Universidad de Concepción
Dirección de Postgrado
Facultad de Ciencias Biológicas - Programa de Magíster en Bioquímica y Bioinformática

Identificación de genes implicados en la biosíntesis de antibióticos en dos genomas bacterianos de sedimentos anóxicos

Tesis para optar al grado de Magíster en Bioquímica y Bioinformática

ALEXIS ARMANDO FONSECA POZA
CONCEPCIÓN-CHILE
2016

Profesor Guía: Alexis Salas Burgos
Dpto. de Farmacología, Facultad de Ciencias Biológicas
Universidad de Concepción

Esta tesis ha sido realizada en el Departamento de Farmacología de la Facultad Ciencias Biológicas, Universidad de Concepción.

Profesor tutor

Dr. Alexis Salas B.
Facultad de Ciencias Biológicas
Universidad de Concepción

Comisión Evaluadora:

Dr. José Martínez O.
Facultad de Ciencias Biológicas
Universidad de Concepción

Dr. Juan Ugalde C.
Facultad de Medicina
Universidad del Desarrollo

Dr. Danilo Pérez P.
Facultad de Ciencias Biológicas
Universidad de Concepción

Director de Programa

Dra. Violeta Morin M.
Facultad de Ciencias Biológicas
Universidad de Concepción

A mis padres y familia por su apoyo incondicional

Tesis de magíster financiada por:
Beca CONICYT para estudios de magíster nacional, concurso regular (2014-2016)

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

5S : subunidad ribosomal con velocidad de sedimentación de 5 unidades Svedberg.

16S : subunidad ribosomal con velocidad de sedimentación de 16 unidades Svedberg.

23S : subunidad ribosomal con velocidad de sedimentación de 23 unidades Svedberg.

A : adenilación

ACP : Proteína transportadora de acilo

ADN : ácido desoxirribonucleico

ARN : ácido ribonucleico

ARNr : ácido ribonucleico ribosómico

ARNt : ARN de transferencia

AT : aciltransferasa

ASCII : American Standard Code for Information Interchange

Bwasw : Burrows-Wheeler Aligner, Smith-Waterman alignment

C : condensation

CDS : secuencia de ADN codificante

CGBs : clúster de genes biosintético

CP : proteína portadora

DH : deshidrogenasa

ER : enoil reductasa

FR : Forward and Reverse

Frg : fragment

G : guanina

Gb : gigabases (mil millones de bases)

GC : guanina-citosina

GTR : *General Time Reversible*

gyrA : girasa subunidad A

gyrB : girasa subunidad B

Indel : inserción-deleción

Kb : Kilobases (1.000 bases)

KR : ceto reductasa

KS : ceto sintasa

Mb : megabases (millon de bases)

MDA **:** Multiple displacement amplification

MRSA **:** Methicillin resistant Staphylococcus aureus

MS : metabolito secundario

N50 : tamaño mínimo del contig que contiene el 50% de los pares de bases de un ensamblaje, si se ordenarán de mayor a menor y se suman las bases.

NGS : Next Generation Sequencing

NRPS : Péptido sintasa no ribosomal

NTC : No template control

OLC : *overlap layout consensus*

OS : Sistema operativo

Pb : Pares de bases

PBS : tampón fosfato salino

pHMM : profile hidden Markov model

PKS : Policétido sintasa

PCR : reacción en cadena de la polimerasa

PCP : proteína portadora de peptidil

sff : *Standard Flowgram Format*

SH : Sulfureto de Humboldt

SNPs : Polimorfismo de nucleótido único

T : timina

TE : tioesterasa

WGS : Whole genome sequencing

RESUMEN

La creciente necesidad de encontrar nuevas moléculas que permitan enfrentar el aumento en la aparición de bacterias patógenas resistentes a antibióticos, ha llevado a la búsqueda de nuevas fuentes de compuestos. Ambientes marinos como el denominado Sulfureto de Humboldt (SH) y sus organismos representan un gran potencial. En bacterias los genes responsables de la biosíntesis, regulación, resistencia y transporte de metabolitos se codifican con frecuencia en un tramo contiguo del genoma, denominado clúster de genes biosintéticos (CGBs). los CGBSs codifican para enzimas claves en la síntesis de metabolitos secundarios (MSs), como la policétido sintasa (PKS) y péptido sintetasa no ribosomal (NRPS). La presente tesis tiene por objetivo identificar CGBs implicados con la biosíntesis de MSs de tipo antibióticos, por medio de herramientas bioinformáticas, en dos genomas bacterianos del SH. La hipótesis establece que los genomas de *Beggiatoa* sp. HS y *Leptospira* sp., parte de la comunidad del SH, poseen CGBs implicados en biosíntesis de MSs de tipo antibiótico. Para determinar aquello, se aislaron las bacterias, luego se extrajo y amplificó su ADN, y se secuenció por medio de la plataforma de secuenciación 454 GS-FLX de Roche. Los genomas de ambas bacterias se reconstruyeron utilizando el método de ensamblaje *de novo* y se realizó la anotación de genes. Como resultado se obtuvo un genoma de 6,2 Mb para *Beggiatoa* sp. HS y de 6,8 Mb para *Leptospira* sp. Se identificaron 3 CGBs en *Beggiatoa* sp. HS, dos de cuales están implicados en biosíntesis de compuestos tipo Terpeno. En *Leptospira* sp. se identificaron 5 CGBs, de los cuales, cuatro estarían implicados en la biosíntesis de productos tipo PKS. En consecuencia, ambos genomas mantienen CGBs que podrían conducir a biosíntesis de MSs con interés farmacológico.

ABSTRACT

The growing need to find new molecules that allow to deal the increase in the appearance of pathogenic bacteria resistant to antibiotics, has led to the search for new sources of compounds. Marine environments such as the so-called Humboldt Sulphide (HS) and their organisms represent great potential. In bacteria the genes responsible for biosynthesis, regulation, resistance and transport of metabolites are frequently coded in a contiguous stretch of the genome, called biosynthetic gene cluster (CGBs). CGBSs encode key enzymes in the synthesis of secondary metabolites (MSs), such as polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS). The present thesis aims to identify CGBs involved in the biosynthesis of MSs of the antibiotic type, through bioinformatic tools, in two bacterial genomes of SH. The hypothesis states that the genomes of Beggiatoa sp. HS and Leptospira sp., Part of the HS community, have CGBs involved in biosynthesis of antibiotic-type MSs. To determine that, the bacteria were isolated, then extracted and amplified their DNA, and sequenced through the 454 GS-FLX sequencing platform. The genomes of both bacteria were reconstructed using the de novo assembly method and gene annotation was performed. As a result, a genome of 6.2 Mb on size was obtained for Beggiatoa sp. HS and 6.8 Mb for Leptospira sp. Three CGBs were identified in Beggiatoa sp. HS, two of which are involved in biosynthesis of Terpene-like compounds. In Leptospira sp. 5 CGBs were identified, of which four would be involved in the biosynthesis of PKS-type products. Consequently, both genomes maintain CGBs that could lead to biosynthesis of MSs with pharmacological interest.

I. INTRODUCCIÓN

1. Necesidad de nuevos fármacos

Un fenómeno generalizado en el uso de fármacos es el proceso de resistencia, que se manifiesta como una disminución de la afinidad del fármaco por su blanco terapéutico. La resistencia a fármacos es un proceso acelerado por las mutaciones del ADN y por mecanismos de transferencia de genes. En efecto, una de las grandes amenazas hoy en día es la aparición de cientos de importantes especies bacterianas patógenas resistentes a múltiples fármacos (Fernandes, 2015), ejemplo de ello es el gran interés que ha suscitado la bacteria resistente a *methicillin*; *Staphylococcus aureus* (MRSA) y la emergencia de otras bacterias Gram-negativas resistentes (Giske *et al*., 2008). Según Fernandes (2015), las principales causas de este problema, son el indiscriminado uso de fármacos en animales y tratamientos de infecciones humanas, sin previos cultivos de diagnóstico y pruebas de susceptibilidad. Esto ha generado una presión selectiva sobre las comunidades bacterianas, implicando evolución acelerada, y por tanto un aumento en las tasas de resistencia a fármacos, lo que tiene un impacto en todos los aspectos de la medicina moderna y disminuye el rendimiento de muchos tratamientos a infecciones. Lo cual está adicionalmente asociado a enormes costos económicos (van Duin y Paterson, 2016).

1.1 *Productos naturales y organismos marinos como fuente de ellos*

Los productos naturales poseen una enorme diversidad estructural, superior a las bibliotecas de moléculas sintéticas actuales, y continúan inspirando nuevos descubrimientos en la química, biología y medicina. Estos productos están optimizados evolutivamente para actuar de forma similar a fármacos y son importantes fuentes para el descubrimiento de prometedoras moléculas de tipo

antibiótico (Newman y Cragg, 2012; Shen , 2016).

De acuerdo a la revisión publicada por Newman y Cragg (2016) sobre agentes terapéuticos entre 1981 y 2014, se han aprobado un total de 1.562 nuevos fármacos. De los cuales, el 67 (4%) corresponden a productos naturales no modificados, 9 (0.06%) a productos naturales botánicos, 320 (21%) a productos naturales modificados, 162 (10%) a compuestos sintéticos con un farmacóforo derivado de un producto natural, y 420 (27%) a compuestos netamente sintéticos. Desde una perspectiva de sus blancos terapéuticos, 174 están dirigidos hacia el cáncer; siendo 23 (13%) de ellos exclusivamente de origen sintético y el 87% de origen natural, natural modificado o de inspiración natural. Por otra parte, de 141 compuestos antibacterianos registrados hasta el 2014, se consigna que; 112 (80%) tienen un origen exclusivamente natural, modificado o inspirados en productos naturales y 29 (20%) con un origen exclusivamente sintético. En conclusión, la mayoría de los fármacos aprobados son de origen natural o inspirados directamente desde compuestos disponibles en la naturaleza.

Actualmente la mayoría de los fármacos derivados de productos naturales son de origen terrestre. Debido principalmente a la relativa facilidad de acceso, en particular a plantas, y siendo las fuentes microbianas especialmente importantes en el área de antibióticos. Sin embargo, un análisis comparativo de Kong *et al*. (2010) demostró que los productos naturales marinos son superiores a los productos naturales terrestres en términos de novedad química. Además, muestran mayor bioactividad en comparación con organismos terrestres (Montaser y Luesch, 2011). Lo cual quedó en evidencia en un análisis de citotoxicidad preclínica conducido por el Instituto Nacional del Cáncer de Estados Unidos, demostrando que aproximadamente el 1% de las muestras marinas ensayadas poseen potencial antitumoral, frente al 0,1% de las muestras terrestres (Munro *et al*., 1999). Sugiriendo que los organismos marinos representan una gran y aún inexplorada fuente de nuevos compuestos naturales (Imhoff *et al*., 2011; Newman and Cragg, 2016).

Inicialmente las investigaciones en torno al descubrimiento de sustancias naturales desde organismos marinos fueron dirigidas sobre algas e invertebrados. No obstante, hoy el foco se ha trasladado hacia los microorganismos (Molinski *et al*, 2009; Mayer *et al.*, 2010; Xiong *et al*., 2013; Gerwick y Fenner, 2013; Romano *et al*., 2016). La evidencia respecto de la capacidad de síntesis de compuestos por parte de bacterias marinas ha ido en constante aumento (Debnath *et al*., 2007; Gulder y Moore, 2009; Rahman *et al*, 2010; Li *et al*. 2013; Machado *et al*., 2015; Britstein *et al*., 2015), y sólo entre 1997 y 2008 se han identificado 659 nuevos compuestos. Originados principalmente a partir de microorganismos pertenecientes a *Actinobacteria* (40%), Cianobacteria (33%) y *Proteobacteria* (12%) (Williams, 2009). Una muestra de la capacidad que poseen los microorganismos marinos quedó en evidencia tras la secuenciación del Actinomiceto *Salinispora tropica* en el año 2007, la cual reveló que el ~10% de su genoma está dedicado a la biosíntesis de productos naturales (Udwary *et al*., 2007). Otros ejemplos de esta capacidad, son el descubrimiento de cuatro nuevos tipos de compuestos; tres macrólidos citotóxicos denominados *levantilida A, B* y C, producidos por cepas de Micromonospora de aguas profundas (A y B) (Gärtner *et al*., 2011) y del Golfo de Corcovado en Chiloé (Fei *et al*., 2013), y *Abisomicina C* (Figura 1), producida por *Verrucosispora maris*,

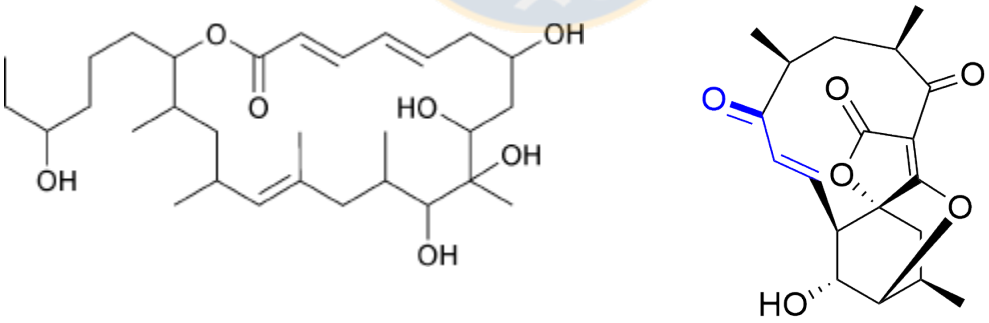


Figura 1. Estructura central del antibiótico antitumoral levantilida C y el antibiótico abisomicina C.

La figura del lado izquierdo muestra la estructura del antitumoral *levantilida* C y la figura del lado derecho la del antibiótico *Abisomicina C.* Fuente: Blunt *et al*., 2015.

una especie de actinomycete recolectada desde sedimentos del mar de Japón (Keller *et al*., 2007) y con potente actividad antimicrobiana, en especial contra MRSA.

Estos antecedentes evidencian la gran capacidad que tendrían las bacterias marinas para producir nuevos compuestos de tipo metabolito secundario (MS), lo que representa un gran potencial como fuente de nuevas moleculas con interes formacologico.

2. Clúster de genes codificantes de enzimas implicadas en la biosíntesis de metabolitos secundarios

La diversidad de MSs conocidos incluyen entre sus principales representantes a compuestos pertenecientes a Policétidos Sintetasas (PKSs), Péptidos Sintetasas No Ribosomales (NRPSs), terpenos, aminoglucósidos, aminocumarinas, indolocarbazoles, lantibióticos, bacteriocinas, nucleósidos, betalactámicos, butirolactonas, sideróforos y melaninas.

En bacterias, la biosíntesis, regulación y transporte de un MSs es controlado por genes que se codifican en un tramo contiguo del genoma, denominado clúster de genes biosintético (CGB). Los CGBs que codifican para enzimas implicadas en vías de generación de un producto natural, generalmente no son esenciales para el crecimiento celular en condiciones ideales y representan elementos genéticos altamente adaptables que evolucionan a través de mutación genética, duplicación de genes, deleción, migración y reordenamientos del genoma en períodos de deriva genética y de selección natural (Wink, 2003; Jenke-Kodama *et al*., 2008; Medema *et al*., 2010).

La Figura 2 esquematiza la arquitectura de dos CGBs implicados en la biosíntesis de un compuesto tipo aril polieno en *Escherichia coli* CFT073 y *Vibrio fischeri* ES114. Estos compuestos son muy similares a carotenoides y el CGB que codifica para las enzimas responsables de su biosíntesis está muy extendido en todas clase de bacterias (Schoner *et al*., 2016).

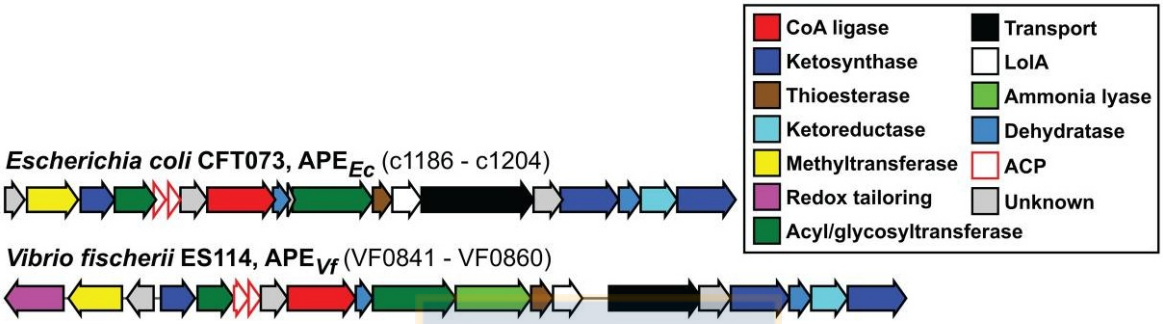


Figura 2. Arquitectura de dos CGBs implicados en la biosíntesis de un compuesto tipo aril polieno en *Escherichia coli* CFT073 y *Vibrio fischeri* ES114. Los segmentos coloreados representan genes biosintéticos, como Ceto Sintasa (azul), genes de transporte (negro) o genes desconocidos (gris). Fuente: Cimermancic *et al*., 2014.

Generalmente los MSs comprenden diversas fracciones químicas, tales como cadenas principales de policétidos, derivados de aminoácidos y azúcares. Las principales enzimas implicadas en biosíntesis de MSs son la policétido sintasa (PKS), la cual puede ser multimodular (tipo I) y con múltiples dominios activos, y la péptido sintetasa no ribosomal (NRPS). Las enzimas responsables de la síntesis de otros compuestos constitutivos, tales como azúcares, generalmente son codificados por genes adyacentes a los genes que codifican para PKSs/NRPSs. A través de eventos como; glicosilación, alquilación y oxidación, se generan diversos y complejos metabolitos. Además, la producción y transporte de MSs está estrictamente controlado por reguladores transcripcionales (Park *et al*., 2010). Como resultado,

todo el conjunto de genes responsables de la biosíntesis de un MS está codificado en un gran grupo de genes que puede abarcar desde 10 hasta 100 Kb.

Los MSs se ensamblan secuencialmente a partir de bloques simples tales como acil-CoA y aminoácidos cuya extensión es catalizada por un conjunto de dominios funcionales, los que colectivamente se denominan módulos, codificados en un PKS/NRPS. Habitualmente en una PKS funcionan como mínimo, un dominio ceto sintasa (KS), proteína portadora de acilo (ACP) y aciltransferasa (AT) (Ichikawa *et al*., 2013).

La Figura 3 muestra 6 sucesivos módulos PKSs tipo I que actúan sucesivamente en la elongación, procesamiento y terminación de una cadena de policétido. Estos módulos están constituidos por diferentes dominios (representaciones esféricas coloreadas) en 4 enzimas PKS tipo I, y que en conjunto generan el antibiótico denominado Picromicina, el cual fue el primer antibiótico macrólido conocido y recientemente sintetizado en su totalidad (Kang, 2012). Por otra parte, las NRPSs son enzimas multimodulares que ensamblan péptidos bioactivos a través de un mecanismo de *tiotemplado* (Ichikawa *et al*., 2013). Un NRPS puede integrar aminoácidos proteinogénicos, así como aminoácidos no proteinogénicos en la cadena en crecimiento, lo que contribuye a la diversidad estructural y están relacionadas con la generación de antibióticos como penicilina y vancomicina, y anticancerígenos como bleomicina. Estructuralmente en un NRPS existen módulos con al menos un dominio funcional de condensación (C), adenilación (A) y dominios de proteína portadora de peptidil (PCP) (Tambadou *et al*., 2014).

La especificidad para cada aminoácido iniciador/extensor se determina por los residuos activos en el dominio A, y los aminoácidos cargados se modifican por dominios opcionales, tales como metiltransferasa, epimerización y dominios reductasa (Stachelhaus *et al*., 1999).

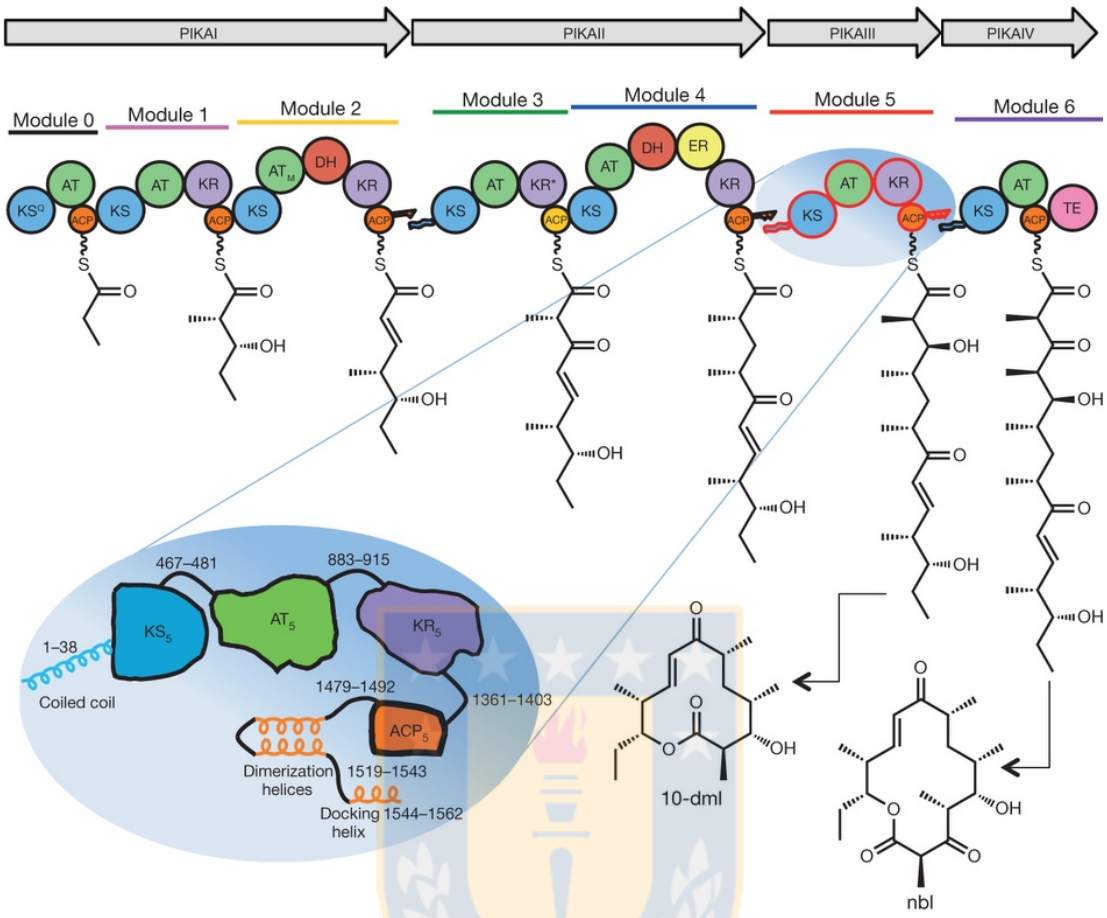


Figura 3. **PKS multimodular tipo I implicado en la biosíntesis del antibiótico picromicina**.

