiversity Information Facility. (2023). *Sphingobium yanoikuyae* (Yabuuchi et al., 1990) Takeuchi et al., 2001. GBIF Backbone Taxonomy. <https://www.gbif.org/es/species/3221339>
- Green, M. L. (2005). Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Research*, 33(13), 4035–4039. <https://doi.org/10.1093/nar/gki711>
- Gregory, S. G. (2005). *Contig Assembly*. In *Encyclopedia of Life Sciences*. Wiley. <https://doi.org/10.1038/npg.els.0005365>
- Gualdrón, J. (2022). *Ensamblaje y anotación del genoma de Vibrio spp. a partir de datos de secuenciación Nanopore (ONT) e Illumina*. . Pontificia Universidad Católica del Ecuador.

- Guo, Y., Ye, F., Sheng, Q., Clark, T., & Samuels, D. C. (2014). Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics*, 15(6), 879–889. <https://doi.org/10.1093/bib/bbt069>
- Gurevich, A. (2021, May 17). *QUAST*. <https://github.com/ablab/quast/blob/master/README.md>
- Illumina. (2021). *MiSeq<sup>TM</sup> System*. Illumina, Inc. <https://emea.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/miseq-data-sheet-m-gl-00006-translations/miseq-system-specification-sheet-m-gl-00006-esp-view.pdf>
- INABIO. (2019). *En el 2021, se ha aumentado al 60% el porcentaje de fuentes de contaminación de la industria hidrocarburífera eliminadas remediadas y avaladas por la Autoridad Ambiental Nacional*. INABIO. <http://inabio.biodiversidad.gob.ec/2019/02/03/10-1-en-el-2021-se-ha-aumentado-al-60-el-porcentaje-de-fuentes-de-contaminacion-de-la-industria-hidrocarburifera-eliminadas-remediadas-y-avaladas-por-la-autoridad-ambiental-nacional/>
- Katta, M. A. V. S. K., Khan, A. W., Doddamani, D., Thudi, M., & Varshney, R. K. (2015). NGS-QCbox and Raspberry for Parallel, Automated and Rapid Quality Control Analysis of Large-Scale Next Generation Sequencing (Illumina) Data. *PLOS ONE*, 10(10), e0139868. <https://doi.org/10.1371/journal.pone.0139868>
- Kumar, G., Ertel, A., Feldman, G., Kupper, J., & Fortina, P. (2020). ISeqQC: A tool for expression-based quality control in RNA sequencing. *BMC Bioinformatics*, 21(1). <https://doi.org/10.1186/s12859-020-3399-8>

- Kundu, R., Casey, J., & Sung, W.-K. (2019). HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies. *BioRxiv*.  
<https://www.biorxiv.org/content/10.1101/2019.12.19.882506v1.abstract>
- Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). *MEGAHIT*: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, *31*(10), 1674–1676.  
<https://doi.org/10.1093/bioinformatics/btv033>
- Li, X., Rehehan, A., Wu, W., Wang, D., Wang, J., Jia, Y., & Yan, Y. (2020). The genome analysis of halotolerant *Sphingobium yanoikuyae* YC-XJ2 with aryl organophosphorus flame retardants degrading capacity and characteristics of related phosphotriesterase. *International Biodeterioration and Biodegradation*, *155*.  
<https://doi.org/10.1016/j.ibiod.2020.105064>
- MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, *5*. <https://doi.org/10.3389/fgene.2014.00013>
- Marínez, E. (2023). *Desarrollo de un pipeline para el análisis de las firmas mutacionales en muestras de pacientes con anemia de Fanconi* [Universidad Nacional Autónoma de México].  
<https://ru.dgb.unam.mx/bitstream/20.500.14330/TES01000834996/3/0834996.pdf>
- Ministerio para la Transición Ecológica y el Reto Demográfico. (2022). *HIDROCARBUROS AROMÁTICOS POLICÍCLICOS (HAP)*. Plan de Recuperación, Transformación y Resiliencia. <https://prtr-es.es/Hidrocarburos-Aromaticos-Policiclicos-HAP,15659,11,2007.html>