Las flechas grises representan enzimas multimodulares PKS tipo I (PIKAI-IV). Las esferas representan dominios; ceto sintasa (KS), aciltransferasa (AT), proteína transportadora de acilo (ACP), ceto reductasa (KR), deshidrogenasa (DH), enoil reductasa (ER) y tioesterasa (TE) contenidas dentro de los 6 módulos. 10-dml y nbl son productos policétidos entregados por el módulo 5 y 6, respectivamente. El ovalo azul contiene el detalle del docking del módulo multidominio 5, derivado de PIKAIII. Fuente: Dutta *et al*., 2014.

En consecuencia, la identificación de genes que codifican para dominios esenciales de enzimas PKSs, NRPSs, además de aquellas implicadas en biosíntesis de compuestos como terpenos, es fundamental en la búsqueda e identificación de CGBs.

3. Secuenciación de nueva generación y descubrimiento de nuevos fármacos

3.1 *Secuenciación de nueva generación*

La secuenciación de nueva generación, mayormente conocida bajo el acrónimo en inglés de NGS (*Next Generation Sequencing*), busca obtener el orden exacto de ocurrencia de nucleótidos en una cadena de ADN y hace referencia a los avances tecnológicos en la instrumentación de secuenciación de ADN (Raza y Ahmad, 2016). El primer método estándar de secuenciación se consiguió en 1975 por Edward Sanger, denominado secuenciación Sanger, con el cual aproximadamente 30 años más tarde, se produjo el primer gran avance en el campo de la secuenciación con la conclusión del Proyecto Genoma Humano en el año 2003, el cual duró 13 años y costó aproximadamente 3 billones de dólares.

En términos generales la tecnología de NGS extiende la metodología de Sanger a un método masivo y en paralelo, además de reducir constantemente su costo (Goodwin *et al*., 2016). Es por ello que el advenimiento de estas tecnologías ha traído una revolución en áreas tales como la ecología microbiana, medicina, medioambiente, agricultura, alimentos, tecnología, etc. Además, con aplicaciones variadas que incluyen el descubrimiento de patógenos, análisis de metagenomas, microbiomas, perfiles transcriptómicos, diagnóstico de enfermedades infecciosas, etc. (Chiu y Miller, 2016).

El primer sistema NGS de segunda generación disponible fue la plataforma de

secuenciación por pirosecuenciación 454 de Roche (Ronaghi, 2001). Luego emergieron sistemas que incluyen a Illumina (derivado del sistema conocido como Solexa) HiSeq/MiSeq/NextSeq, ABI SOLiD y Life Technologies Ion Torrent (Chiu y Miller, 2016). Sin embargo, hoy en día se están desarrollando tecnologías NGS denominadas de “tercera generación” o *Single-molecule real-time long reads*, tales como Pacific BioSciences RS II (secuencias de ~20 Kb) y Oxford Nanopore MK 1 MinION (Secuencias de hasta 200 Kb) (Goodwin *et al*., 2016).

El tipo de técnicas y tecnologías descritas anteriormente, son fundamentales en proyectos relacionados con la búsqueda de CGBs implicados en biosíntesis de MSs, utilizando métodos bioinformáticos. El tipo de tecnología en particular, dependerá de los objetivos concretos del proyecto. Como podría ser la secuenciación de un genoma bacteriano completo, con el fin de identificar genes de interés.

3.2 *Proyecto de secuenciación de genoma bacteriano completo*

La secuenciación de un genoma bacteriano completo, incluye la secuenciación de su cromosoma y plásmidos en un mismo momento, con la finalidad de determinar variabilidad genética o genes de interés, entre otros objetivos. La irrupción de tecnologías NGS, su constante reducción en costos y aumento en velocidad de secuenciación, ha dado paso a un gran aumento en el número de genomas bacterianos depositados en las bases datos, ya sean borradores de genomas, tambien denominados *drafts* (genomas fragmentados), o genomas finiquitados (genoma sin fragmentación). En consecuencia, el número de genomas bacterianos depositados en GenBank hasta el año 2015 alcanzan a 31.252 genomas. Siendo el filo Proteobacteria el que cuenta con el mayor número de genomas secuenciados (14.268 genomas) seguido por Firmicutes (9.628 genomas) y Actinobacteria (4.059 genomas) (Land *et al*., 2015).

La Figura 4 detalla el flujo general de trabajo de un proyecto de secuenciación de un genoma bacteriano completo y aislado directamente desde el ambiente. Comenzando por la toma de muestras, seguido del trabajo de laboratorio, incluyendo la extracción y fragmentación del ADN, preparación de las bibliotecas y la secuenciación de estas con alguna plataforma adecuada, según sea el propósito del proyecto. Luego, por medio de técnicas bioinformáticas se controla la calidad de secuenciación, se filtran y remueven bases y lecturas no deseadas, todo ello con el objetivo de reconstruir el genoma de la forma más precisa posible, tanto por ensamblaje *de novo* (reconstrucción del genoma sin un genoma de referencia) o mapeamiento (reconstrucción usando un genoma como referencia) (Zerbino *et al*., 2009; Miller, 2010; Compeau *et al*., 2011). Finalmente, se realiza la predicción de genes (Anotación) utilizando bases de datos disponibles, permitiendo identificar genes de interés si fuera el propósito.

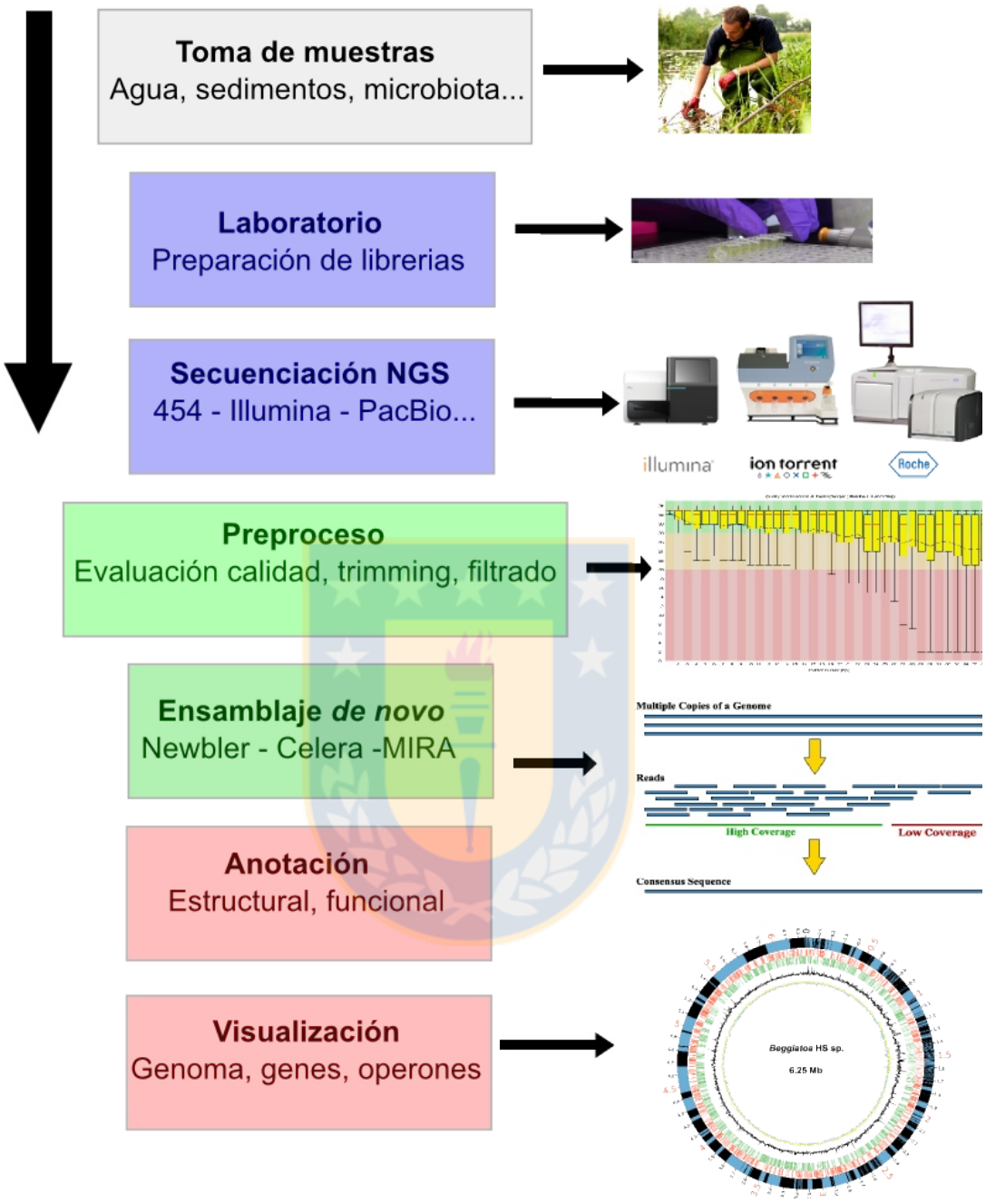


Figura 4. Flujo de trabajo general de un proyecto de secuenciación de un genoma bacteriano completo.

El esquema considera desde la toma de muestras ambientales, hasta el análisis bioinformático que permitirá identificar genes de interés. Fuente: Elaboración propia.

3.3 *Minería de datos de Metabolitos secundarios*

La minería de datos, hace referencia a la determinación de patrones y extracción de conocimientos desde grandes cantidades de datos. En este contexto el descubrimiento de MSs está estrechamente ligada al desarrollo de tecnologías NGS y la eficacia en la identificación *in silico* de objetivos prometedores dentro de los genomas y la gran cantidad de datos de secuencias disponibles. En consecuencia, el continuo y creciente flujo de secuenciación de genomas microbianos ha revelado un potencial biosintético de diversos y nuevos MSs, mucho mayor del que se había pensado (Winter *et al*., 2011; Zotchev *et al*., 2012; Weber y King, 2016).

Las estrategias bioinformáticas para el descubrimiento de CGBs implicados en biosíntesis de productos naturales y en particular de antibióticos se centran en gran medida en búsquedas basadas en homología (Fedorova *et al*., 2012). Por otra parte, a pesar del éxito de las estrategias de bioensayo, basadas en el cultivo tradicional para descubrir nuevos productos naturales, los análisis genéticos han revelado que estos enfoques han proporcionado acceso a sólo una pequeña fracción de la capacidad biosintética codificada en genomas microbianos, y ha quedado en evidencia que la mayoría de las vías biosintéticas no siempre se expresan en condiciones de laboratorio, pasando por alto los productos de estas vías. Además, solo una pequeña fracción de microorganismos son cultivables, ~1% (Weber y King, 2016). Es por ello que la minería genómica busca explotar el potencial oculto de las rutas biosintéticas.

Como se ha mencionado anteriormente, genes de interés corresponden a menudo a aquellos que codifican para enzimas PKSs/NRPSs, los cuales son clave en la síntesis de antibióticos y agentes antitumorales (Zazopoulos *et al*., 2003). Un ejemplo de predicción de MSs por medio de minería de datos genómicos fue el descubrimiento de salinosporamida K (Reed *et al*., 2007), así como también la amplificación de un fragmento de gen *pks*, identificado en una cepa de

Pseudomonas sp. asociada a una esponja marina, el cual género Mupirocina (Imhoff *et al*., 2011) y la síntesis del antibiótico ECO-050, desde la bacteria *Amycolatopsis orientalis* (Banskota *et al*., 2006). En consecuencia, la identificación de genes codificantes de enzimas implicadas en biosíntesis de MSs, por medio de métodos computacionales, puede ser utilizada para dirigir un enfoque de preselección en el aislamiento de microorganismos que probablemente sean capaces de producir algún MS (Schneemann *et al*., 2010).

La anotación estructural y funcional de un genoma es un paso esencial para la identificación de CGBs relacionados con biosíntesis de algún tipo de MS. Fedorova y colaboradores (2012) entrega una buena descripción de herramientas disponibles en la identificación de CGBs, y más recientemente Weber y Kim (2016) han realizado una revisión actualizada de herramientas y portales bioinformáticos que asisten en la tareas de búsqueda, identificación y caracterización de CGBs. En este contexto se han desarrollado diversas herramientas con enfoques basados en búsqueda específica de secuencias de enzimas o dominios, tales como; antiSMASH (Weber *et al*. 2015), el cual es uno de los programas más populares y precisos para la identificación de un amplio rango de CGBs implicados en biosíntesis de compuestos de tipo MS. También existen alternativas como CLUSEAN (Weber *et al*., 2009) (incluido en antiSMASH 3.0), capaz de predecir dominios, predicción de genes y predicción de especificidad de substrato de los posibles CGBs.

Complementariamente existen algunas bases de datos relacionadas con vías metabólicas como KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa y Goto, 2000) (http://www.kegg.jp/) y DoBISCUIT (Database of BIoSynthesis clústers CUrated and InTegrated) (http://www.bio.nite.go.jp/pks/) (Ichikawa *et al*., 2013), la cual es una base de datos curada y actualizada de información sobre CGBs implicados en vías metabólicas de biosíntesis de MSs, que además proporciona descripciones estandarizadas de gen/módulo y dominios relacionados con grupos de genes especializados desde bacterias (Ichikawa *et al*., 2013).

De esta manera, la minería de datos genómicos ha impactado positivamente en la explotación biotecnológica de la microbiota, provocando un creciente interés en el desarrollo de este enfoque en la investigación de productos naturales.

4. Bacterias *Beggiatoa* sp. HS y *Leptospira* sp. como potenciales fuentes de nuevos Metabolitos secundarios

En este contexto de necesidad de nuevas fuentes de MSs de tipo antibiótico, la comunidad bacteriana del denominado “Sulfureto de Humboldt” (SH) frente a las costas centro-norte de Chile (Gallardo *et al*., 2013a) posee un potencial enorme, aún por explorar. Un Sulfureto está definido por una biota bentónica principalmente microbiana, en la cual el principal factor modelador de la estructura comunitaria es el alto nivel de hidrógeno sulfurado, ante la presencia limitada o total ausencia de oxígeno disuelto (Baas-Becking, 1925).

Grandes bacterias sulfuro-oxidantes de los géneros; *Candidatus* Thioploca y *Candidatus* Beggiatoa pueblan los sedimentos del SH, frente a Chile central (Gallardo, 1977; Schulz *et al*, 2001; Salman *et al*., 2011). Las cuales pueden utilizar nitrato u oxígeno para oxidar el sulfuro acumulado de manera intracelular.

Desde que se demostró que el metabolismo en base al azufre tiene una data aproximada de 3,5 mil millones de años, y que aún se mantiene en zonas en las cuales las condiciones ambientales lo permiten, tal como en el SH, es que se piensa que probablemente la comunidad bacteriana de este sulfureto podría existir desde siempre en los océanos del planeta (Gallardo *et al* 2013b). Similitudes morfológicas consistentes entre fósiles del Arcaico y macro-megabacterias filamentosas del SH respaldan esta proposición (Schopf, 2006; Schopf *et al*., 2015).

Mußmann *et al*. (2007) descubrieron que filamentos de *Ca*. Beggiatoa sp.

recolectados desde sedimentos de la bahía de Eckernförde (Alemania), probablemente albergan el potencial para sintetizar MSs. En efecto, se identificaron numerosos genes que codifican para NRPSs y PKSs. Derivados de estos PKSs poseían alta similitud con proteínas implicadas en síntesis de toxinas y antibióticos. Por otra parte, recientemente desde una forma dominante del SH, "*Ca.* Thioploca araucae" (Salman *et al*., 2011) se aíslo un nuevo macrólido bioactivo denominado "Macplocimina A" (Li *et al.*, 2013), el cual es una molécula estructuralmente cercana a lactonas del ácido resorcílico.

Además de las grandes bacterias filamentosas anteriormente mencionadas, otros tipos bacterianos, tales como *Spirocheta,* forman parte de la comunidad del SH. Filo que tiene representantes en ambiente marinos descubiertos recientemente, tal como *Spirochaeta lutea* sp. Nov, aislada desde muestras marinas cerca de India (Shivani *et al*., 2015). Por otra parte, se ha registrado la presencia de genes codificantes de enzimas relacionadas con síntesis de MSs en este filo (Abt *et al*., 2012). No obstante, los registros en la literatura aún son escasos. Además, el conocimiento de la comunidad bacteriana del SH es limitado, aunque resultados previos indican una diversidad enorme.

El presente trabajo de tesis tiene como objetivo la identificación de CGBs implicados con la biosíntesis de MSs de tipo antibiótico, por medio de herramientas bioinformáticas, en dos genomas bacterianos aislados directamente desde los sedimentos anóxicos del SH. Bacterias aquí identificados como *Beggiatoa* sp. HS y *Leptospira* sp. (Perteneciente al filo Spirochaeta).

5. Hipótesis

Las bacterias *Beggiatoa* sp. HS y *Leptospira* sp., parte de la comunidad del SH, poseen CGBs implicados en la biosíntesis de MSs de tipo antibiótico, los cuales pueden ser identificados mediante análisis de secuencias por métodos computacionales.

6. Objetivo general

Identificar por medio de métodos computacionales de análisis de secuencias la presencia de CGBs implicados en vías metabólicas de biosíntesis de MSs de tipo antibiótico, en los genomas de *Beggiatoa* sp. HS y *Leptospira* sp.

7. Objetivos específicos

1) Analizar y pre-procesar las lecturas de ADN genómico generadas por secuenciación de alto rendimiento en la plataforma de secuenciación 454 GS-FLX de Roche.

2) Reconstruir los genomas de *Beggiatoa* sp. HS y *Leptospira* sp., utilizando la estrategia de ensamblaje *de novo*.

3) Realizar la anotación funcional y estructural de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp., con especial atención en la identificación CGBs relacionados en biosíntesis de MSs de tipo antibiótico.

II. MATERIALES Y MÉTODOS

1. Toma de muestras

La toma de muestras se llevó a cabo el 13 de diciembre de 2008 en el punto de muestreo denominado Estación 7 (lat. -36.64, Long. -73,04), situado en la boca de la Bahía de Concepción (Chile central), a 35 m de profundidad (Figura 1). Las muestras de sedimento se tomaron a bordo de la embarcación R/L "Otilia" de 8,2 m y motor-*Whale*, utilizando un dispositivo de toma muestras tipo *mono-corer,* que consta de un tubo de plexiglás de 1 m de largo y 5 cm de diámetro.

Figura 5. Locación desde donde se aislaron las bacterias *Beggiatoa* sp. HS y *Leptospira* sp.

El punto de muestreo denominado Estación 7, está ubicado en la boca de la bahía de Concepción a 35 m de profundidad. Fuente: Elaboración propia.

2. Micromanipulación de las bacterias, amplificación y secuenciación del ADN

La aislación por micromanipulación y amplificación del ADN de las bacterias seleccionados se realizó siguiendo el método de Ishoey (2008), el cual consiste en la amplificación del ADN, con el objetivo de obtener la cantidad de microgramos de ADN requerido como molde, a partir de una única bacteria, mediante un método denominado amplificación de desplazamiento múltiple (MDA) (Shoaib *et al*., 2008). Se aislaron siete filamentos individuales de *Beggiatoa* sp. HS (Figura 2) y uno de tipo *Spirochaeta (Leptospira* sp.). Microcapilares de aproximadamente 10 µm fueron utilizados para los aislamientos.

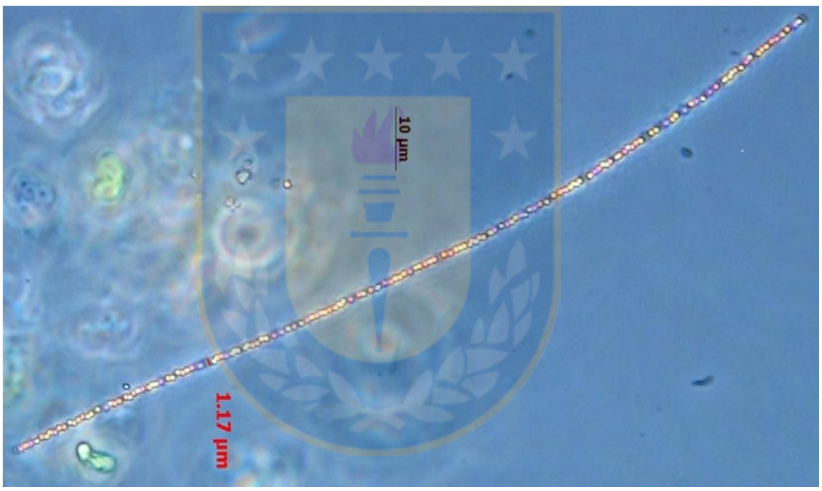


Figura 6. **Microfotografía de un filamento de *Beggiatoa* sp. HS.**
Filamento de la macrobacteria *Beggiatoa* sp. HS de 1,17 µm de diámetro denominado MDA1. Fuente: Elaboración propia.

Luego del aislamiento por micromanipulación, los filamentos se transfirieron a 0.5 µL o 1 µl de PBS estéril en un tubo de PCR de 200 µL para la amplificación de todo el genoma. La amplificación se realizó usando el kit de GenomiPhi HY (GE Healthcare) por el método de lisis alcalina y se terminó después de 6 horas a 30° C.

Los productos del WGA se diluyeron 2 veces en tampón TE (almacenado a -20° C) y una dilución de 20 veces se preparó como solución de trabajo para el análisis de PCR y la cuantificación. La pureza del ADN amplificado fue examinado por PCR/secuenciación del gen 16S de ARNr utilizando cebadores universales (Figura 7).

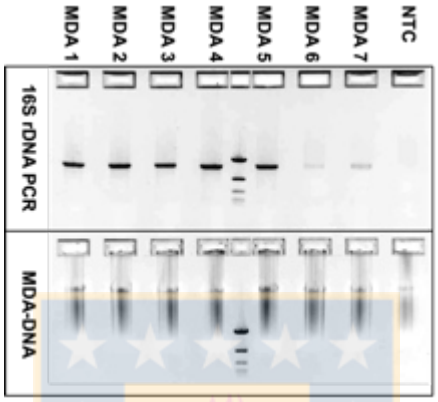


Figura 7. Productos de la amplificación del gen 16S de ADNr y del genoma completo de 7 filamentos bacterianos aislados (MDA 1-7).
La parte superior de la figura muestra los productos de la amplificación del gen 16S de ARNr de los filamentos bacterianos micromanipulados MDA1 a MDA 7. La parte inferior de la figura muestra los productos de la amplificación total del genoma (WGA) de los filamentos MDA1 a MDA 7 y el control sin molde (NTC). Todos los productos se visualizaron por electroforesis en gel de agarosa. Fuente: Elaboración propia.

El análisis de las secuencias de 16S de ADN de *Beggiatoa* sp. HS y *Leptospira* sp. (MDA-010709), a través de la base de datos RDP II (Ribosomal Database Project, RDP II) y del valor de S_ab_score, el cual es el porcentaje de 7-oligómeros compartido entre dos secuencias (Tabla 1), se confirmó que MDA1, MDA2 y MDA5 contenían especies idénticas. Esto En consecuencia, MDA1 y MDA2 fueron seleccionados para análisis de genoma mediante la secuenciación de bibliotecas *mate-pair* con un tamaño de inserto de 3 kb y MDA5 para análisis de

biblioteca s*ingle-end,* al igual que para el material amplificado de MDA-010709 (*Leptospira* sp.), a través de la plataforma de secuenciación 454 GS-FLX de Roche. La micromanipulación de filamentos, amplificación y secuenciación del ADN se realizó en Synthetic Genomics, Inc., La Jolla, CA, EE.UU.

Tabla 1. Identificación preliminar de *Beggiatoa* sp. HS y *Leptospira* sp., por medio de la función SeqMatch (Ribosomal Database Project, RDP II), utilizando el gen de ARNr 16S.

Fecha	Query nombre	S_ab_score	Nombre de secuencia (RDP SeqMatch to cultured isolates)	Forma
12/08/08	MDA1	0,615	*Beggiatoa* sp. MS-81-1c;AF1102276	Gammaproteobacterium
12/08/08	MDA2	0,617	*Beggiatoa* sp. MS-81-1c;AF1102276	Gammaproteobacterium
12/08/08	MDA3	0,797	Maorithyas hadalis gill thioautotrophic symbiont II; AB188780	Gammaproteobacterium
12/08/08	MDA5	0,613	*Beggiatoa* sp. MS-81-1c;AF1102276	Gammaproteobacterium
12/08/08	MDA6np	0,615	*Desulfonema magnum* (T); DSM2077	Deltaproteobacterium
01/07/09	MDA-010709	0,542	*Leptospira* genom osp. 1 serovar Sichuan; 79601; ATCC 700521	Spirochaetes

Fuente: Elaboración propia.

3. Pre-proceso de lecturas

El control de calidad de las secuencias se realizo a través de la herramienta FastQC Versión 0.11.2 (www.bioinformatics.babraham.ac.uk/projects/), la que proporciona una manera rápida y fácil de verificar la calidad de las secuencias en bruto (tras la secuenciación) y monitorear el avance del tratamiento de las lecturas

(pre-proceso), a través de un conjunto modular de utilidades. Entre los principales esta el módulo de puntaje de calidad de secuenciación, el que brinda una visión general de la gama de valores de calidad de todas las bases en cada posición de las lecturas. La calidad de secuenciación es valorada según la denominada escala de Phred o Q, normalmente expresada en código ASCII, el cual representa la probabilidad de que una base determinada haya sido erróneamente asignada y se representa por un número entero.

Si P es la probabilidad de error, entonces:

$$P = Q\text{-}10/10$$

$$Q = -10 \log10 (P)$$

Este módulo se reporta como fallido si el cuartil inferior para cualquier base es menor que 5 o si la mediana para cualquier base es inferior a 20. Otro módulo de gran importancia es el contenido de bases por posición. El cual evalúa la aleatoriedad en la distribución de los cuatro nucleótidos de ADN a través de las secuencias.

La ejecución del tratamiento de lecturas en bruto se llevó a cabo en primer término con el subprograma SffToCA del programa de ensamblaje Celera (Myers *et al*., 2000). El cual permitió la remoción de secuencias de adaptadores, lecturas con baja calidad, cortas (lecturas < 60 pb) y duplicaciones. Además, detectó el *linker* de las secuencias *mate-pair*, escindiendo estás en dos (Las secuencias tipo *mate-pair* de 454 están contenidas en una sola secuencia unidas por un *linker* de ~20 pb) y agregando un *flag* en ellas para su reconocimiento posterior. Por último, se transformaron las secuencias desde el formato de salida nativo de la plataforma 454 GS-FLX de Roche (sff - *Standard Flowgram Format*) al formato fastq y *fragment* (frg), el cual es el formato particular de entrada que sa Celera. sffToCA se ejecutó por

medio de línea de comandos, sobre las secuencias *mate-pair,* de la siguiente forma:

$ *sffToCA* *-insertsize* *3000 300* ***-library*** *nombre_biblioteca* ***-trim chop -clear*** *454* ***-linker*** *flx* ***-out*** *nombre_salida.frg input_archivo.sff*

insertsize = Tamano de Inserto y desviación en pb de las bibliotecas *mate-pair* originales.

trim chop = Remueve secuencias que están fuera de rango (Ejemplo *linker*)

clear 454 = Declara que el rango a "aclarar" es de tipo 454.

linker flx = Detecta específicamente el linker flx: "GTTGGAACCGAAAGGGTT TGATATTCAAACCCTTTCGGTTCCAAC".

out = Prefijo que se utilizara para los archivos de salida (fastq y frg).

Los archivos s*ingle-end* fueron tratados de la misma forma, excepto que no se utilizan las opciones de tamaño de inserto (-*insertsize*) ni tipo de *linker* (-*linker)*. Posteriormente, cuando fue necesario se filtraron y cortaron las bases y secuencias que mantuvieron un índice de calidad *Phred* (Ewing y Green, 1998) menor o igual a 30 utilizando el *software* Prinseq-lite (http://prinseq.sourceforge.net/). De esta forma, en los casos requeridos se preparó Prinseq-lite para filtrar por calidad las secuencias con una media menor a 30 (*min_qual_mean* = 30) y de ser requerido se cortaron los extremos 3' o 5´ (entre 5 – 20 pb).

4. Ensamblaje de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp.

Un ensamblaje es una estructura de datos jerárquica que asigna datos de secuencia a una reconstrucción. Generando una secuencia consenso a partir de las lecturas que solapan entre ellas, denominada c*ontig.* Luego estos *contigs* se pueden ordenar y unir dentro de super estructuras denominadas *scaffolds* (Miller et al. 2010).

La reconstrucción de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp. se llevó a cabo utilizando la aproximación bioinformática denominada ensamblaje *de novo*. El cual es un proceso de fusión y superposición de secuencias contiguas, basado en la metodología; OLC (*overlap layout consensus*). El ensamblaje del genoma de *Beggiatoa* sp. HS y de *Leptospira* sp. se realizaron utilizando como entrada las bibliotecas de lecturas ya tratadas; Beggiatoa_V1 (*single-end)*, Beggiatoa_V2 (*mate-pair)* y Beggiatoa_V3 (*single-end)*, y Leptospira (*single-end),* a través de las herramientas de ensamblaje *de novo*; Celera (Myers *et al*., 2000), Newbler y MIRA (Chevreux *et al*., 1999).

4.1 *Ensamblaje de novo con Celera*

Celera es una herramienta modular de ensamblaje *de novo*, la cual utiliza el subprograma runCA para ejecutar el ensamblaje, el que además y por defecto posee ciertos filtros tales como remover adaptadores de tipo 454 (*merTrim*), identificar secuencias duplicadas y *mate-pairs*, detectar secuencias quiméricas por comparación con otras y no considera los fragmentos menores a 64 pb. El último módulo de Celera realiza la unión de *contigs* dentro de scaffolds (scaffolding) y cierra *contigs* dentro de ellos (cierre de gaps).

La línea de comandos ejecutada en terminal y el archivo de configuración (Contiene instrucciones para llevar a cabo el ensamblaje) utilizado para el ensamblaje del genoma de *Beggiatoa* sp. HS fue el siguiente:

$ ***runCA -p*** *Beggiatoa* ***-d*** *Beggiatoa-Bogart* ***-s*** *Beggiatoa.spec* ***unitigger****=bogart Beggiatoa.frg*

-p = Prefijo que llevarán los archivos de salida.
-d = Directorio en el que se alojarán los archivos de salida.

-s = Archivo de configuración para ejecutar el ensamblaje (*Beggiatoa.spec).*
unitigger = método para construir las gráficas de *overlaping* durante el ensamblaje.

El archivo de configuración (Beggiatoa.spec) utilizado en el ensamblaje de *Beggiatoa* sp. HS con runCA (Celera) se muestas en el anexo 1.

4.2 *Ensamblaje de novo con Newbler*

El segundo programa utilizado fue Newbler, usando como entrada las bibliotecas preprocesadas de *Beggiatoa* sp. HS y *Leptospira* sp. Newbler se ejecutó por medio de la interfaz gráfica que posee, y las opciones activadas en todos los casos fueron las siguientes:

vt trimmingFile.fasta = Cortar colas polyA desde el comienzo y final de las secuencias (Además de adaptadores si los hubieran).

a NUM = Largo mínimo de *contig* a construir, por defecto: 100.

l num = Tamaño mínimo de *contig* considerado como “largos”, por defecto: 500

Newbler entrega como salida el ensamblaje a nivel de *contigs* y de *scaffolds* al igual que Celera, valiéndose de la información de distancia de las secuencias *mate-pair* identificadas, para construir los *scaffolds* (súper estructuras compuesta por fusión de *contigs* ordenados).

4.3 *Ensamblaje de novo con MIRA*

El tercer programa utilizado para llevar a cabo la reconstrucción de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp. se realizó a través de la herramienta de ensamblaje *de novo* MIRA versión 4.0.2. La cual utilizó como entrada las secuencias previamente tratadas (pre-procesadas). Las cuales fueron extraídas utilizando el *script sff_extract 0.3.0.py*. desde los archivos nativos. MIRA entrega el ensamblaje del genoma solo a nivel de *contigs* y no es capaz de hacer scaffolding por sí mismo. La línea de ejecución en la terminal y el archivo de configuración (*manifest file*) con las instrucciones para el ensamblaje de las bibliotecas de *Beggiatoa* sp. HS con MIRA fue la siguiente:

$ ***mira*** *manifiesto_beggiatoa.conf >&proj.log*

Archivo de configuración para ensamblar genoma de *Beggiatoa* sp. HS se muestra en el anexo 2. El archivo de configuración para ejecutar el ensamblaje de la biblioteca de *Leptospira* sp. fue idéntico al de *Beggiatoa* sp. HS excepto por la definición del tipo de data.

5. Conciliación de ensamblajes *de novo* utilizando el *software* CISA

Tras los ensamblajes con los tres programas mencionados anteriormente, se evaluó la calidad de las reconstrucciones generadas por cada uno de ellos, evaluando el estadístico denominado N50, el cual se define como el tamaño de *contig* o *scaffold* mínimo en el que está contenido el 50% de las bases del genoma obtenido y es el estadístico más utilizado para describir la calidad de un genoma reconstruido, tanto a nivel de *contigs* como *scaffolds.* Además, se evaluó el tamaño máximo y medio de los *contigs* y *scaffolds*. Por último, se considera el número de *contigs* y *scaffolds* obtenidos, así como el nivel de GC.

Con el objetivo de mejorar las métricas de ensamblaje para cada organismo, principalmente el N50, Tamaño máximo de *contig*, tamaño medio de *contigs* y número de *contigs* se procedió a conciliar los contigs generados por Celera, Newbler y MIRA, tanto para Beggiatoa sp. HS como para Leptospira sp. Este proceso se realizo con la herramienta de conciliación de ensamblajes llamada CISA (http://sb.nhri.org.tw/CISA/en/CISA).

Como el resultado de ensamblaje tanto para el genoma de *Beggiatoa* sp. HS como para *Leptospira* sp. fue rendido por Celera. Se utilizó este como base para la conciliación de los ensamblajes (considerando *contigs* > 1.000 pb) y una versión más exigente para Newbler y MIRA, considerando *contigs* > 2.000 pb, de forma de reducir el desmedido tamaño de los genoma generados y el número de *contigs*.

CISA debe generar nuevos archivos de los ensamblajes y adecuarlos para la unión, lo cual realiza con el subprograma Merge.py, indexando y generando un archivo con los ensamblajes de *contigs* utilizados. La línea de instrucción en terminal utilizada para el programa en conjunto con el archivo de configuración se ejecutó de la siguiente forma:

$ python ***Merge.py*** *Merge_Leptospira.config*

Donde *Merge_Leptospira.config* establece:

count = 3
data = Newbler_2000.fa,title=Contig_m1
data = Spiro_celera.fasta,title=Contig_m2
data = Mira_2000.fa,title=Contig_m3
Master_file = Contigs_merge.fa
min_length = 100
Gap = 11

El archivo de configuración indica el número de ensamblajes a unir (count), ensamblajes a nivel de *contigs* a unir (data), el archivo de salida (Contigs_merge.fa), un mínimo de tamaño de *contig* a considerar en la unión (min_length) y el nivel de *gap* tolerado (Gap - divide los ensamblajes en *contigs* si encuentra >10 Ns). Luego de generar el archivo con los ensamblajes a unir, se ejecutó la unión de ellos utilizando el programa CISA.py y el archivo de configuración, como se indica a continuación:

$ *python* ***CISA.py*** *Leptospira_cisa.config*

Archivo de configuración (*Leptospira_cisa.config):*

genome = 6700000 # Tamaño de genoma esperado
infile = Contigs_merge_1.fa # Archivo generado por Merge.py
R2_Gap = 0.95 # Umbral utilizado en la fase 2 de CISA
outfile = CISA_beggiatoa.fa # Nombre de archivo de salida
nucmer = /home/alfon/Documentos/MUMmer3.23/nucmer
R2_Gap = 0.95
CISA = /home/alfon/CISA1.3
makeblastdb = /usr/bin/makeblastdb
blastn = /usr/bin/blastn

Este proceso generó un archivo en formato fasta con *contigs*, producto de la conciliación de los ensamblajes ensayados con Celera, Newbler y MIRA.

6. *Scaffolding* de los *contigs* unidos con el *software* CISA

El scaffolding consiste en utilizar la información de distancia entre las secuencias *mate-pair* para ordenar y orientar los *contigs* dentro de *scaffolds.*

CISA recibe y genera el resultado en *contigs*, por tanto, el resultado obtenido en primera instancia en cuanto a métricas es potencialmente mejorable, si se llevan los *contigs* al nivel estructural de *scaffolds*. Por otra parte, Celera y Newbler tienen la capacidad de realizar *scaffolding* por sí mismos. Lo que implica un nivel de organización superior, el cual es configurado por *contigs* ordenados y orientados. Por lo tanto las métricas de ensamblaje a nivel de *scaffold* son potencialmente superiores que el nivel de *contigs*, dependiendo de la eficiencia del *scaffolding*, ya que la fusión de *contigs* genera estructuras de mayor tamaño, lo que conlleva a un aumento en el N50, una disminución en la fragmentación y probablemente un aumento en las restantes métricas.

En consecuencia, se procedió a realizar el *scaffolding* de los *contigs* generados por CISA producto de la conciliación de los ensamblajes de *Beggiatoa* sp. HS y evaluar una posible mejora en las métricas con respecto al resultado preliminar del mismo y con respecto a los ensamblajes a nivel de *scaffold* generados por Celera y Newbler. Por otra parte, no es posible realizar *scaffolding* del genoma de *Leptospira* sp. generado con cualquier herramienta de ensamblaje *de novo*, debido a que se carece de lecturas tipo *mate-pair*, lo que imposibilita poder ordenar y orientar *contigs* dentro de *scaffolds*.

El *scaffolding* de los *contigs* generados por CISA se llevó a cabo con el programa SSPACE (Boetzer *et al*., 2011), el cual requirió de las bibliotecas *mate-pair* de *Beggiatoa* sp. HS generadas con anterioridad por sffToCA, resultantes de la remoción del *linker* y escisión de las lecturas *mate-pair* en dos lecturas, contenidas cada una en un archivo distinto en formato fastq.

La linea de comandos para ejecutar SPPACE.pl fue la siguiente:

$ **SSPACE_v3.0.pl -l** library_Beggiatoa.txt **-s** CISA_beggiatoa.fa **-x** 1

-l = Archivo que contiene la información de las bibliotecas.
-s = Archivo de *contigs* en formato fasta a ser sometido a *scaffolding.*
-x = Hacer extensión de *contigs* (x = 1).

Archivo **library_Beggiatoa.txt:**

v1 bwa v1_beggiatoa.1.fastq v1_beggiatoa.2.fastq 3000 0.25 FR
v2 bwa v2_beggiatoa.1.fastq v2_beggiatoa.2.fastq 3000 0.25 FR

donde:

1.ª columna = Identificador para las bibliotecas (v1 y v2).
2.ª columna = Mapeador utilizado con las bibliotecas (bwa).
3.ª y 4.ª columna = bibliotecas *mate-pair* (fastq) un extremo en cada archivo.
5.ª columna = Tamaño de inserto de secuencias *mate-pair.*
6.ª columna = Error mínimo esperado.
7.ª columna = Orientación de las secuencias (→←).

Todos los análisis se realizaron por medio de software libre en sistemas OS Mac, y apoyados en un clúster ROCKS de 64 núcleos para realizar procesos de alta complejidad como ensamblaje *de novo*.

7. Anotación de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp.

La anotación de genomas es un proceso multinivel que incluye la predicción

de genes codificadores de proteínas, así como otras unidades funcionales del genoma tales como ARNs estructurales, ARNt, ARN pequeños, pseudogenes, regiones de control, repeticiones directas e invertidas, secuencias de inserción, transposones y otros Elementos móviles.

7.1 *Anotación general*

La predicción de genes se realizó a través simultáneamente entre la plataforma RAST (Aziz *et al*., 2008) y el programa de anotación Prokka (*rapid prokaryotic genome annotation*) (Seemann, 2014), utilizado de forma local. Los genes de ARN ribosomal fueron identificados con Barrnap versión 3 (*BAsic Rapid Ribosomal RNA Predictor*) (**http://www.vicbioinformatics.com/software.barrnap.shtml**), que utiliza la herramienta de búsqueda HMMER 3.1. Complementariamente se utilizaron BLASTN y BLASTP, consultando sus bases de datos no redundantes, y la base de datos Pfam (Finn *et al*., 2008) para determinar dominios funcionales y familias de proteínas.

7.2 *Análisis filogenético utilizando el gen 16S de ARNr*

A pesar que desde el momento de la micromanipulación de los filamentos bacterianos se presumió su taxonómica aproximada, complementada posteriormente con la amplificación, secuenciación y búsqueda en la base de datos Ribosomal Database Project del gen 16S de ARNr, se realizaron análisis más detallados que los anteriores. En consecuencia, se llevó a cabo la reconstrucción filogenética de ambas bacterias, utilizando la secuencia del gen 16S de ARNr, identificado desde la anotación.

Después del ensamblaje de ambos genomas, se tomaron los genes 16S de

ARNr (~1.500 pb de longitud) identificados con Barrnap en los borradores de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp. Desde búsquedas BLAST (McGinnis y Madden 2004) se identificaron y tomaron los primeras cincuenta secuencias de 16S ARNr homólogas a la secuencia de 16S de ARNr de *Beggiatoa* sp. HS para la construcción de su filogenia. Además, como grupo externo se incluyeron dos genes de 16S del rRNA de *Ruegeria atlantica* y una de *Ruegeria* sp. (*Alphaproteobacteria*). Por otro lado, para construir la filogenia de *Leptospira* sp. se tomaron 50 secuencias del gen 16S de ARNr pertenecientes a diversos géneros del filo *Spirochaeta* desde GenBank y se incluyeron 3 secuencias de 16S de ARNr de *Deferribacter* spp. (Deferribacteres), como grupo externo. Las secuencias identificadas por las búsquedas BLAST se recuperaron de forma automática utilizando el paquete Bio.Blast de Biopython. EL alineamiento de secuencias se realizó a través del programa MUSCLE v3.8.31 (Edgar 2004), estableciendo 1.000 iteraciones. La filogenia fue construida utilizando el enfoque bayesiano a través de la herramienta MrBayes (Huelsenbeck et al., 2001) durante dos ejecuciones. En donde cada ejecución constó de cuatro cadenas de 1.000.000 de generaciones. La probabilidad a posterior se estimó utilizando el método de cadenas de Markov Monte Carlo (MCMC) y el modelo de sustitución usado fue el de GTR (*General Time Reversible*), con una tasa de sustitución=6. El método de reconstrucción utilizado fue el de máxima verosimilitud y se descartó el 25% de los árboles muestreados.

7.3 *Identificación y anotación funcional de Clúster de genes biosintéticos implicados con la biosíntesis de Metabolitos secundarios*

La identificación, anotación funcional y estructural de los CGBs relacionados con biosíntesis de MSs se llevó a cabo con el software antiSMASH versión 3.0 (Medema et al., 2011; Weber *et al*. 2015) y la plataforma NapDos (Ziemert *et al*., 2012) capaz de detectar dominios de Condensación (C) y ceto sintasa (KS). Se utilizaron los genomas de *Beggiatoa* sp. HS y *Leptospira* sp. previamente anotados

con RAST en formato GenBank como entrada en antiSMASH, formato de entrada altamente recomendado. La lógica de detección de antiSMASH es detectar los Perfiles de modelos de Markov ocultos (pHMM) en secuencias genes, y utilizar dominios conservados típicos de enzimas implicadas en MSs como marcadores, para la identificación de un posible CGB. Además, antiSMASH utilizo BlastP para comparar los genes predichos en los CGBs, análisis de las familias de genes (COG), identificando regiones conservadas de secuencias características de genes, y análisis con la base de datos Pfam. Las predicciones de antiSMASH se contrastaron con los genes predichos por Prokka y se determinó un consenso de los genes que compondrían los CGBs candidatos implicados en la biosíntesis de MSs en los genomas de *Beggiatoa* sp. HS y *Leptospira* sp.

Las secuencias de genes identificados, fueron confirmado como tal, cuando cumplieron con los criterios de (a) E-*value* $<1e^{-8}$, (b) cobertura > 60% y (c) identidad de secuencia > 30%, en contraste con la base no redundante de Blastp.