Mitra, M., Nguyen, K. M. A. K., Box, T. W., Gilpin, J. S., Hamby, S. R., Berry, T. L., & Duckett, E. H. (2020). Isolation and characterization of a novel *Sphingobium yanoikuyae* strain variant that uses biohazardous saturated hydrocarbons and aromatic compounds as sole carbon sources. *F1000Research*, 9. <https://doi.org/10.12688/f1000research.25284.1>

NCBI. (2017). *Sphingobium yanoikuyae*. NCBI. <https://www.ncbi.nlm.nih.gov/datasets/taxonomy/13690/>

NCBI. (2023a). *Whole genome sequencing of Sphingobium yanoikuyae HAMBI\_1842; long read* (SRR27033679). SRR27033679. [https://trace.ncbi.nlm.nih.gov/Traces/?view=run\\_browser&acc=SRR27033679&display=metadata](https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR27033679&display=metadata)

NCBI. (2023b). *Whole genome sequencing of Sphingobium yanoikuyae HAMBI\_1842; short read* (SRR27033680). SRR27033680. [https://trace.ncbi.nlm.nih.gov/Traces/?view=run\\_browser&acc=SRR27033680&display=metadata](https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR27033680&display=metadata)

Ní Chadhain, S. M., Moritz, E. M., Kim, E., & Zylstra, G. J. (2007). Identification, cloning, and characterization of a multicomponent biphenyl dioxygenase from *Sphingobium yanoikuyae* B1. *Journal of Industrial Microbiology and Biotechnology*, 34(9). <https://doi.org/10.1007/s10295-007-0235-3>

Oxford Nanopore Technologies. (2024a). *How nanopore sequencing works*. <https://nanoporetech.com/platform/technology>

Oxford Nanopore Technologies. (2024b). *What is assembly polishing (consensus improvement)?*