8. Visualización de los genomas

Para observar el resultado de los genomas ensamblados se utilizó la herramienta de visualización de código abierto Hawkeye del proyecto AMOS versión 3.1.0 (Schatz *et al*., 2013), el cual permite apreciar las lecturas ensambladas, la profundidad de ensamblaje y las lecturas de cada *contig* o *scaffold*, entre otras características. Además, se utilizó Artemis (Rutherford *et al*., 2000), un software de visualización y análisis de datos de secuenciación de alto rendimiento, el cual permitió observar los clúster de genes y genes adyacentes a ellos en los genomas. Los genomas se representaron a través del programa Circos (Krzywinski *et al*., 2009), por medio del cual se representaron los cromosomas bacterianos en forma circular, además de representación de datos asociados, como nivel de GC y secuencias de genes codificantes.

III. RESULTADOS

1. Tratamiento de las lecturas de *Beggiatoa* sp. HS y *Leptospira* sp.

El tratamiento de lecturas en bruto o preproceso consistió en controlar la calidad de secuenciación, removiendo adaptadores, partidores, filtrando lecturas cortas (< 100 pb), bases/lecturas de baja calidad, de acuerdo a la escala de Phred (< 20 puntaje Phred).

A partir de la secuenciación de dos bibliotecas *mate-pair* y una *single-end* desde tres filamentos de *Beggiatoa* sp. HS se obtuvieron tres archivos en formato sff; Beggiatoa_V1, Beggiatoa_V2 y Beggiatoa_V3 (Tabla 2).

Tabla 2. **Características de los archivos obtenidos tras la secuenciación de ADN genómico de *Beggiatoa* sp. HS y *Leptospira* sp. con la plataforma de secuenciació 454 GS-FLX de Roche.**

Archivo	Total de lecturas	Longitud	%GC	Tipo de lecturas
Beggiatoa_v1	343.049	54 - 1453	45	mate-paired
Beggiatoa_v2	342.409	58 - 1122	45	mate-paired
Beggiatoa_v3	490.095	53 - 2048	43	Single-end
Leptospira	777.836	51 - 1200	33	Single-end

Fuente: Elaboración propia.

Por otra parte, tras la secuenciación de una biblioteca *single-end* construida a partir de ADN de *Leptospira* sp. se obtuvo un archivo sff (Tabla 2). En todos los casos se utilizó una plataforma de pirosecuenciación 454 GS-FLX de Roche (Ronaghi, 2001).

Como se mencionó anteriormente, previo a la reconstrucción de los genomas, fue necesario determinar que las secuencias a utilizar tengan la calidad apropiada.

A través de la herramienta de control de calidad fastQC, se detectaron los aspectos de calidad considerados como correctos y deficientes a través de las lecturas. La Figura 8 muestra los módulos de calidad marcados como deficientes en las lecturas en bruto de *Beggiatoa* sp. HS El primer módulo marcado como deficiente es el de puntaje de calidad a través de las bases (Figura 8 primera columna: A-D-G y J). Este módulo se reporta como deficiente debido a la existencia de bases con alta probabilidad de haber sido asignadas de forma errónea, según el puntaje de *Phred*, teniendo en consideración que la zona verde de la gráfica, puntaje de *Phred* sobre 30, se considera de óptima calidad, la zona naranja, entre 20 y 29, como aceptable y la zona roja, bajo 20, como de calidad inaceptable. Siendo notoria la baja en puntaje de calidad que se presenta hacia el final de las secuencias. El segundo módulo identificado como deficiente corresponde al contenido de Timina (T = rojo), Citosina (C = azul), Adenina (A = verde) y Guanina (Guanina = negro) a través de las bases (Figura 8. B-E-H y K), marcando un desbalance entre tetrámeros, enfatizado en los extremos de las secuencias, las cuales lucen un fuerte desequilibrio. Este módulo es considerado fallido cuando existen diferencias sobre el 20% entre A y T, o G y C. La última columna (Figura 8 = C-F-I y L) muestra el módulo de ocurrencia de K-meros, las cuales representan pequeñas secuencias de 7 pb, sobrerrepresentadas y anormalmente distribuidas. Este módulo es reportado como erróneo cuando cualquier k-mero presenta un desbalance con un valor binomial p-value $< 10^{-5}$.

Tras el filtrado de las bases con bajo puntaje de calidad (puntaje de Phred < 20), corte de secuencias, remoción de duplicaciones y primers con la sub-rutina SffToCA y Prinseq-lite se lograron adecuar las secuencias a la calidad requerida para pasar al siguiente paso de reconstrucción de los genomas.

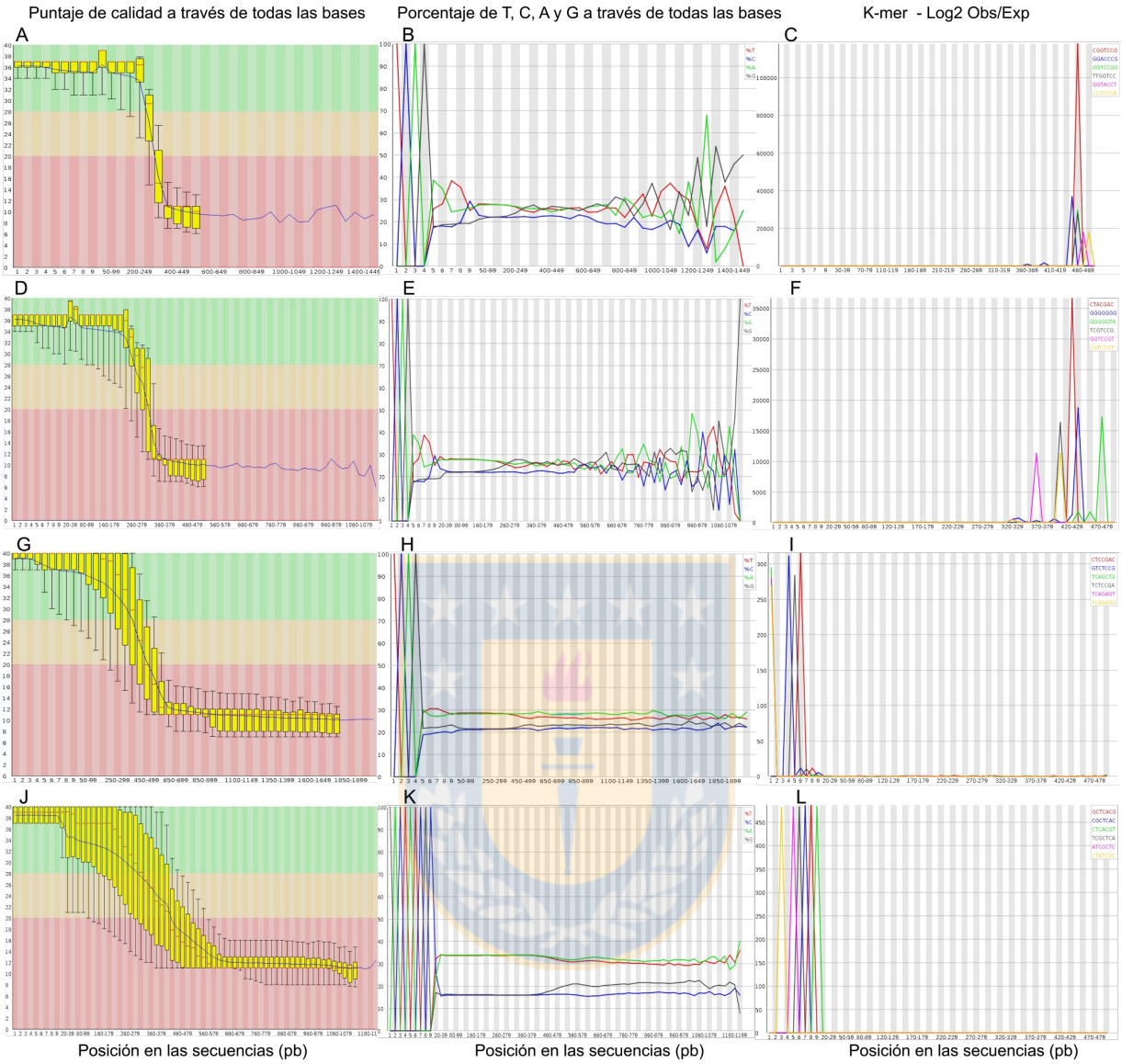


Figura 8. Módulos identificados como deficientes en las lecturas en bruto de *Beggiatoa* sp. HS y de *Leptospira* sp.

Siendo Beggiatoa_V1: A-B-C, Beggiatoa_V2: D-E-F, Beggiatoa_V3: G-H-I y Leptospira: J-K-L. La primera columna muestra el puntaje de calidad de secuenciación a través de todas las bases (*Phred* de 0 a 40). La segunda columna muestra el contenido de T (rojo), C (azul), A (verde) y G (negro) en porcentaje y la tercera columna muestra el contenido de K-meros a través de todas las bases. Fuente: Elaboración propia.

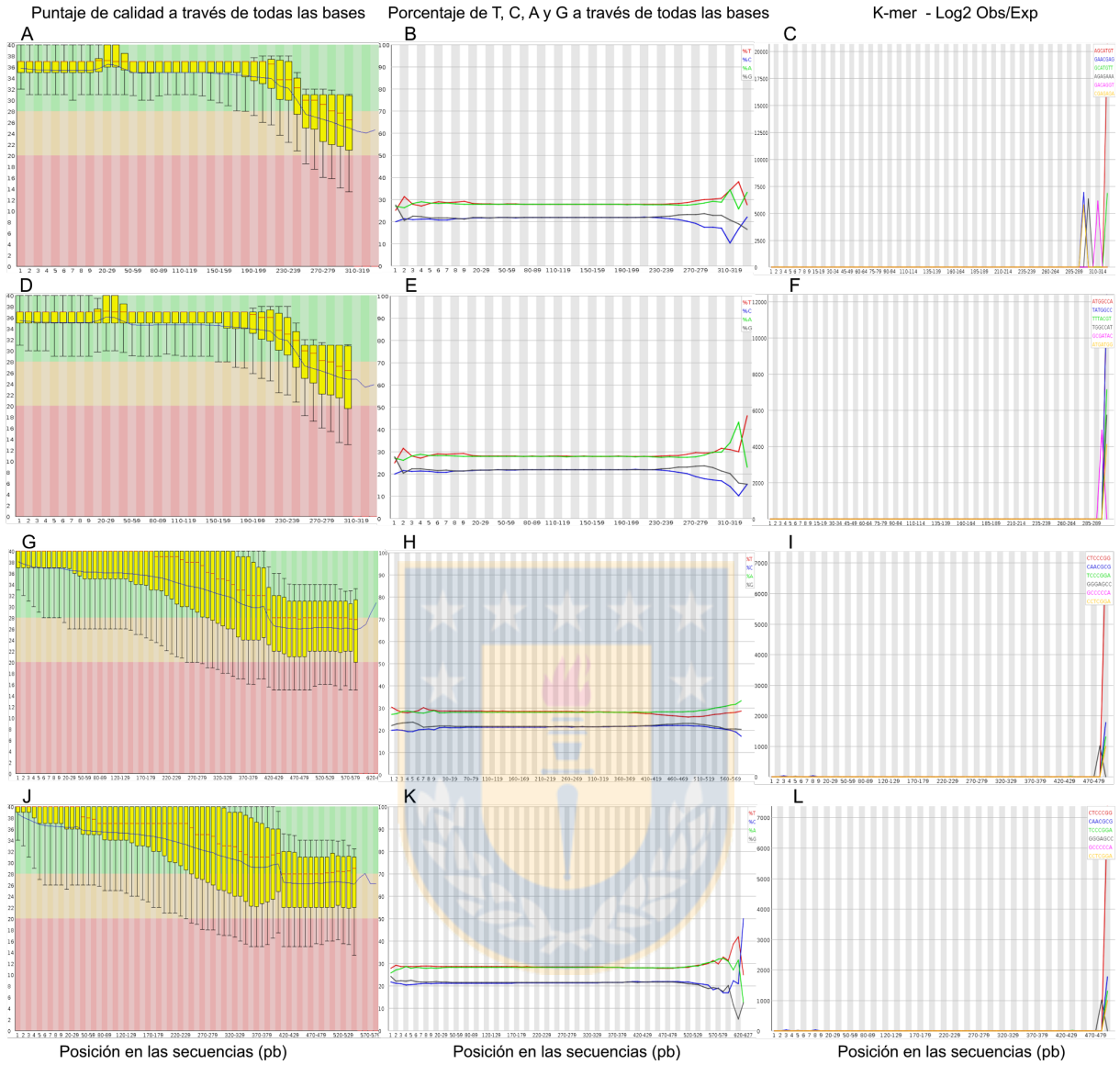


Figura 9. Módulos tras pre-proceso de las lecturas de *Beggiatoa* sp. HS y de *Leptospira* sp.

Siendo Beggiatoa_V1: A-B-C, Beggiatoa_V2: D-E-F, Beggiatoa_V3: G-H-I y Leptospira: J-K-L. La primera columna muestra el puntaje de calidad de secuenciación a través de todas las bases (*Phred* de 0 a 40). La segunda columna muestra el contenido de T (rojo), C (azul), A (verde) y G (negro) en porcentaje y la tercera columna muestra el contenido de K-meros a través de todas las bases. Fuente: Elaboración propia.

Las Figuras 9. A-D-G-J, muestran que la media en puntaje de calidad (Línea roja dentro de cada barra) a través de todas las bases está por sobre 30 (Zona verde), en casi la totalidad de los casos, los inter-cuartiles (25-75%) representados por las cajas amarillas están en la zona aceptable >20 y mayoritariamente en la zona óptima de calidad >30 (Zona verde), al igual que la media del puntaje de calidad, representada por la línea azul. La segunda columna de la Figura 9. B-E-H-K, muestra una normalización de las diferencias entre las proporciones de A-T y G-C, pasando estas diferencias a un valor menor al 10%, calificado como adecuado, a pesar de los pequeños desbalances en los extremos 5' de las gráficas (Figura 9. B-E-H y K). Por último, el tercer módulo identificado previamente como fallido; contenido de K-meros, en la tercera columna de la Figura 9. C-F-I-L, paso de un estado fallido a un estado de alerta, teniendo una desviación binomial de p-value <0,01. Sin embargo, esto no debiera afectar significativamente los análisis posteriores. La tabla 3 muestra las lecturas remanentes tras el tratamiento de pre-proceso sobre ellas, el rango de longitud, y el nivel de GC en porcentaje de las lecturas.

Tabla 3. L**ecturas obtenidas tras el preprocesamiento en los set de datos de *Beggiatoa* sp. HS y *Leptospira* sp.**

Archivo	Total de lecturas	Longitud	%GC	Tipo de lecturas
Beggiatoa_v1	244.880	64 - 339	43	mate-paired
Beggiatoa_v2	245.912	64 - 338	43	mate-paired
Beggiatoa_v3	476.756	64 - 743	43	Single-end
Leptospira	670.233	375 - 595	32	Single-end

Fuente: Elaboración propia.

2. Reconstrucción de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp.

Luego del tratamiento de las lecturas, se llevó a cabo la reconstrucción del genoma de *Beggiatoa* sp. HS y *Leptospira* sp. a través de la estrategia bioinformática denominada ensamblaje *de novo*, utilizando tres ensambladores distintos; Celera, MIRA y Newbler, para luego conciliarlos con la herramienta de unión de ensamblajes CISA, buscando mejorar las estadísticas de los genomas ensamblados**.** Entre ellas el N50, número, tamaño medio y máximo de *contigs* o *scaffolds* que conforman el genoma.

2.1 *Ensamblaje de novo del genoma de Beggiatoa sp. HS*

La Figura 10 muestra las principales métricas del ensamblaje del genoma de *Beggiatoa* sp. HS a nivel de *contigs*, obtenidos con Celera, Newbler y MIRA, y la conciliación de todos ellos con CISA. Los resultados muestran que Celera rinde el N50 (29 Kb), tamaño máximo (113 Kb) y medio (7,6 Kb) de *contig* más altos (Figura 10. A-B-C). Además, un número menor de *contigs* (794) en comparación con Newbler y MIRA (Figura 10. D). Siendo generalmente deseable un genoma con menor fragmentación. Por otro lado, los genomas ensamblados con Newbler (5,8 Mb) y Celera (6,19 Mb) rindieron un tamaño muy similar. Mientras que MIRA está muy por sobre ambos (9,8 Mb) (Figura 10. E). Finalmente, al ejecutar la unión de los tres ensamblajes con CISA, se observa que las métricas mejoraron, aumentando el N50 a 44,7 Kb y el tamaño medio y máximo de *contigs* a 28,3 y 139, 6 Kb, respectivamente. Además, se redujo el número de *contigs* a 214 en un genoma de 6,0 Mb.

Tanto Celera como Newbler tienen la capacidad de utilizar la información de distancia de las secuencias *mate-pair* para ordenar y unir *contigs* dentro de super-estructuras denominadas *scaffolds*. Alcanzando de esta forma un nivel superior en

orden y disminuyendo la fragmentación del genoma reconstruido.

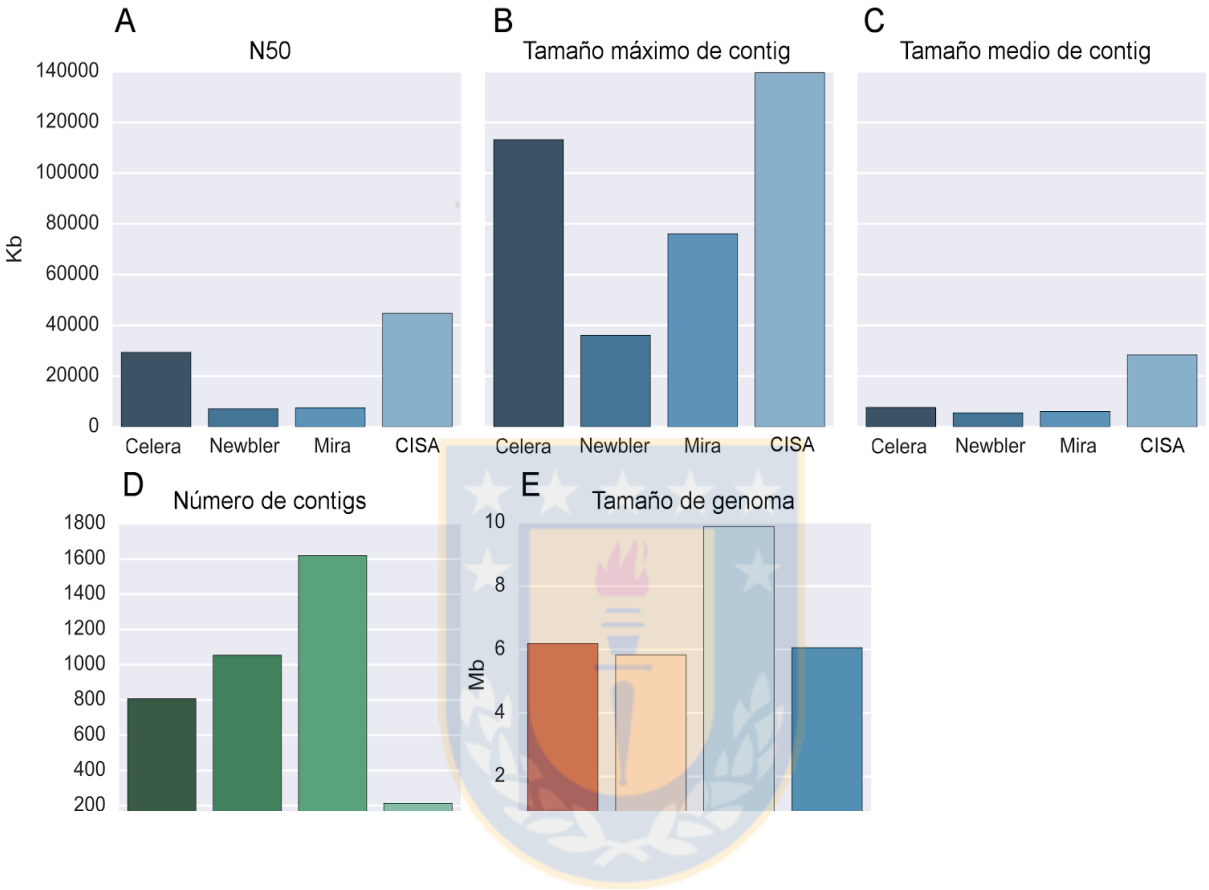


Figura 10. Métricas del ensamblaje *de novo* del genoma de *Beggiatoa* sp. HS a nivel de *contigs*, llevado a cabo con Celera, Newbler y MIRA, y la conciliación de ensamblajes con CISA.

A. Valor de N50. **B.** Tamaño máximo de *contig* y **3.** Tamaño medio de *contig* en kilobases (Kb). **D.** Número de *contigs* obtenidos. **E.** Tamaño del genoma rendido por cada herramienta en millones de bases (Mb). Fuente: Elaboración propia.

Valiéndose de 53.085 fragmentos *mate-pair* correctos (existen 176.122 fragmentos parciales) de la biblioteca Beggiatoa_V1 y 52.125 fragmentos *mate-pair*

correctos (Existen 175.754 fragmentos parciales) de la biblioteca Beggiatoa_V2, Celera y Newbler lograron ordenar y unir *contigs* dentro de 560 (Figura 11) y 487 *scaffolds*, respectivamente (Figura 11. D). Además, por medio del programa externo

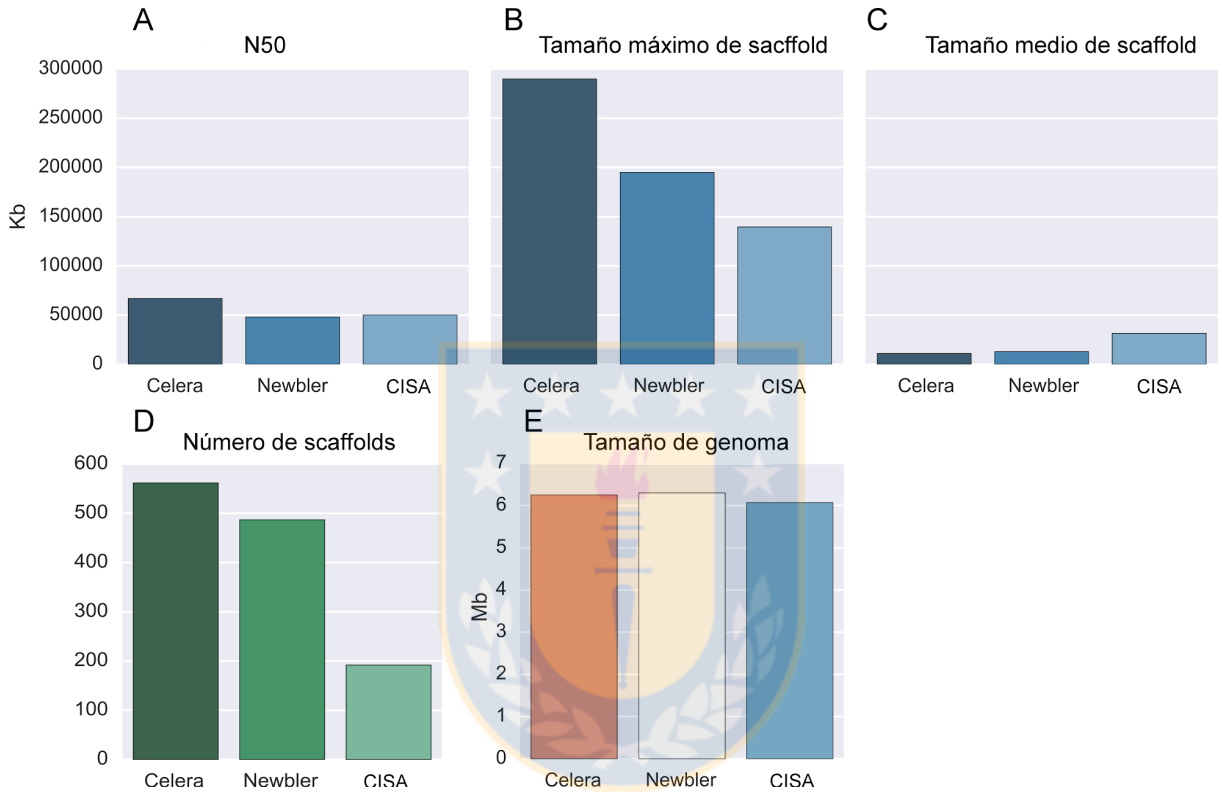


Figura 11. Métricas del ensamblaje *de novo* del genoma de *Beggiatoa* sp. HS a nivel de *scaffolds*, llevado a cabo con Celera, Newbler y *scaffolding* realizado con SSPACE sobre los *contigs* conciliados por CISA.

A. Valor de N50. **B.** Tamaño máximo de *scaffold*. **C.** Tamaño medio de *scaffold* en kilobases (Kb). **D.** Número de *scaffolds* obtenidos. **E.** Tamaño del genoma rendido por cada herramienta en millones de bases (Mb). Fuente: Elaboración propia.

de *scaffolding* SSPACE se ordenaron y unieron *contigs* rendidos por CISA dentro de 192 *scaffolds* (No se considero el ensamblaje previo de MIRA). Las métricas de

ensamblaje a nivel de *scaffolds* entregadas por Celera en estadísticos claves, como el N50 (67 Kb) y tamaño máximo de *scaffold* (289,8 Kb), fueron superiores a las entregadas por Newbler (N50 = 48 Kb; tamaño máximo de *scaffold* = 195 Kb), e incluso a los entregados por SSPACE a partir de los *contigs* conciliados por CISA (N50 = 50,1 Kb; tamaño máximo de *scaffold* = 139,5 Kb) (Figura 11. A y B). Por otra parte, el *scaffolding* de los *contigs* conciliados por CISA, presenta métricas superiores a los ensamblajes individuales, en tamaño medio de *scaffold* (31,6 Kb) (Figura 11. C) y una menor fragmentación (192 *scaffolds*) (Figura 11. D), en un genoma de 6,1 Mb, muy similar a los rendidos por Celera (6,2 Gb) y Newbler (6,3 Gb).

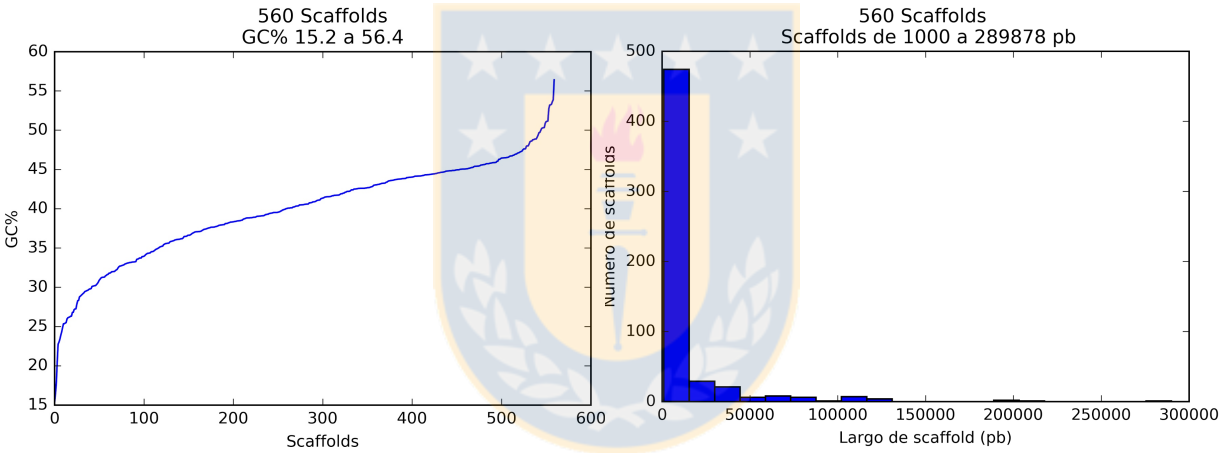


Figura 12. **Contenido de GC y distribución del tamaño de *scaffolds* en el genoma de *Beggiatoa* sp. HS obtenido con Celera.**
La gráfica del lado izquierdo muestra el contenido de Guanina-Citosina (GC) en porcentaje a través de todos los *scaffolds.* La figura de la derecha muestra un histograma de la distribución del tamaño de los *scaffolds.* Fuente: Elaboración propia.

Por otro lado, el contenido de GC del ensamblaje de *Beggiatoa* sp. HS (GC = 43.5%) obtenido del *scaffolding* sobre los *contigs* conciliados por CISA se desvía en un 3,7% del contenido de GC del ensamblaje rendido por Celera (GC = 39,8%)

(Figura 12).

Tomando en consideración las métricas de ensamblaje a nivel de *scaffolds* del genoma de *Beggiatoa* sp. HS, en particular del N50 y tamaño máximo de *scaffold*, además del contenido de GC. Se consideró que el ensamblaje de mejor calidad fue el generado por Celera (Ver métricas en Tabla 4).

2.2 *Ensamblaje de novo del genoma de Leptospira sp.*

El procedimiento para la reconstrucción del genoma de *Leptospira* sp. fue el mismo que el realizado con *Beggiatoa* sp. HS. Se llevaron a cabo ensamblajes con Celera, Newbler y MIRA, para luego unirlos con CISA. El ensamblaje de mejor calidad, considerando todos los *contigs* generados, fue el realizado con Celera. Obteniendo un N50 (17,9 Kb) y tamaño máximo de *contig* (87,5 Kb) superior al resto, un tamaño medio de *contig* (6,4 Kb) similar al rendido por Newbler (6,6 Kb) y un menor número de *contigs* (1.057 *contigs*). La Figura 13 muestra las métricas originales del ensamblaje obtenido con Celera (*contigs* >1000 pb) y las métricas de Newbler y MIRA luego de editar sus ensamblajes, para mejorar su calidad, considerando solo *contigs* > 2000 pb. De tal forma que fueran útiles para unirlos al ensamblaje de Celera con CISA.

El resultado de la conciliación de ensamblajes con CISA, mejoro las métricas del ensamblaje de mejor calidad, llevado a cabo con Celera. De este modo, se obtuvo un N50 de 34,5 Kb (Figura 13. A), un tamaño máximo de *contig* de 127,7 Kb (Figura 13. B) y tamaño medio de *contig* de 26,8 Kb (Figura 13. C). Además, CISA género un genoma de 6,8 Mb (Figura 13. E), menos fragmentado, con 257 *contigs* (Figura 13 D). .

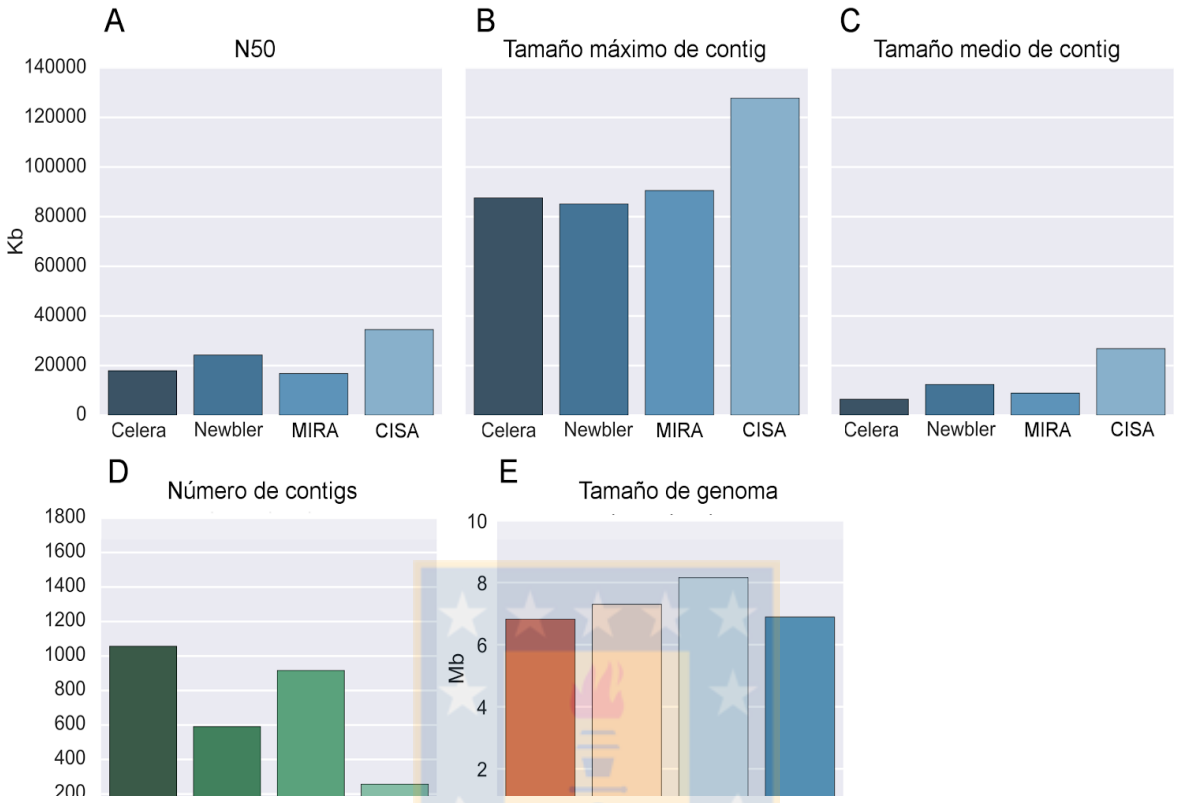


Figura 13. Métricas del ensamblaje *de novo* del genoma de *Leptospira* sp. a nivel de *contigs*, con Celera (contigs >1000 pb), Newbler y MIRA (*contigs* >2000 pb), y de la herramienta de conciliación de ensamblajes CISA.

A. Valor de N50. **B.** Tamaño máximo de *contig*. **C.** Tamaño medio de *contig* en Kilobases (Kb). **D.** Número de *contigs* obtenidos. **E.** Tamaño del genoma generado por cada herramienta en millones de pares de bases (Mb). Fuente: Elaboración propia.

Por último, el porcentaje medio de GC tras la unión de los tres ensamblajes con CISA fue de 32,3% (Figura 14), el mismo obtenido con Celera (GC = 32,3%).

Por otro lado, a diferencia de *Beggiatoa* sp. HS, *Leptospira* sp. solo cuenta

con una biblioteca de tipo *single-end*, no siendo posible ordenar y conciliar *contigs* dentro de *scaffolds*, debido a no contar con la información de distancia que entregan las secuencias *mate-pair*.

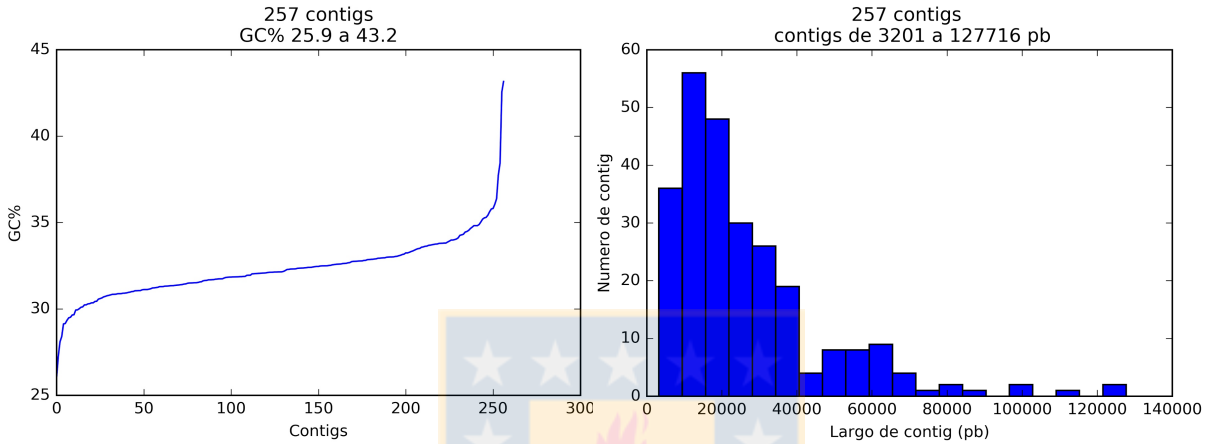


Figura 14. **Contenido de GC y distribución del tamaño de *contigs* en el genoma de *Leptospira* sp. obtenido tras la unión de ensamblajes con CISA.**
La gráfica del panel izquierdo muestra el contenido de Guanina-Citosina (GC) en porcentaje a través de todos los *contigs.* La figura de la derecha muestra un histograma de la distribución del tamaño de *contigs.*Fuente: Elaboración propia.

La Tabla 4 contiene las métricas de los genomas ensamblados de *Beggiatoa* sp. HS y *Leptospira* sp. considerados de mayor calidad, utilizando las métricas de ensamblaje como parámetro de comparación. En el caso de *Beggiatoa* sp. HS a pesar que el *scaffolding* de la conciliación de ensamblajes ejecutado con CISA rindió un genoma mucho menos fragmentado, el N50 y tamaño máximo de *scaffold* fue inferior al ensamblaje rendido por Celera a nivel de *scaffold*.

Tabla 4. Métricas del genoma de mayor calidad para *Beggiatoa* sp. HS y *Leptospira* sp.

Métrica	*Beggiatoa* HS sp.	*Leptospira* sp.
	Celera	CISA
contigs	807 (794 > 1000)	**257**
tamaño genoma (Mb)	6,25	**6,89**
N50 (Kb)	29,3	**34,5**
tamaño máximo de *contig* (kb)	113	**127,1**
promedio del tamaño de *contig* (kb)	7,6	**26,8**
GC (%)	41,6	**32,3**
secuencias ensambladas	691.263 (65%)	ND
Coverage	32X	ND
scaffolds	**560**	-
N50 (Kb)	**67.0**	-
tamaño máximo de *scaffold* (kb)	**289.2**	-
promedio tamaño de *contig scaffold* (kb)	**11.0**	-

Mb: millones de pares de bases. Kb: miles de pares de bases.
ND: Datos no disponibles.

Fuente: Elaboración propia.

En el caso de *Leptospira* sp. el mejor ensamblaje también fue generado por Celera. Sin embargo, al ejecutar la conciliación de ensamblajes con CISA, mejoraron todas las métricas y no distorsionó el contenido de GC.

3. Anotación de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp.

La anotación de genomas hace referencia a la identificación de secuencias de genes codificantes de proteínas, así como de otras estructuras funcionales, tales como ARNs, pseudogenes, regiones de control, entre otras. La anotación general de genes se realizó con RAST y Prokka, y la anotación funcional de los CGBs se llevó a cabo con antiSMASH.

3.1 *Anotación general del genoma de Beggiatoa sp. HS*

Tras ejecutar la anotación del borrador del genoma de *Beggiatoa* sp. HS se identificaron 5.220 secuencias de genes codificantes, cinco ARNr; un 16S, un 23S y tres 5S, además de 43 ARNt y un ARNtm. Los tres ARNr 5S tienen un tamaño de 109 pb y el ARNr 23S está dividido en dos piezas, en una región contigua del *scaffold* scf31540, separados por una región de unión de *gap*, siendo ambas secuencias de un tamaño de ~1600 pb. Del total de genes, 2.159 genes predichos corresponden a proteínas hipotéticas, dejando 3.061 genes candidatos con una función conocida. La Figura 15, esquematiza el borrador del genoma de *Beggiatoa* sp. HS anotado, el cual posee un tamaño de 6.2 Mb. Los *scaffolds* destacados en coloreados en negro y azul. Además, las regiones codificantes de proteínas (Con función conocidas y sin función conocida) de la cadena sentido y antisentido del ADN, identificadas en el.

La Figura 16 muestra una representación de la anotación funcional de los genes identificados en el borrador del genoma de *Beggiatoa* sp. HS se encontraron 2.351 secuencias dentro de 372 colecciones de familias de proteínas funcionalmente relacionadas (subsistemas). Del total, el ~12% (273) están ligadas al metabolismo de

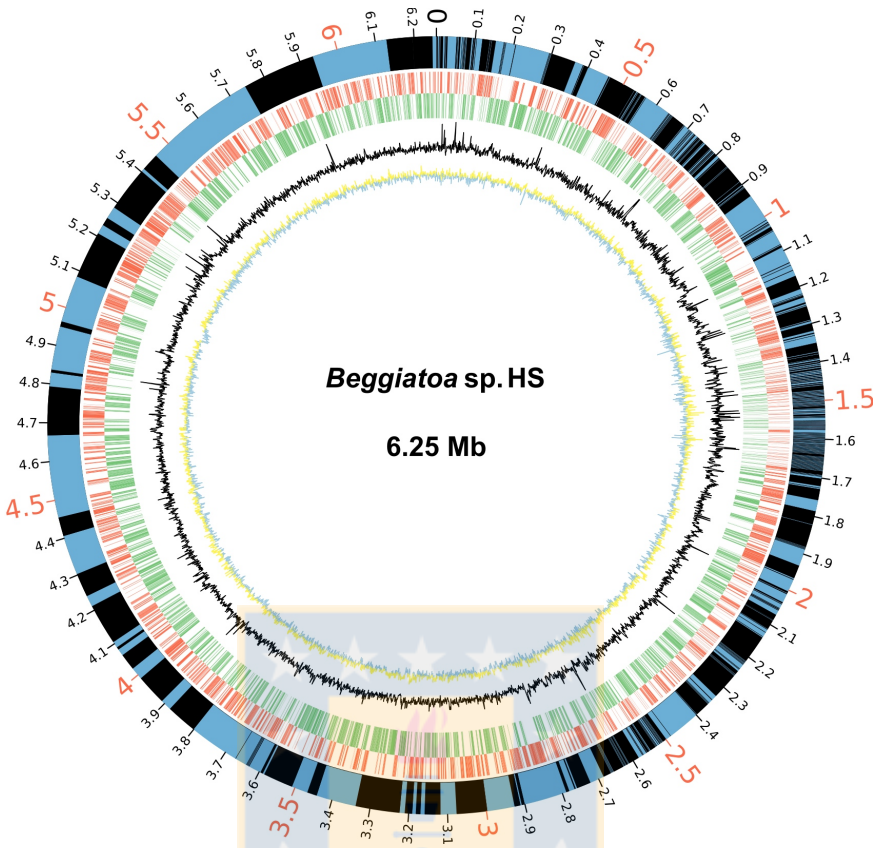


Figura 15. Representación circular del *draft* del genoma de *Beggiatoa* sp. HS. Desde el anillo más interno hacia fuera: el anillo amarillo (positivo) y azul (negativo) muestra el GC *skew* en una ventana = 1000 pb; el anillo negro describe el contenido de GC% en una ventana = 1000 pb; el anillo verde representa la predicción de genes de la cadena de ADN antisentido y el anillo rojo las CDSs de la cadena de ADN sentido; el círculo más externo muestra todos los *scaffolds*, alternados entre azul y negro. Los niveles en negro están espaciados cada 100 kb y en rojo cada 0,5 Mb. Fuente: Elaboración propia.

proteínas, el ~10% al metabolismo del ADN (233), cofactores, vitaminas, grupos prostéticos y pigmentos (228). Además, las secuencias ligadas al metabolismo de carbohidratos, sulfuro, fósforo y nitrógeno, sindicados como importantes en el

metabolismo central de este tipo de bacterias sulfuro oxidantes, aparecen con un ~6% (144), 1,6% (39), 0,9% (20) y un ~3% (69), respectivamente. Por otro lado, esta herramienta identificó 5 secuencias de genes relacionados con metabolismo secundario.

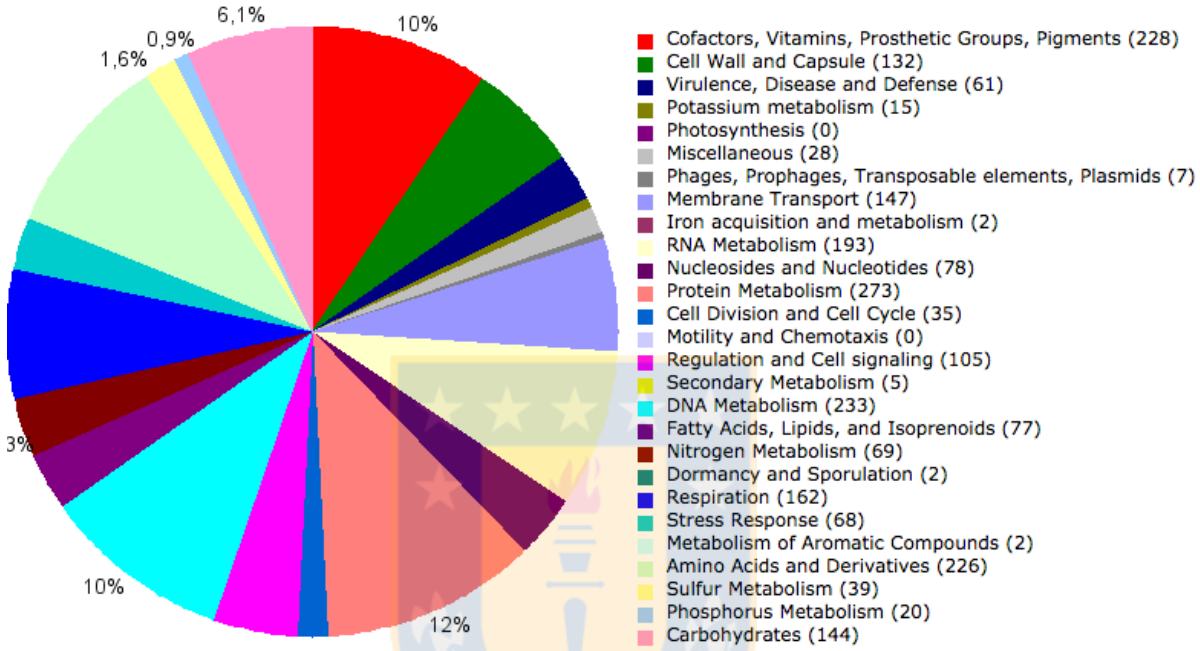


Figura 16. Anotación funcional de genes predichos en el genoma de *Beggiatoa* sp. HS por subsistema.