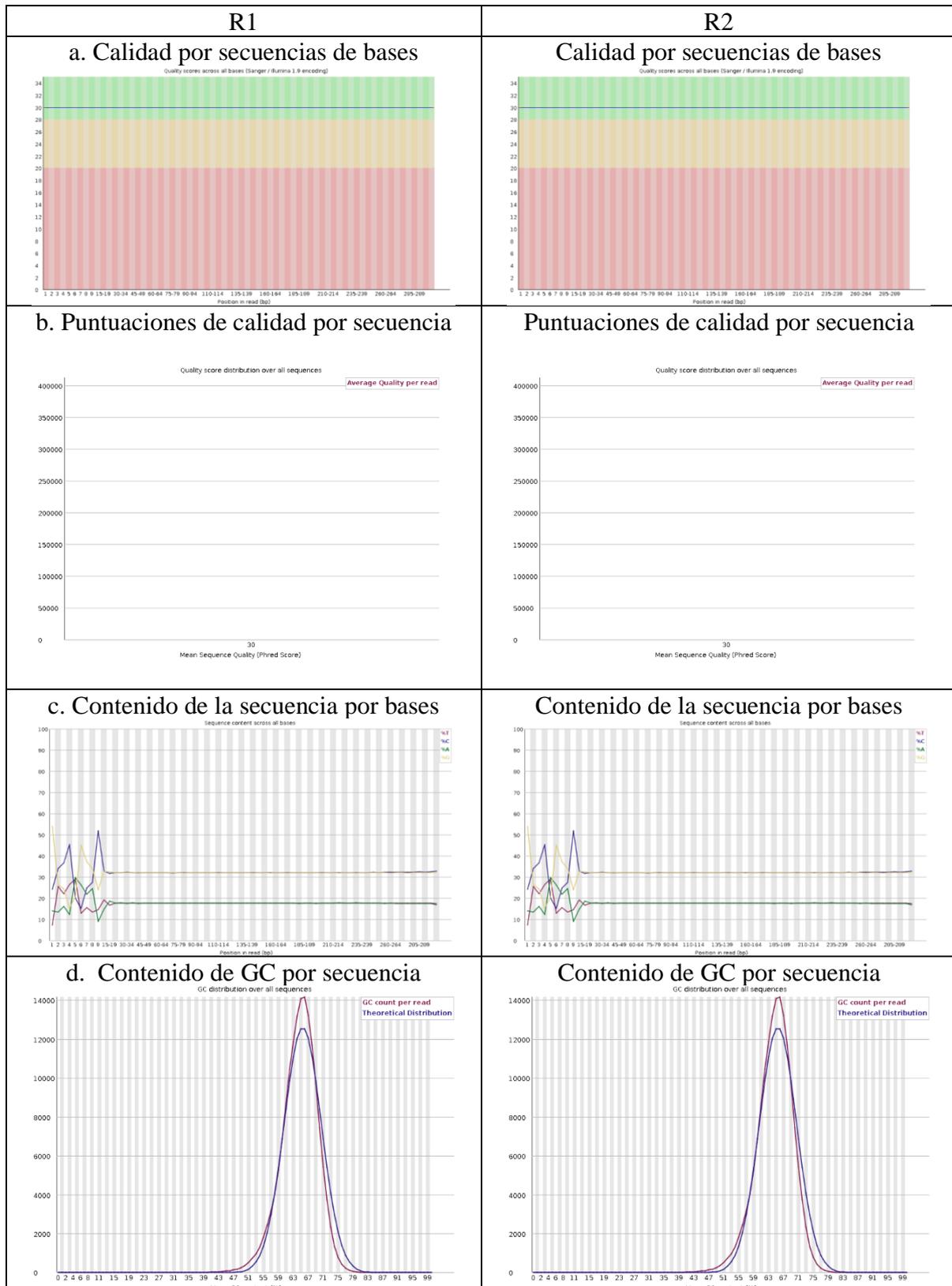
- PacBio. (2024). *LATEST SYSTEM RELEASE Sequel IIe*. PacBio.  
<https://www.pacb.com/technology/hifi-sequencing/sequel-system/latest-system-release/>
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7), 1043–1055.  
<https://doi.org/10.1101/gr.186072.114>
- PRINSEQ. (2024). *Control de calidad y preprocesamiento de datos fácil y rápido*.  
<https://prinseq.sourceforge.net/index.html>
- Prijbelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, 70(1).  
<https://doi.org/10.1002/cpbi.102>
- Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1). <https://doi.org/10.1186/1471-2164-13-341>
- Sánchez, E., Oluyomi, T., Bustos, P., Mendoza, C. P., Mendoza-Herrera, A., & Guo, X. (2022). Complete Genome Report of a Hydrocarbon-Degrading *Sphingobium yanoikuyae* S72. *Applied Sciences*, 12(12), 6201. <https://doi.org/10.3390/app12126201>
- Saunders, J. R., Pickup, R. W., Morgan, J. A., Winstanley, C., & Saunders, V. A. (1996). XyleE as a marker gene for microorganisms. In *Molecular Microbial Ecology Manual*.  
[https://doi.org/10.1007/978-94-009-0215-2\\_12](https://doi.org/10.1007/978-94-009-0215-2_12)
- Seeman, T. (2020). *Shovill*. GitHub. <https://github.com/tseemann/shovill>

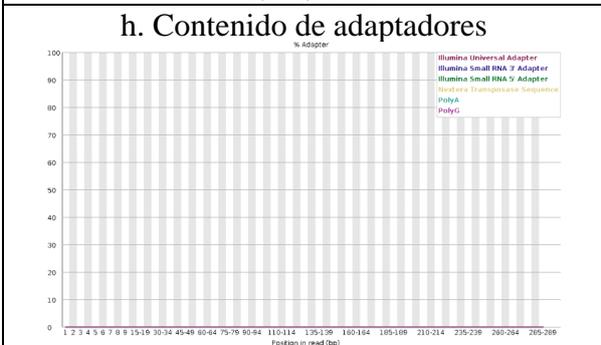
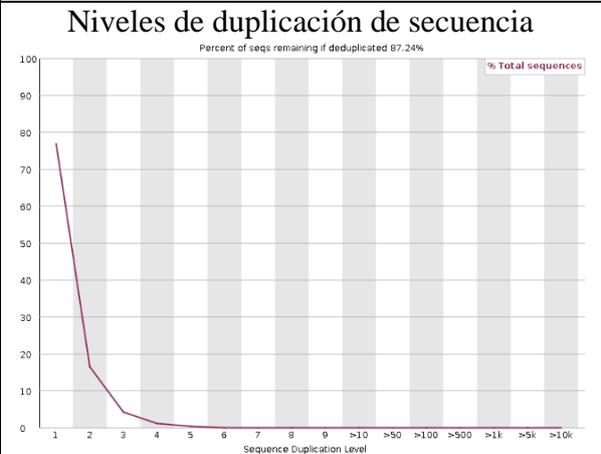
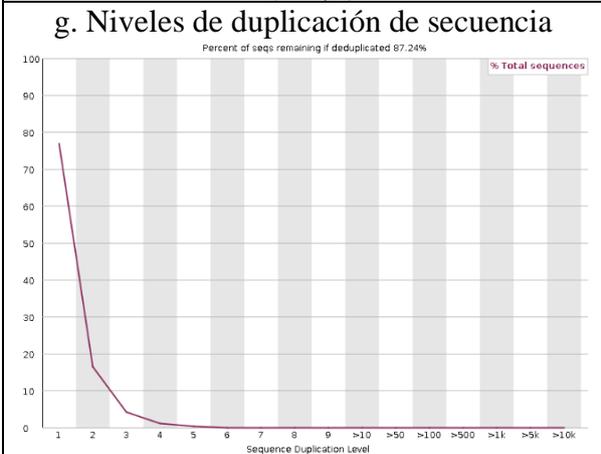
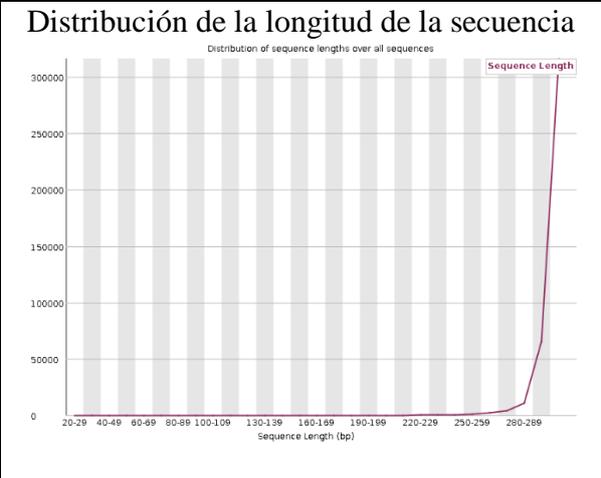
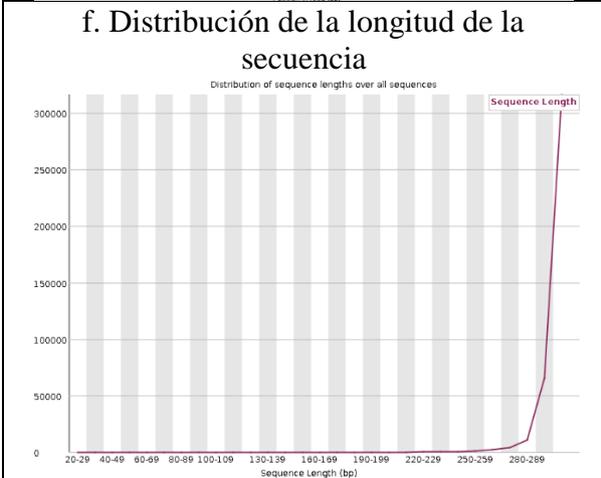
- Seemann, T. (2014). *Prokka*: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>
- Souvorov, A., & Agarwala, R. (2021). SAUTE: sequence assembly using target enrichment. *BMC Bioinformatics*, 22(1), 375. <https://doi.org/10.1186/s12859-021-04174-9>
- Souvorov, A., Agarwala, R., & Lipman, D. J. (2018). SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biology*, 19(1), 153. <https://doi.org/10.1186/s13059-018-1540-z>
- Stein, L. (2001). Genome annotation: from sequence to biology. *Nature Reviews Genetics*, 2(7), 493–503. <https://doi.org/10.1038/35080529>
- Sun, K. (2020). Ktrim: an extra-fast and accurate adapter- and quality-trimmer for sequencing data. *Bioinformatics*, 36(11), 3561–3562. <https://doi.org/10.1093/bioinformatics/btaa171>
- Thermo Fisher Scientific. (2024). *Ion Torrent*. <https://www.thermofisher.com/ec/en/home/brands/ion-torrent.html>
- Thrash, A., Hoffmann, F., & Perkins, A. (2020). Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics*, 21(S4), 249. <https://doi.org/10.1186/s12859-020-3382-4>
- Vaca, I. (2024). *Análisis de expresión génica en genes del maní (Arachis hypogaea L.), relacionados a reacciones alérgicas en el ser humano*. [PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR]. <https://repositorio.puce.edu.ec/server/api/core/bitstreams/7cd56ba5-7150-4d57-834c-30c6b15fed7b/content>