Gráfico circular de la proporción de secuencias predichas por cada colección de familias de proteínas funcionalmente relacionadas (subsistema) identificado con RAST. La leyenda muestra el nombre del subsistema y entre paréntesis el número de secuencias predichas para cada uno. Fuente: Elaboración propia.

3.2 *Anotación general del genoma de Leptospira sp.*

Tras la anotación del borrador del genoma de *Leptospira* sp. se identificaron 7.151 CDSs, tres ARNr; un 16S y dos 5S, 35 ARNt y un ARNtm. Los dos 5S ARNr

están en *contigs* diferentes (*contigs* 78 y 154) y tienen un tamaño de 111 y 110 pb.

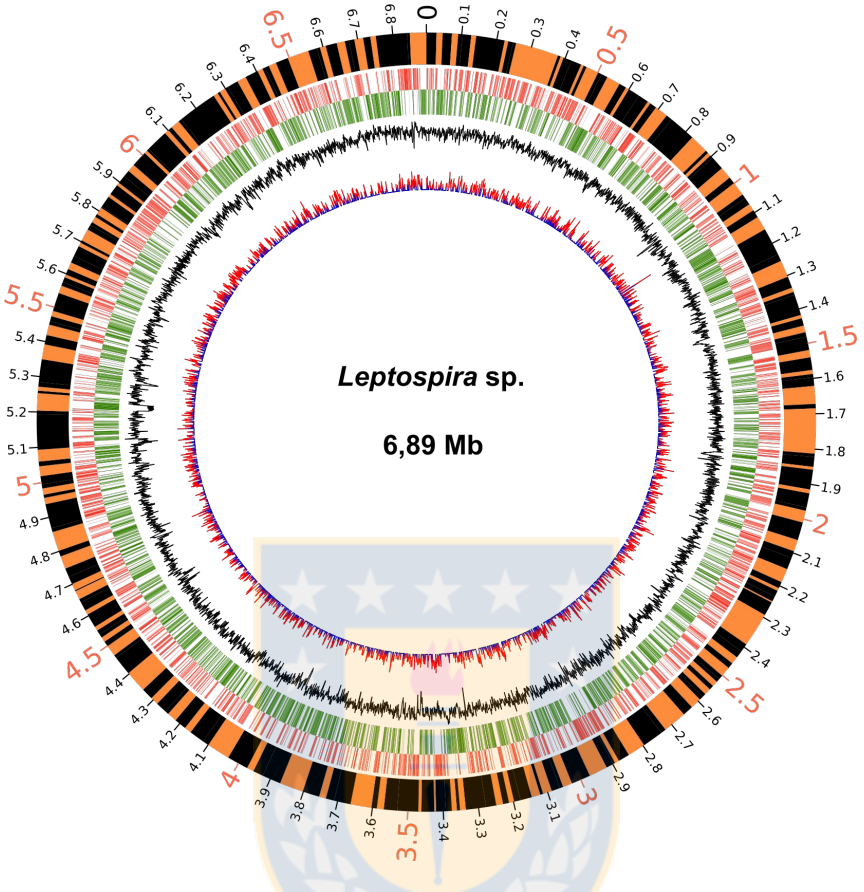


Figura 17. Representación circular del *draft* del genoma de *Leptospira* sp. Desde el anillo más interno hacia fuera: el anillo rojo (positivo) y azul (negativo) muestra el GC *skew* en una ventana =1000 pb; el anillo negro describe el contenido de GC% en una ventana =1000 pb; el anillo verde representa la predicción de genes de la cadena de ADN antisentido y el anillo rojo las CDSs de la cadena de ADN sentido; el anillo más externo muestra todos los *contigs*, alternados entre naranjo y negro. Los niveles en negro están espaciados cada 100 kb y en rojo cada 0,5 Mb. Fuente: Elaboración propia.

Por su parte el gen 16S de ARNr está ubicado en el *contig* 185 y tiene un tamaño de 1.515 pb. De los 7.151 genes, 3.366 están identificados como codificantes

de proteínas hipotéticas, dejando 3.785 secuencias codificantes con un producto de función conocida. La Figura 17 esquematiza al borrador del genoma de *Leptospira* sp. obtenido, destacando en el anillo más externo los *contigs* que lo conforman, e inmediatamente al interior los genes codificantes identificados en la cadena sentido y antisentido del ADN.

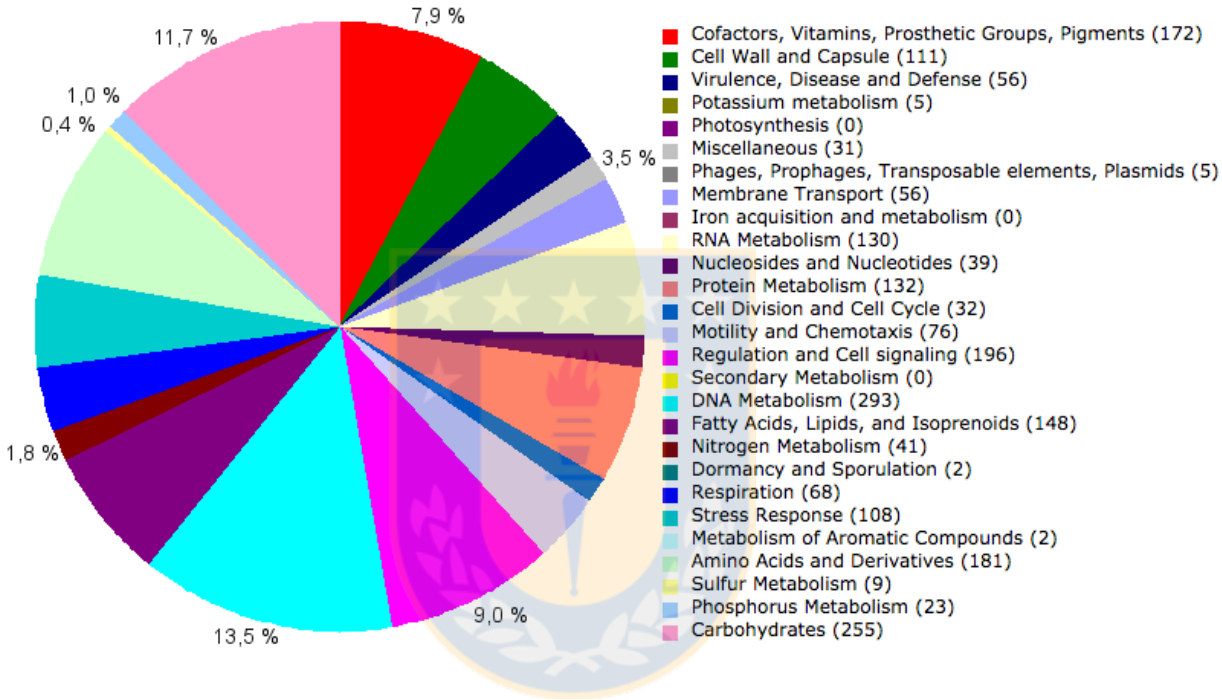


Figura 18. Anotación funcional de genes predichos en el genoma de *Leptospira* sp. por subsistema.

Gráfico circular de la proporción de secuencias predichas por cada colección de familias de proteínas funcionalmente relacionadas (subsistema) identificado con RAST. La leyenda muestra el nombre del subsistema y entre paréntesis el número de secuencias predichas para cada uno. Fuente: Elaboración propia.

Por otra parte, la anotación con RAST identificó a 2.171 secuencias dentro de 372 subsistemas de proteínas (Figura 18). El mayor número de secuencias está relacionado con el metabolismo del ADN con un ~13% (293 secuencias), seguido del

metabolismo relacionado a carbohidratos con un 11,7% (255 secuencias), el de regulación y señal celular con un ~9% (196 secuencias) y el metabolismo relacionado a cofactores, vitaminas, grupos prostéticos y pigmentos con un ~8% (172 secuencias). Además, un ~6% (148 secuencias) aparecen ligado a sintesis de acido grasos, lípidos y isoprenoides. También aparecen como relevantes la movilidad y quimiotaxis con un 3,5% y virulencia, enfermedad y defensa con un 2,5%. Contrariamente, no se identificaron secuencias dentro de subsistemas relacionados al metabolismo secundario, fotosíntesis, adquisición de hierro y solo 2 secuencias asociadas a compuestos aromáticos.

3.3 *Filogenia de Beggiatoa sp.* HS *y Leptospira sp. basado en el gen 16S de ARNr*

Utilizando los genes de 16S del ARNr identificados en el *draft* de genoma de *Beggiatoa* sp. HS (*scaffold* 540; 1.537 pb) y de *Leptospira* sp. (*contig* 185; 1.515 pb) como marcadores filogenéticos, se reconstruyó la filogenia para ambas bacterias. La Figura 19 muestra el árbol filogenético de *Beggiatoa* sp. HS construido utilizando los primeros 50 encuentros en BLAST y tres secuencias de *Ruegeria* (dos secuencias de *Ruegeria atlantica* y uno de *Ruegeria* sp.; un tipo de *Alphaproteobacteria*) como grupo externo. *Beggiatoa* sp. HS fue incluido en un clado no monofilético (probabilidad a posterior = 0,88) conformado por várias *Beggiatoa* spp., no cultivadas, *Beggiatoa* sp. MS-81-6 y *Beggiatoa* sp. Arauama II.

La Figura 20 muestra el árbol filogenético de *Leptospira* sp. construido a partir de la secuencia del gen de 16S ARNr identificada en el *draft* de genoma de *Leptospira* sp. y 50 secuencias tomadas arbitrariamente desde genbank, pertenecientes a distintos géneros del filo *Spirochaeta*, además de tres secuencias del filo *Deferribacter* como grupo externo. El resultado muestra que *Leptospira* sp. se inserta dentro de un multi-clado de *Leptospira* spp. (con 100% de probabilidad) y con un 60% en un clado monofilético de *Leptospira interrogans*.

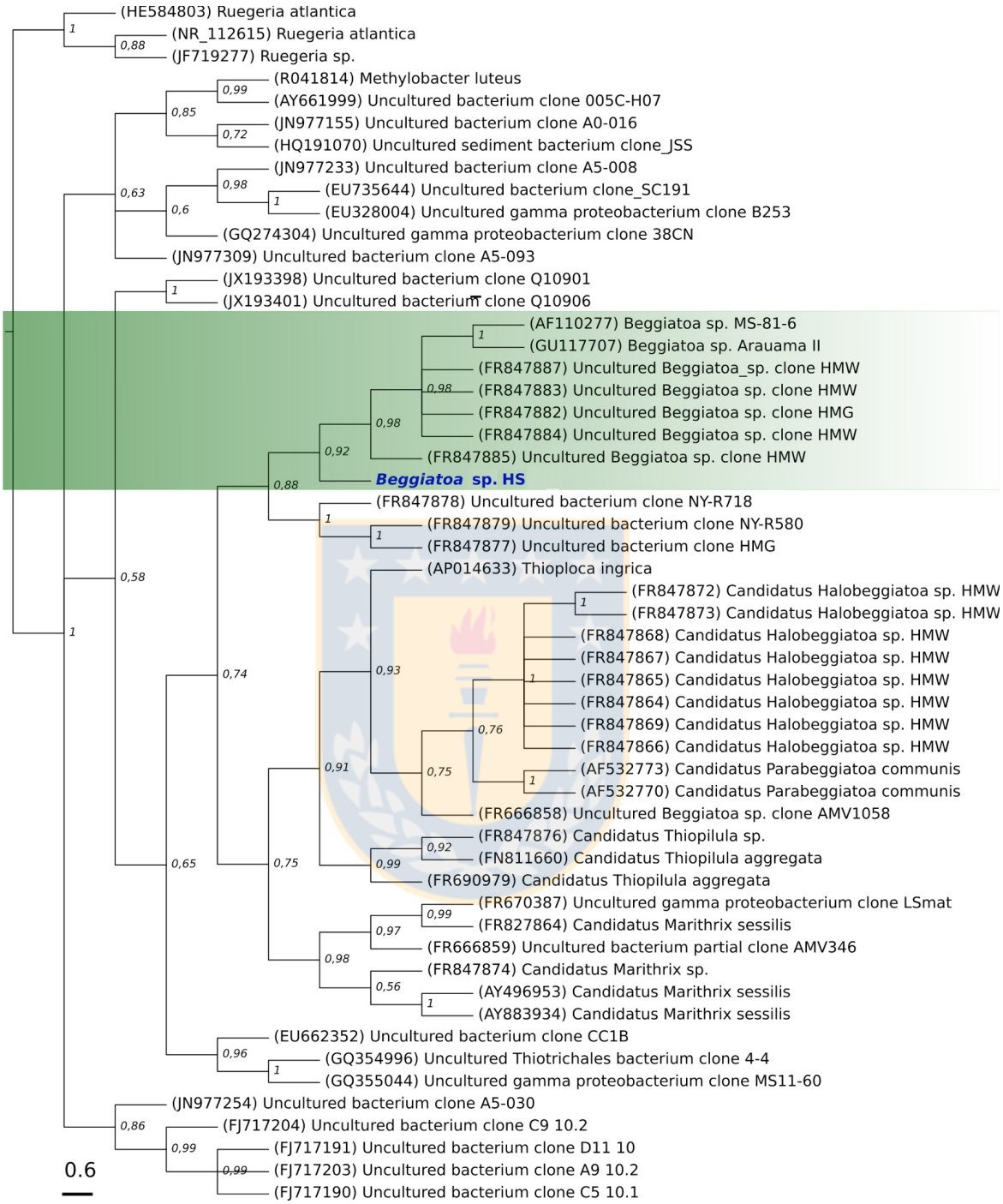


Figura 19. Árbol filogenético de *Beggiatoa* sp. HS.

Construido en base al gen 16S de ARNr y cincuenta secuencias del gen 16S ARNr recuperados desde BLAST. Las secuencias fueron alineadas utilizando MUSCLE y el

árbol filogenético fue construido a través del enfoque bayesiano con la herramienta MrBayes, utilizando el método de cadenas de Markov Monte Carlo (MCMC) para estimar la probabilidad posterior. Se ejecutaron 1.000.000 de generaciones y se utilizó como modelo de sustitución el de GTR (*General Time Reversible*) con una tasa de sustitución = 6. El método de reconstrucción utilizado fue el de máxima verosimilitud y se descartó el 25% de los árboles muestreados. Como grupo externo se utilizaron tres secuencias de *Ruegeria* spp. (*Alphaproteobacteria*). Fuente: Elaboración propia.

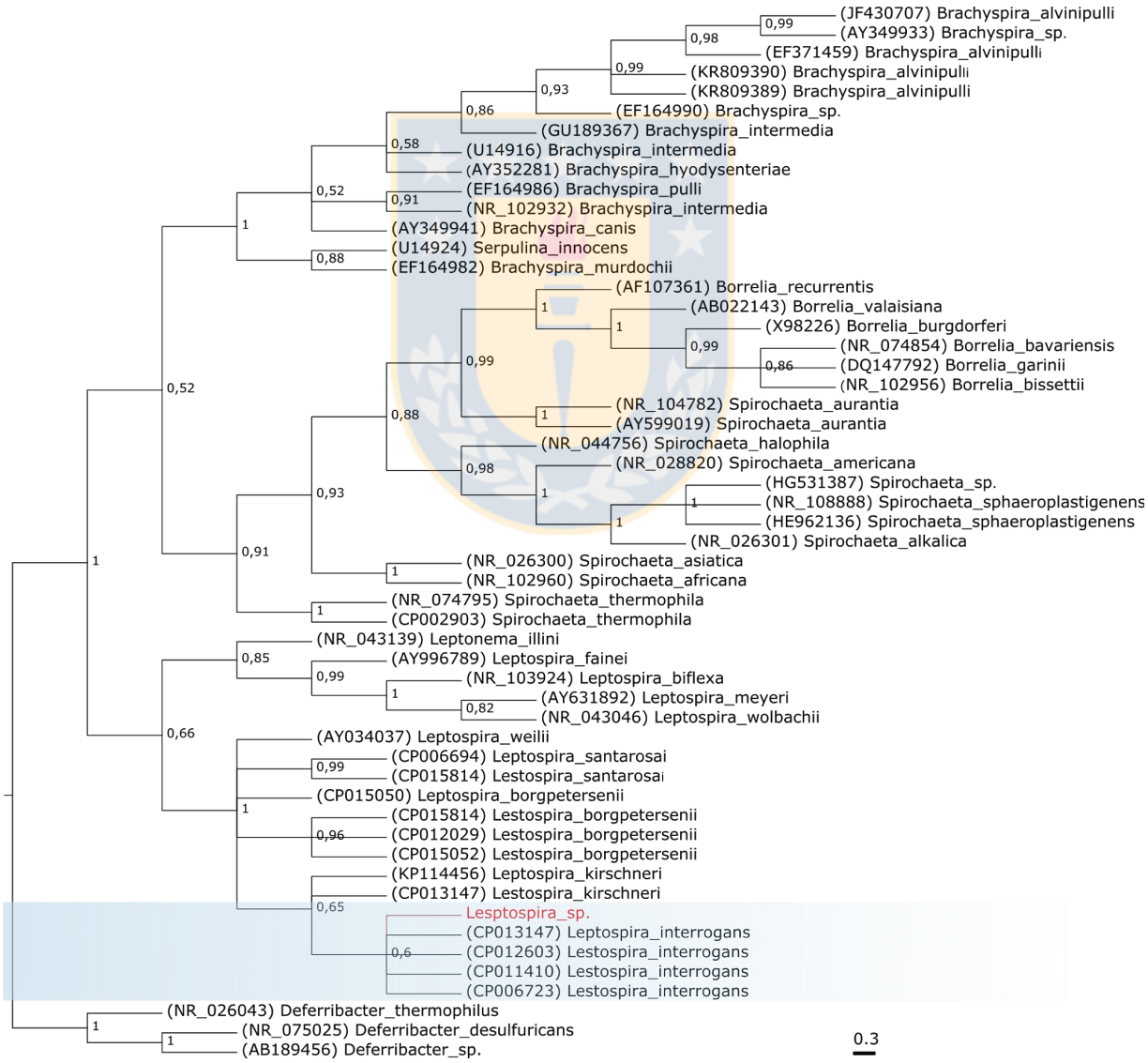


Figura 20. Árbol filogenético de *Leptospira* sp.

Construido usando la secuencia del gen 16S ARNr y cincuenta secuencias del gen 16S ARNr de géneros pertenecientes a Spirochaetae. Las secuencias fueron alineadas utilizando MUSCLE y el árbol filogenético fue construido a través del enfoque bayesiano con la herramienta MrBayes, utilizando el método de cadenas de Markov Monte Carlo (MCMC) para estimar la probabilidad posterior. Se ejecutaron 1.000.000 de generaciones y se utilizó el modelo de sustitución GTR (*General Time Reversible*) con una tasa de sustitución de 6. El método de reconstrucción fue el de máxima verosimilitud y se descartó el 25% de los árboles muestreados. Tres secuencias de *Deferribacter* spp. (*Filo =Deferribacteres*) se utilizaron como grupo externo. Fuente: Elaboración propia.

3.4 *Identificación y anotación funcional de los CGBs implicados con la biosíntesis de MSs en el genoma de Beggiatoa sp. HS*

Se identificaron tres CGBs candidatos implicados en biosíntesis de MSs en el borrador del genoma de *Beggiatoa* sp. HS (Figura 21). Dos idénticos CGBs estarían relacionados con la síntesis de compuestos tipo Terpeno, ubicados en los scaffolds scf-561 y scf-564, ambos de 21.082 pb. El tercer clúster sería de un tipo indeterminado (no canónico) de 41.958 pb, identificado en el *scaffold* scf-559. La Figura 22 muestra la arquitectura y disposición de los genes que forman parte del CGB de tipo indeterminado en el scf-559. Genes que codificarían para dominios biosintéticos aciltransferasa (AT), aminotransferasa (amino), adenilación (A) y proteína portadora (CP; tiolación) fueron identificados. El dominio CP identificado posee subdominios tanto de PCP (proteína portadora de peptidil en NRPSs) y de ACP (proteína portadora de acilo en PKSs). La presencia de los dominios A-CP se utilizó como identificador de los pHMM por antiSMASH, para otorgar el carácter de CGB tipo indeterminado. Por otra parte, la combinación de los dominios esenciales de los módulos NRPS; A y CP (tiolación), junto al dominio AT, típico de enzimas PKS,

indican una arquitectura poco convencional, posiblemente híbrida entre NRPS-PKS. Además, se identificaron genes de transporte como *arpB* (*Antibiotic efflux pump membrane transporter ArpB*) y el gen *rpoD*, el cual codifica para un factor sigma, relacionado con regulación de la transcripción .

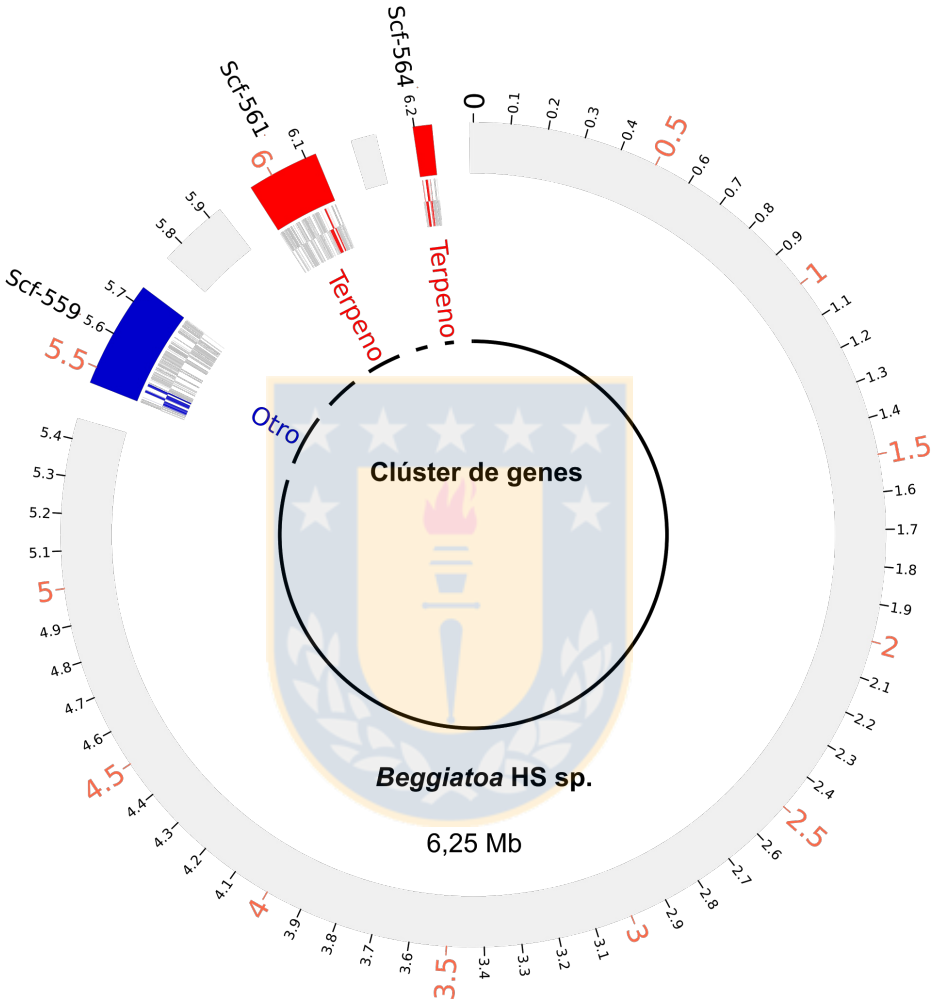
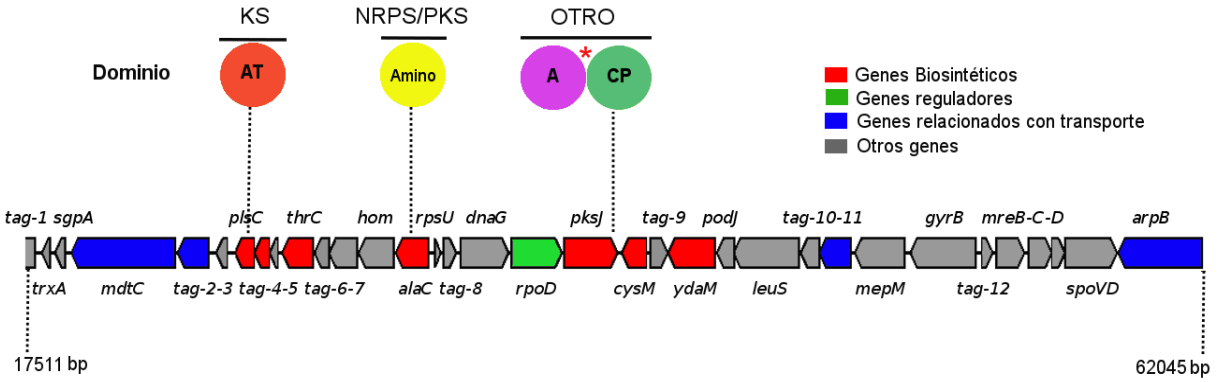


Figura 21. Tipo y ubicación de los CGBs candidatos implicados en biosíntesis de MSs, identificados en el *draft* del genoma de *Beggiatoa* sp. HS.

Los segmentos en rojo contienen la identificación de los CGBs implicados en biosíntesis de compuestos tipo Terpeno. El segmento en azul contiene al CGB tipo indeterminado. En el mismo color que los *scaffolds*, se muestran las secuencias codificantes que forman parte de los CGB en la cadena sentido y antisentido de ADN. Fuente: Elaboración propia.



Gen		Producto
tag-1	-	hypothetical protein
trxA	-	Thioredoxin-1
sgpA	-	Sulfur globule protein CV1 precursor
mdtC	-	Multidrug resistance protein MdtC
tag-2	-	putative efflux pump membrane fusion protein
tag-3	-	hypothetical protein
plsC	-	1-acyl-sn-glycerol-3-phosphate acyltransferase
tag-4	-	D-glycero-beta-D-manno-heptose-1,7-bisphosphate 7-phosphatase
tag-5	-	MazG nucleotide pyrophosphohydrolase domain protein
thrC	-	Threonine synthase
tag-6	-	hypothetical protein
tag-7	-	recombination protein F
hom	-	Homoserine dehydrogenase
alaC	-	Glutamate-pyruvate aminotransferase AlaC
rpsU	-	30S ribosomal protein S21
tag-8	-	Yqey-like protein
dnaG	-	DNA primase
rpoD	-	RNA polymerase sigma factor RpoD
pksJ	-	Polyketide synthase PksJ
cysM	-	Cysteine synthase B
tag-9	-	Rhomboid family protein
ydaM	-	putative diguanylate cyclase YdaM
podJ	-	Localization factor PodJL
leuS	-	Leucine--tRNA ligase
tag-10	-	hypothetical protein
tag-11	-	ABC transporter ATP-binding protein
mepM	-	Murein DD-endopeptidase MepM
gyrB	-	DNA gyrase subunit B
tag-12	-	hypothetical protein
mreB	-	Rod shape-determining protein MreB
mreC	-	Cell shape-determining protein MreC
mreD	-	Rod shape-determining protein MreD
spoVD	-	Stage V sporulation protein D
arpB	-	Antibiotic efflux pump membrane transporter ArpB

Figura 22. Arquitectura del CGB candidato tipo indeterminado identificado en *Beggiatoa* sp. HS (scf 559).

Los segmentos en rojo muestran los genes que codifican para dominios enzimáticos biosintéticos, en azul las que codifican para proteínas de transporte, en verdes las asociadas a regulación y en gris otro tipo de genes. Las esferas muestran los dominios aciltransferasa (AT), aminotransferasa (Amino), adenilación (A) y proteína portadora (CP; tiolación). Sobre las esferas se indica la clase de compuesto al cual se asocian los dominios. EL asterisco en rojo indica los dominios utilizados como identificador del CGB, de acuerdo a las características del sitio activo. Bajo la representación del CGB se listan los genes y productos que están dentro de los límites de clúster. Los genes no identificados se catalogan como tag. Fuente: Elaboración propia.

La Figura 23 muestra la representación de la estructura del CGB relacionado con biosíntesis de un compuesto tipo terpeno, identificado en el *scaffold* scf-561 de *Beggiatoa* sp. HS Este CGB tipo terpeno cuenta con genes asociados a regulación y biosíntesis, entre ellos el gen que codifica para el dominio enzimático de la fitoeno sintasa, utilizado como firma por antiSMASH para clasificar el clúster, y cuyo sitio activo da el carácter identificatorio al dominio (Pythoen_synt - Tipo terpeno).

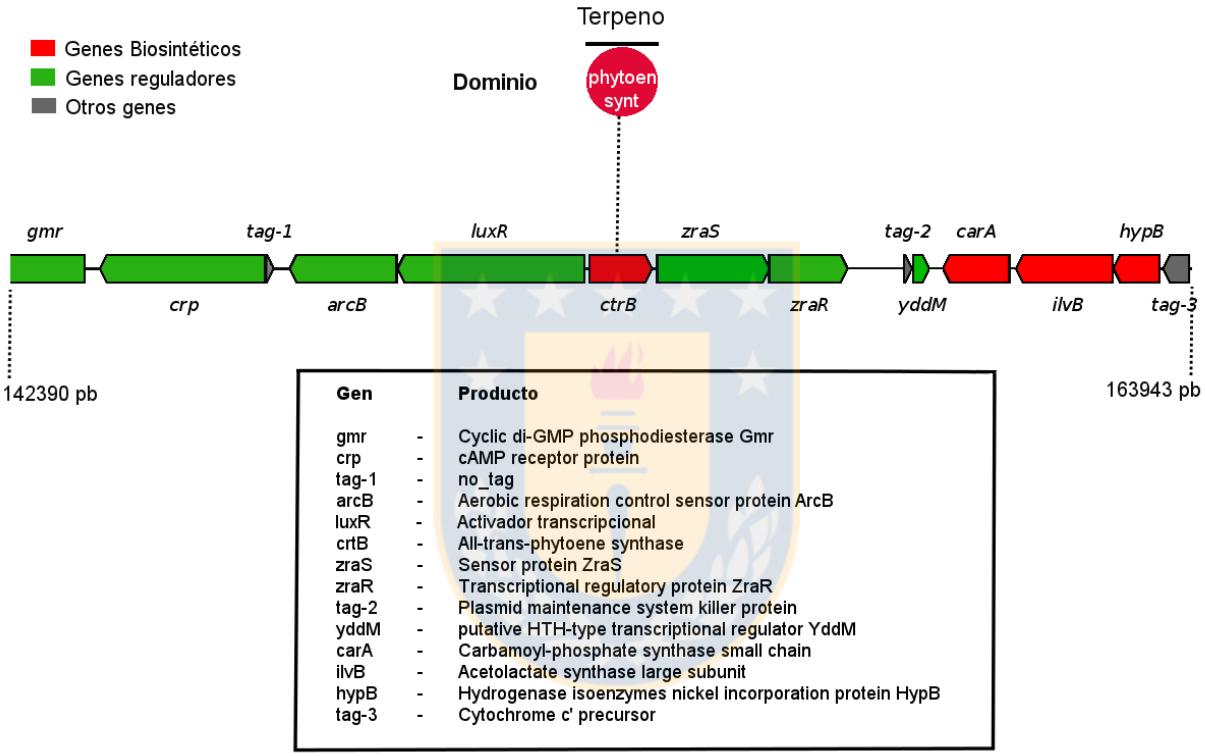


Figura 23. Arquitectura del CGB candidato tipo terpeno identificado en *Beggiatoa* sp. HS (scf 561). Los segmentos en rojo muestran las secuencias de ADN que codifican para dominios enzimáticos biosintéticos, en verde las secuencias que codifican para proteínas reguladoras y en gris a otro tipo de genes. La esfera muestra el dominio identificado. Sobre la esfera se indica la clase de compuesto al cual se asocia el dominio. Bajo la representación del CGB se listan los genes y productos que están dentro de los límites de clúster (*tag* = genes no no conocidos). Fuente: Elaboración propia.

3.5 *Identificación y anotación funcional de los CGBs implicados con la biosíntesis de MSs en el genoma de Leptospira sp.*

Se identificaron cinco CGBs candidatos posiblemente relacionados con biosíntesis de MSs en *Leptospira* sp. (Figura 24). De los cuales, uno estaría envuelto en biosíntesis de homoserina lactona de 21.885 pb en el *contig* 144; uno con arquitectura PKS tipo III de 30.216 pb en el *contig* 172, relacionado a síntesis de Chalconas; dos PKS tipo indeterminados de 28.694 y 10.427 pb en el contig-185 y 200 respectivamente y un PKS tipo indeterminado-I de 43.457 pb en el contig-187. El CGB relacionado con biosíntesis de homoserina lactona (Figura 25) posee un gen (*tag-15*) que codifica para un dominio biosintético identificado como *Autoinducer synthase* (signature pHMM=Autoind_synth), el cual es utilizado como identificador del clúster, además existen tres genes relacionados con regulación (*Crp*, *kinE* y *rpfC*). Por otra parte, el CGB envuelto en biosíntesis de Policétido sintasa tipo III (PKS III) (Figura 26), contiene la predicción de nueve dominios biosintéticos funcionales, entre ellos; acyltransferase (*tag-2*), adenilación (*tag-8*; signature pHMM = AMP-binding), Aminotransferasa (*dapL*; signature pHMM = Aminotran_1_2) y los dominios C y N terminal de *Chalcone synthase* (*tag-9*; signature pHMM = Chal_sti_synt_N y Chal_sti_synt_C) utilizados como firma para clasificar el clúster. Además, un dominio glicosil transferasa (*epsD*; signature pHMM=Glycos_transf_1), dos posibles dominios de epimerización (*rfbE*; signature pHMM=Epimerase) y un gen envuelto en regulación (*pleD*; *Response regulator PleD*). La Figura 27 muestra la arquitectura del CGB candidato PKS tipo otro del *contig* 185, el cual contiene una secuencia que codificaría para una enzima multidominio similar a PKS tipo I (*pksj*; *Phenolphthiocerol synthesis polyketide synthase type I Pks15/1*): KS, AT, hglE y CP. Además, otros tres genes; *pksL* (*Polyketide synthase PksL*), *tag-3* (proteína hipotética) y *pksE* (*Polyketide biosynthesis protein PksE*) que codifica para dominios KR, DH y AT. Arquitectura similar a un híbrido entre PKS tipo I y PKS tipo II. Además, el gen contiguo *fabA* codifica para un tipo de CP, no obstante, podría ser un intermediario de síntesis de ácidos grasos. Todos los dominios tienen relación con producción de

productos derivados de biosíntesis de PKSs.

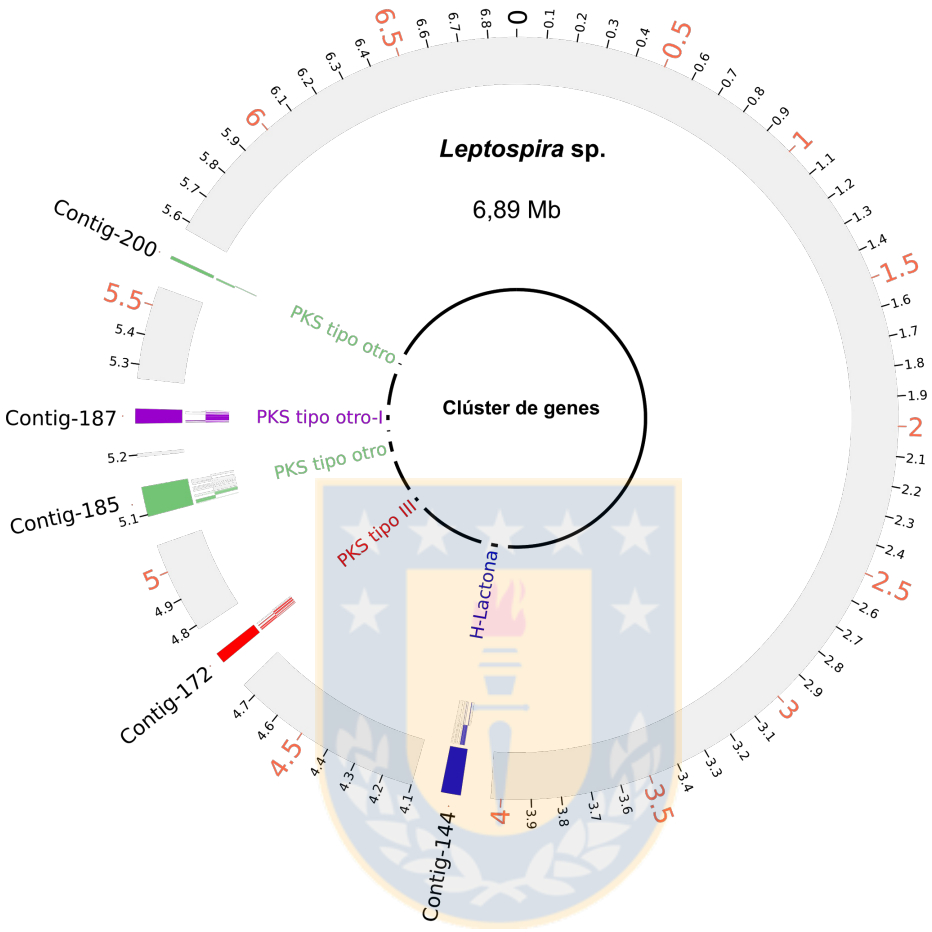
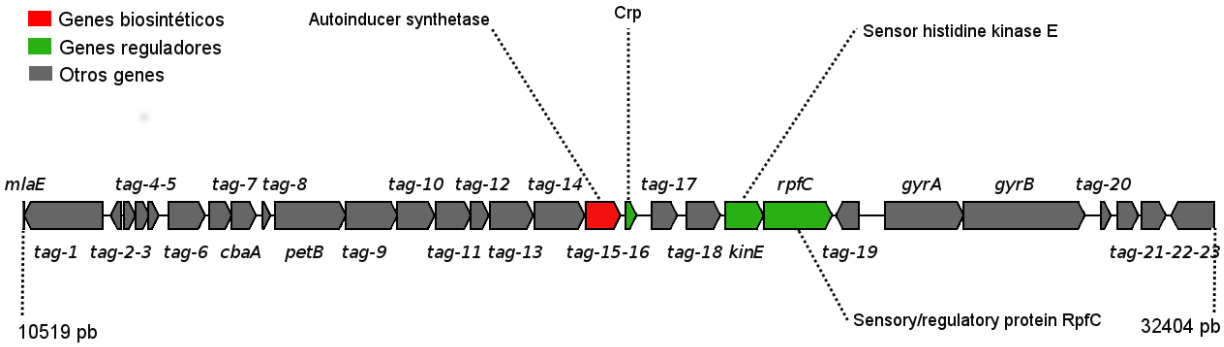


Figura 24. Tipo y ubicación de los CGBs candidatos implicados en biosíntesis de MSs, identificados en el *draft* del genoma de *Leptospira* sp.

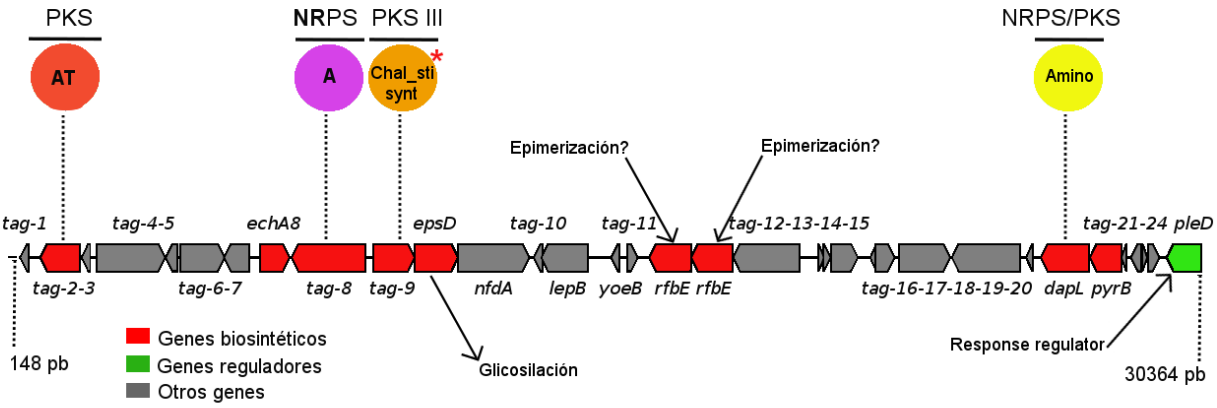
Los segmentos coloreados del anillo exterior indican el *contig* contenedor de un CGB. Hacia el interior, en gris se muestran todas las CDSs en los *contigs* y se destacan las secuencias que forman parte de los CGBs, en la cadena sentido y antisentido de ADN, con el mismo color del *contig* contenedor. Fuente: Elaboración propia.



Gen		Producto	Gen		Producto
mlaE	-	putative phospholipid ABC transporter permease protein MlaE	tag-13	-	hypothetical protein
tag-1	-	hypothetical protein	tag-14	-	Leucine carboxyl methyltransferase
tag-2	-	no_match	tag-15	-	Autoinducer synthetase
tag-3	-	hypothetical protein	tag-16	-	Crp; Activador transcripcional
tag-4	-	hypothetical protein	tag-17	-	hypothetical protein
tag-5	-	hypothetical protein	tag-18	-	hypothetical protein
tag-6	-	Cytochrome C and Quinol oxidase polypeptide I	kinE	-	Sensor histidine kinase E
tag-7	-	no_match	rpfC	-	Sensory/regulatory protein RpfC
cbaA	-	Cytochrome c oxidase subunit 1	tag-19	-	c hypothetical protein
tag-8	-	hypothetical protein	gyrA	-	DNA gyrase subunit A
petB	-	Cytochrome b6	gyrB	-	DNA gyrase subunit B
tag-9	-	Methyl-viologen-reducing hydrogenase, delta subunit	tag-20	-	hypothetical protein
tag-10	-	Tetratricopeptide repeat protein	tag-21	-	hypothetical protein
tag-11	-	hypothetical protein	tag-22	-	hypothetical protein
tag-12	-	hypothetical protein	tag-23	-	hypothetical protein

Figura 25. Arquitectura del CGB candidato tipo homoserina lactona identificado en *Leptospira* sp. (*contig* 144).

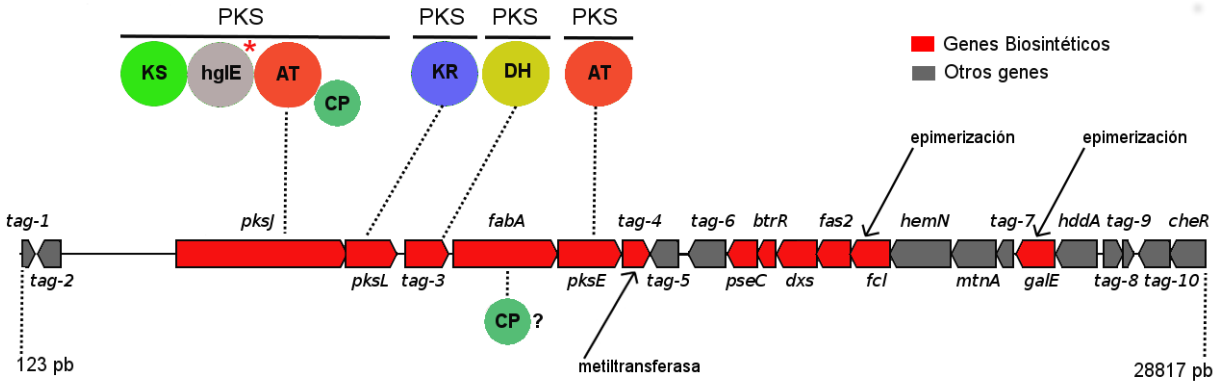
Los segmentos en color rojo muestran las secuencias de ADN que codifican para dominios enzimáticos biosintéticos, en verde las secuencias que codifican para proteínas reguladoras y en gris a otro tipo de genes. Las líneas punteadas indican la presencia de un dominio biosintético y de genes reguladores. Bajo la representación del CGB se listan los genes y productos que están dentro de los límites de clúster. Los genes con identificación indeterminada se catalogan como tag. Las secuencias predichas por antiSMASH que no registran encuentros en las bases de datos aparecen como producto “no_match”. Fuente: Elaboración propia.