- Vera, F. (2014). *LPS: un algoritmo de ensambles de secuencias cortas de ADN* [Instituto Nacional de Astrofísica, Óptica y Electrónica]. <https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/164/1/VeraVF.pdf>
- Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., ... Haussler, D. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature*, *604*(7906), 437–446. <https://doi.org/10.1038/s41586-022-04601-8>
- Wang, Y., Liu, H., Peng, Y., Tong, L., Feng, L., & Ma, K. (2018). Characterization of the diethyl phthalate-degrading bacterium *Sphingobium yanoikuyae* SHJ. *Data in Brief*, *20*, 1758–1763. <https://doi.org/10.1016/j.dib.2018.09.033>
- Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). *Unicycler*: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, *13*(6), e1005595. <https://doi.org/10.1371/journal.pcbi.1005595>
- Zerbino, D. (2008). *Velvet Manual - version 1.1*.
- Zhao, Q., Hu, H., Wang, W., Peng, H., & Zhang, X. (2015). Genome sequence of *Sphingobium yanoikuyae* B1, a polycyclic aromatic hydrocarbon-degrading strain. *Genome Announcements*, *3*(1). <https://doi.org/10.1128/genomeA.01522-14>

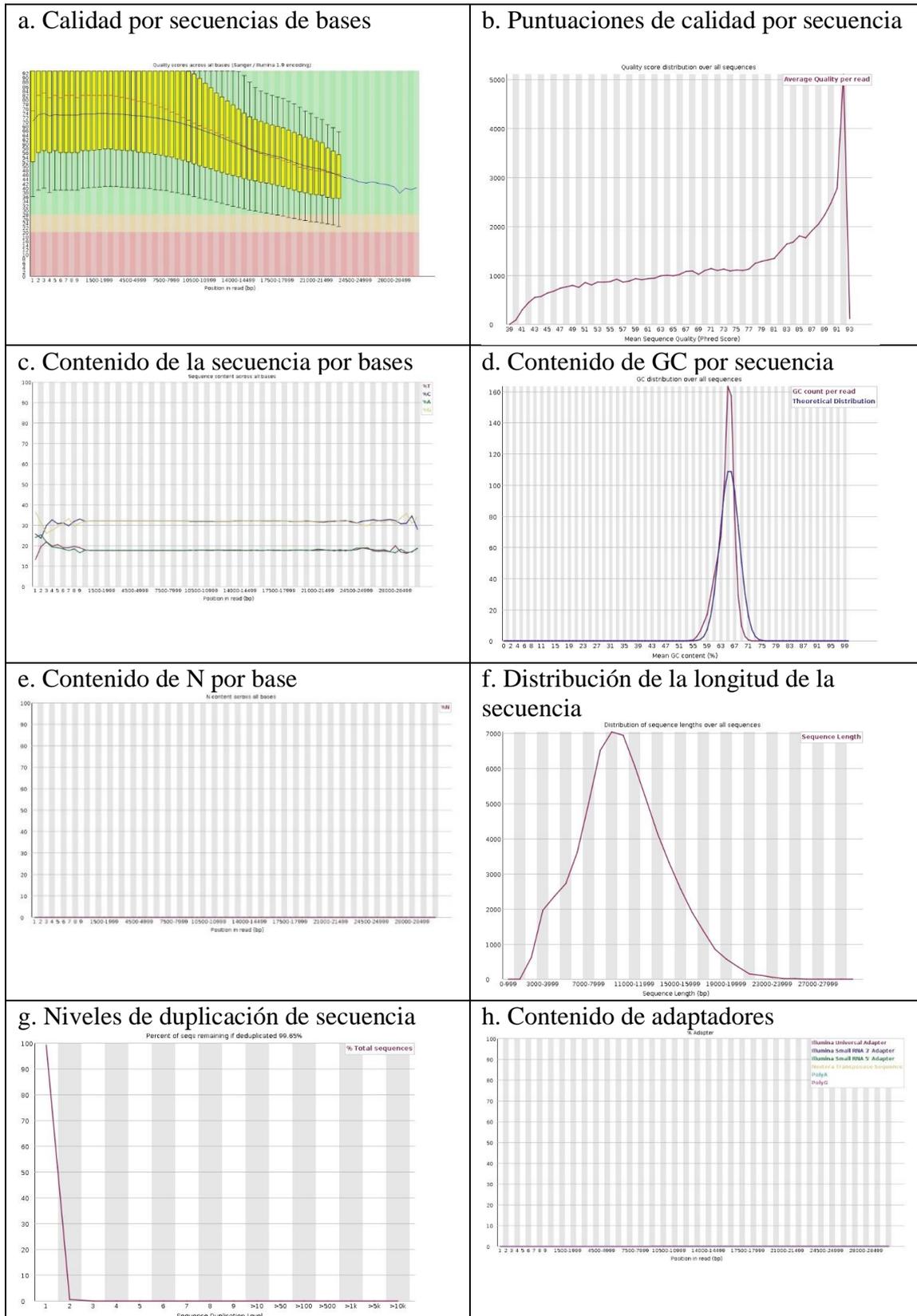
## 7 Anexos

Anexo 1. Resultados de FASTQC de las secuencias crudas de Illumina R1 y R2.





Anexo 2. Resultados de FASTQC de la secuencia cruda de PACBIO.



Anexo 3. Resultados de FASTQC de las secuencias de Illumina R1 y R2 antes y después de Trimmomatic

Comparación general de los resultados obtenidos de FASTQC	
<b>Summary</b>	<b>Summary</b>
 <a href="#">Basic Statistics</a>	 <a href="#">Basic Statistics</a>
 <a href="#">Per base sequence quality</a>	 <a href="#">Per base sequence quality</a>
 <a href="#">Per tile sequence quality</a>	 <a href="#">Per tile sequence quality</a>
 <a href="#">Per sequence quality scores</a>	 <a href="#">Per sequence quality scores</a>
 <a href="#">Per base sequence content</a>	 <a href="#">Per base sequence content</a>
 <a href="#">Per sequence GC content</a>	 <a href="#">Per sequence GC content</a>
 <a href="#">Per base N content</a>	 <a href="#">Per base N content</a>
 <a href="#">Sequence Length Distribution</a>	 <a href="#">Sequence Length Distribution</a>
 <a href="#">Sequence Duplication Levels</a>	 <a href="#">Sequence Duplication Levels</a>
 <a href="#">Overrepresented sequences</a>	 <a href="#">Overrepresented sequences</a>
 <a href="#">Adapter Content</a>	 <a href="#">Adapter Content</a>

Anexo 4. Resultados de FASTQC de la secuencia de PACBIO antes y después de correr Trimmomatic.

Comparación general de los resultados obtenidos de FASTQC	
<b>Summary</b>	<b>Summary</b>
 <a href="#">Basic Statistics</a>	 <a href="#">Basic Statistics</a>
 <a href="#">Per base sequence quality</a>	 <a href="#">Per base sequence quality</a>
 <a href="#">Per sequence quality scores</a>	 <a href="#">Per sequence quality scores</a>
 <a href="#">Per base sequence content</a>	 <a href="#">Per base sequence content</a>
 <a href="#">Per sequence GC content</a>	 <a href="#">Per sequence GC content</a>
 <a href="#">Per base N content</a>	 <a href="#">Per base N content</a>
 <a href="#">Sequence Length Distribution</a>	 <a href="#">Sequence Length Distribution</a>
 <a href="#">Sequence Duplication Levels</a>	 <a href="#">Sequence Duplication Levels</a>
 <a href="#">Overrepresented sequences</a>	 <a href="#">Overrepresented sequences</a>
 <a href="#">Adapter Content</a>	 <a href="#">Adapter Content</a>

Anexo 5. Resultado de las secuencias de Illumina recortadas.



Anexo 6. Resultados de la secuencia de PACBIO recortada.