Gen		Producto	Gen		Producto
tag-1	-	hypothetical protein	rfbE_5	-	CDP-paratose 2-epimerase ((Epimerasa - familia)
tag-2	-	Acyltransferase	tag-12	-	hypothetical protein
tag-3	-	hypothetical protein	tag-13	-	hypothetical protein
tag-4	-	Papain family cysteine protease	tag-14	-	hypothetical protein
tag-5	-	no_tag	tag-15	-	hypothetical protein
tag-6	-	putative permease YjgP/YjgQ family protein	tag-16	-	hypothetical protein
tag-7	-	hypothetical protein	tag-17	-	no_tag
echA8	-	putative enoyl-CoA hydratase echA8	tag-18	-	hypothetical protein
tag-8	-	AMP-binding (Adenilación)	tag-19	-	putative AAA-ATPase
tag-9	-	Alpha-pyrone synthesis polyketide synthase-like Pks11	tag-20	-	hypothetical protein
epsD	-	Putative glycosyltransferase EpsD	dapL	-	LL-diaminopimelate aminotransferase
nfdA	-	N-substituted formamide deformylase precursor	pyrB	-	Aspartate carbamoyltransferase
tag-10	-	Putative F0F1-ATPase subunit (ATPase_gene1)	tag-21	-	no_tag
lepB	-	Signal peptidase I	tag-22	-	DNA utilization protein GntX
yoeB	-	Toxin YoeB	tag-23	-	no_tag
tag-11	-	hypothetical protein	tag-24	-	no_tag
rfbE	-	CDP-paratose 2-epimerase (Epimerasa - familia)	pleD	-	Response regulator PleD

Figura 26. Arquitectura del CGB candidato PKS tipo III identificado en *Leptospira* sp. (*contig* 172).

Los segmentos en color rojo muestran las secuencias de ADN que codifican para dominios enzimáticos biosintéticos, en verde las secuencias que codifican para proteínas reguladoras y en gris a otro tipo de genes. Las esferas muestran los dominios de aciltransferasa (AT), adenilación (A), Chalcona sintasa (Chal_sti_syny - que incluye Los dominios C y N) y aminotranserasa (Amino). Sobre las esferas se indica la clase de compuesto al cual se asocian los dominios. Las flechas indican posibles dominios opcionales del clúster. El asterisco rojo indica los dominios utilizados como firma del CGB, de acuerdo a las características del sitio activo. Bajo la representación del clúster se listan los genes y productos que están dentro de los límites de este. Los genes no identificados se catalogan como *tag*. Fuente: Elaboración propia.



Gen		Producto	Gen		Producto
tag-1	-	hypothetical protein	dxs	-	1-deoxy-D-xylulose-5-phosphate synthase
tag-2	-	hypothetical protein	fas2	-	Ferredoxin fas2 (transketolase)
pksJ	-	Phenolphthiocerol synthesis polyketide synthase type I Pks15/1	fcl_2	-	NAD-dependent epimerase/dehydratase
pksL	-	Polyketide synthase PksL	hemN	-	Radical SAM domain protein
tag-3	-	hypothetical protein	mtnA	-	Methylthioribose-1-phosphate isomerase
fabA	-	3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase	tag-7	-	hypothetical protein
pksE	-	Polyketide biosynthesis protein PksE	galE	-	UDP-glucose 4-epimerase
tag-4	-	Methyltransferase domain protein	hddA	-	GHMP_kinases_N
tag-5	-	hypothetical protein	tag-8	-	hypothetical protein
tag-6	-	hypothetical protei	tag-9	-	hypothetical protein
pseC	-	DegT/DnrJ/EryC1/StrS aminotransferase	tag-10	-	hypothetical protein
btrR	-	L-glutamine:2-deoxy-scyllo-inosose aminotransferase	cheR	-	Chemotaxis protein methyltransferase

Figura 27. Arquitectura del CGB candidato PKS tipo indeterminado en *Leptospira* sp. (*contig* 185).

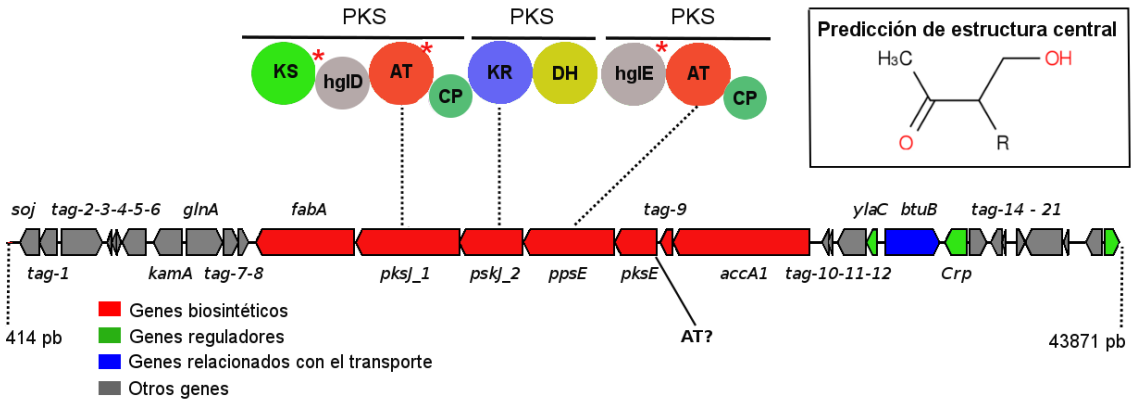
Los segmentos en color rojo muestran las secuencias de ADN que codifican para dominios enzimáticos biosintéticos y en gris a otro tipo de genes. Las esferas muestran los dominios ceto sintasa (KS), aciltransferasa (AT), inusual PKS HglE-like (HglE), proteína portadora (CP; portadora de acilo), ceto reductasa (KR) y deshidratasa (DH). Sobre las esferas se indica la clase de compuesto al cual se asocian los dominios. El signo de interrogación indica que el dominio posee residuos similares a CP, pero no existe consenso. EL asterisco en rojo indica el dominio utilizado como firma del CGB. Bajo la representación del CGB se listan los genes y productos que están dentro de los límites de clúster. Los genes no identificados se catalogan como *tag*. Fuente: Elaboración propia.

La Figura 28 muestra la arquitectura del CGB tipo indeterminado-I ubicado en

el *contig* 187. Los genes *pksJ_1* (*Polyketide synthase PksJ*) y *ppsE* (*Phthiocerol synthesis polyketide synthase type I PpsE*), codificarían para enzimas multidominio similares a PKS tipo I, los cuales forman un módulo individual tipo PKS. El gen *pksj_1* codificaría para los dominios KS, AT, hglD y CP. Siendo hglD un dominio inusual similar a KS. El gen *pksJ_2* codificaría para dominios opcionales KR y DH y el gen *ppsE* codificaría para los dominios hglE, AT y CP. Los dominios inusuales hglE y hglD, son utilizado como firma y otorgan el carácter de otro (desconocido) al CGB y los dominios KS-AT, el carácter canónico de PKS tipo I a *pksj_1* y *ppsE*. Además, se incluyen predicciones de genes relacionados con transporte (*btuB*) y otros involucrados en regulación; *ylac*, *tag-21* y *Crp* un activador de la transcripción de metabolismo secundario al algunas bacterias. Adicionalmente, antiSMASH entregó una predicción aproximada de la estructura central que produciría el CGB, que corresponde a una cetona con cadena lateral (recuadro en Figura 28).

La Figura 29 muestra la estructura del CGB candidato PKS tipo indeterminado identificado en el *contig* 200. El cual está constituido por ocho genes que codifican para dominios biosínteticos. El gen *ppsC* (*Phthiocerol synthesis polyketide synthase type I PpsC*) y *ppsA* (*Phthiocerol/phenolphthiocerol synthesis polyketide synthase type I PpsA*) codificarían para enzimas multidominio; ER-KR y KS-hgIE, respectivamente. Además, los genes *pksN* (*Polyketide synthase PksN*), *eryA* (*Erythronolide synthase*), *pksL* (*Polyketide synthase PksL*), *pksJ* (*Polyketide synthase PksJ*), los cuales codificab para enzimas tipo PKS no multidominio. Por otro lado, *tag-3* y *tag-4* codifican para dominios *Short-chain dehydrogenase,* asociado a compuestos de tipo Lantipeptidos; péptidos sintetizados por ribosomas y modificados postraduccionalmente, producidos por microorganismos. Los dominios identificados son comunes en estructuras que generan compuestos de tipo PKS y Lantipeptidos. El dominio Amino puede estar tanto en PKSs o NRPSs.

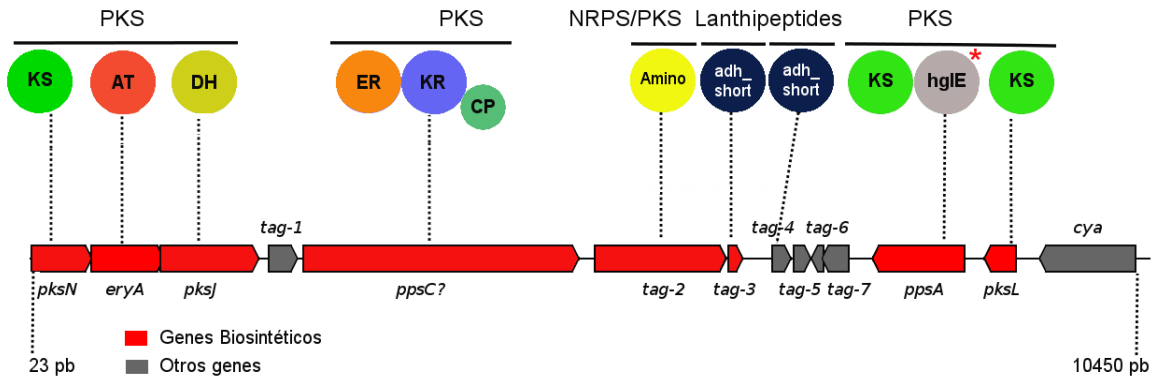
El dominio inusual hgIE fue utilizado como firma en la detección del clúster, y excepto por ppsC, el CGB es similar a un PKS tipo II.



Gen		Producto	Gen		Producto
soj	-	Sporulation initiation inhibitor protein Soj	tag-9	-	Methylmalonyl-CoA carboxyltransferase 12S subunit
tag-1	-	hypothetical protein	accA1	-	Acetyl-/propionyl-coenzyme A carboxylase alpha chain
tag-2	-	hypothetical protein	tag-10	-	Rhodopirellula transposase
tag-3	-	hypothetical protein	tag-11	-	hypothetical protein
tag-4	-	hypothetical protein	tag-12	-	FecR protein
tag-5	-	No_tag	ylaC	-	RNA polymerase sigma factor YlaC
tag-6	-	hypothetical protein	btuB	-	Vitamin B12 transporter BtuB precursor
kamA	-	L-lysine 2,3-aminomutase	Crp	-	DNA-binding transcriptional dual regulator Crp
glnA	-	Glutamine synthetase	tag-14	-	hypothetical protein
tag-7	-	hypothetical protein	tag-15	-	hypothetical protein
tag-8	-	hypothetical protein	tag-16	-	hypothetical protein
fabA	-	3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase	tag-17	-	hypothetical protein
pksJ	-	Polyketide synthase PksJ	tag-18	-	hypothetical protein
pksJ	-	Polyketide synthase PksJ	tag-19	-	No_tag
ppsE	-	Phthiocerol synthesis polyketide synthase type I PpsE	tag-20	-	hypothetical protein
pksE	-	Polyketide biosynthesis protein PksE	tag-21	-	RNA polymerase sigma factor

Figura 28. Arquitectura del CGB candidato PKS tipo indeterminado-I identificado en *Leptospira* sp. (*contig* 187).

Los segmentos en rojo muestran los genes biosintéticos, en azul los relacionados con transporte, en verde los relacionados con regulación y en gris otro tipo de genes. Las esferas muestran los dominios ceto sintasa (KS), aciltransferasa (AT), inusual PKS HglD-*like* (HglD), ceto reductasa (KR), deshidratasa (DH) y inusual PKS Hgle-*like* (HglE). El asterisco rojo indica los dominios utilizados como firma del clúster. En el extremo superior derecho se muestra la predicción de la estructura central producida por el primer y último módulo PKS. Bajo la representación del CGB se listan los genes y productos que están dentro de los límites de clúster (tag = genes no identificados; Not_tag = sin encuentros en bases de datos). Fuente: Elaboración propia. Fuente: Elaboración propia.



Gen		Producto
pksN	-	Polyketide synthase PksN
eryA	-	Erythronolide synthase, modules 3 and 4
pksJ	-	Polyketide synthase PksJ
tag-1	-	hypothetical protein
ppsC	-	Phthiocerol synthesis polyketide synthase type I PpsC
tag-2	-	Aminotran_1_2
tag-3	-	3-ketoacyl-(acyl-carrier-protein) reductase
tag-4	-	short chain dehydrogenase
tag-5	-	hypothetical protein
tag-6	-	no_tag
tag-7	-	hypothetical protein
ppsA	-	Phthiocerol/phenolphthiocerol synthesis polyketide synthase type I PpsA
pksL	-	Polyketide synthase PksL
cya	-	Adenylate cyclase

Figura 29. Arquitectura del CGB candidato tipo indeterminado identificado en *Leptospira* sp. (*contig* 200).

Los segmentos en rojo indican los genes biosintéticos y en gris a otro tipo de genes. Las esferas muestran los dominios ceto sintasa (KS), aciltransferasa (AT), deshidratasa (DH), ceto reductasa (KR), proteína portadora (portadora de acilo: CP), aminotransferasa (Amino), *Short-chain dehydrogenase* (adh_short), enoil reductasa (ER) y inusual PKS HglE-*like* (HglE). El asterisco rojo indica dominio utilizado como firma del clúster. Bajo la representación del CGB se listan los genes y productos que están dentro de los límites de clúster (tag = genes no identificados). Fuente: Elaboración propia.

IV. DISCUSIÓN

1. Pre-proceso de las lecturas en bruto tras la secuenciación

Cualquier flujo de trabajo con datos provenientes de tecnología de secuenciación NGS requiere preprocesar las bibliotecas secuenciadas (Grada y Weinbrecht, 2013) antes de continuar con cualquier análisis, lo que incluye; transformar los archivos a los formatos adecuados para análisis posteriores, filtrar secuencias y bases con baja calidad de secuenciación (bajo puntaje de Phred), remover adaptadores, *primers*, y si corresponde; *linkers* y secuencias identificadoras. El pre-proceso de las cuatro bibliotecas se realizó con el subprograma sffToCA, en cooperación con el programa Pinseq-lite. A pesar que el subprograma ssfToCA no está en si catalogado como una herramienta de pre-proceso, está optimizado (opciones: “-clear 454”, “-trim chop”) para trabajar con datos generados por tecnología 454 GS-FLX.

La tecnología de secuenciación 454 GS-FLX rinde secuencias con una media de longitud de hasta 700 pb, siendo superior a otras tecnologías de secuencias “cortas” (Ej. Illumina). Esto proporciona ventajas para sortear el problema de secuencias repetidas o de ADN complejo (Goodwin *et al*., 2016). No obstante, este tipo de técnica de secuenciación, referenciada como “secuenciación por síntesis” (454 de Roche y Ion Torrent), tienen el problema común de la falta de precisión en la lectura de homopolímeros de extensión mayor a 6-8 pb (Chiu y Miller, 2016), siendo frecuente deleciones e inserciones (*indel*) (Loman *et al*., 2012). Sin embargo, la tasa de error global está a la par con otras plataformas de NGS en regiones libres de homopolímeros (~0,1–15%), siendo la tasa promedio de error cercano al 1% (Gilles *et al*., 2011). Pero mientras la plataforma Ion Torrent se ha mantenido en el campo de NGS en rápida evolución, la plataforma 454 ha sido incapaz de competir con otras plataformas en términos de rendimiento o coste. Esta limitación ha llevado a Roche a descontinuar la plataforma en 2016 (Goodwin *et al*., 2016).

La secuenciación con la plataforma 454 GS-FLX de Roche, de dos bibliotecas *mate-pair* y una *single-end* derivadas de la amplificación de ADN de tres filamentos de *Beggiatoa* sp. HS rindió un total de 1.175.553 lecturas, con un promedio de longitud de 240 y 550 pb, en las bibliotecas *mate-pair* y *single-end*, respectivamente. Por otra parte, la biblioteca *single-end* de *Leptospira* sp. rindió 777.836 lecturas con un tamaño medio de 580 pb. Resultado concordante con las características de la plataforma de secuenciación y las bibliotecas preparadas. Como era previsible, el control de calidad inicial de las lecturas en bruto mostró que la calidad de secuenciación por base, el contenido de bases por secuencia y el contenido de k-meros resultaron deficientes. En consecuencia, el puntaje de calidad, evaluado a través de la herramienta fastQC, indico alta probabilidad de error en la asignación de bases (*Base-calling*) en posiciones por sobre 260 pb en ambas bibliotecas *mate-pair* de *Beggiatoa* sp. HS y en posiciones sobre 450 pb en las bibliotecas *single-end* tanto de *Beggiatoa* sp. HS como en la de *Leptospira* sp. probablemente producto de inserciones o deleciones, no siendo posible determinar cuál es la fuente de error dominante desde este análisis (Beuf *et al*., 2012). Este patrón de irregularidad en el puntaje de calidad en amplicones generados con tecnología 454 ha sido documentado en análisis previos, tal como la revisión realizada por Fuellgrabe *et al*. (2015), en el cual muestra que la calidad de puntaje a través de las bases de amplicones 454 GS-FLX cae después de la posición ~300. Por otra parte, las bibliotecas *mate-pair* de *Beggiatoa* sp. HS muestran un marcado desequilibrio del contenido de bases en ambos extremos, en la posición 1-6 del extremo 5’ y por sobre la posición ~800 en el extremo 3’. Algo esperable, teniendo en cuenta que la distribución de bases en los extremos no es aleatoria y está marcada por la presencia de un *barcode* en el extremo 5’ (Ej. “tcag”), secuencia cebadora y adaptadores en ambos extremos, utilizados en la secuenciación bidireccional de las lecturas. En las bibliotecas *single-end* el desequilibrio sólo se produce en las primeras 10 bases del extremo 5’, debido la presencia de un *barcode*, adaptador y cebador, utilizado en el ciclo único de secuenciación en dirección 5’ a 3’. Además, en las bibliotecas *mate-pair* se identificaron ~50.000 (~15%) secuencias con el *linker* flx

de 44 pb, lo que incide en el desbalance en zonas medias de las lecturas. El alto contenido de *k-meros* (segmentos anormalmente repetidos) en los extremos de las lecturas, está también directamente relacionado con la presencia de *barcodes*, adaptadores y cebadores.

El resultado del pre-proceso controlado a través de fastQC mostró que sffToCA, además de transformar los archivos desde el formato sff a fastq y frg, detectar los *linkers* y dividir las lecturas *mate-pair de* forma exitosa, mejoró el puntaje de calidad por base y removió lecturas cortas, y duplicadas de forma muy efectiva. Las instrucciones para realizar estas tareas están codificadas en los archivos nativos 454 (archivos binarios sff), y basta con indicar a sffToCA que se trata de lecturas tipo 454 para habilitar las instrucciones de pre-proceso. No obstante, esto no fue siempre suficiente, en particular en la biblioteca de *Leptospira* sp., y el puntaje de calidad por base fue mejorado con Prinseq-lite (>30 en puntaje Phred), además se ejecutaron pequeños cortes progresivos en los extremos 5' y 3' de las secuencias, para adecuar el contenido de bases, cuando fue necesario. Un problema frecuente en la etapa de pre-proceso, es la remoción de un porcentaje demasiado alto de secuencias, lo que implicaría perder información que puede afectar los análisis posteriores. Al cabo del pre-proceso, el porcentaje de secuencias sobrevivientes fue de un 72% para las bibliotecas *mate-pair* y de 61% para la biblioteca *single-end* de *Beggiatoa* sp. HS. Por otra parte, el 86% de las lecturas de la biblioteca *single-end* de *Leptospira* sp. sobrevivieron. En el presente trabajo se consideró que la remoción de secuencias fue apropiado y no supuso un riesgo para los análisis posteriores, ya que todas las bibliotecas conservaron a lo menos el 72% de las secuencias, a excepción de la biblioteca Beggiatoa_V3 (61%), tras el tratamiento de las lecturas. Un pre-proceso satisfactorio de las lecturas, aumenta la probabilidad de reconstruir los genomas de manera más precisa y no perder información valiosa que puede provocar un ensamblaje defectuoso.

2. Reconstrucción de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp. a través de ensamblaje *de novo*

Un ensamblaje es una estructura de datos jerárquica que asigna secuencias en *contigs* y *contigs* dentro de *scaffolds*. Donde los *scaffolds* son estructuras con contigs ordenados y orientados (Miller *et al*., 2010). Por tanto, un genoma ensamblado, puede estar a nivel de *contigs* o a nivel de *scaffolds*, siempre y cuando se disponga de bibliotecas *paired-end* o *mate-pair* que entreguen información de distancia entre lecturas, que permitan unir y ordenar *contigs*, siendo en ambos niveles de ordenamiento posible predecir genes. Por otra parte, el ensamblaje *de novo* es la alternativa estándar a la reconstrucción de genomas denominada mapeo, no utilizando un genoma como referencia.

La reconstrucción de los cromosomas tanto de *Beggiatoa* sp. HS y *Leptospira* sp. son borradores del genoma o *draft*, los cuales pueden contener errores y fragmentación. En este contexto, solo los genomas que tras la reconstrucción contienen a lo más 1 error por cada 10.000 bases y no presentan fragmentación son considerados como genomas finiquitados (Campbell *et al*., 2007).

El ensamblaje *de novo* de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp. se llevó a cabo utilizando las herramientas Celera, Newbler y MIRA. Como antecedente es importante señalar que los ensambladores Celera y Newbler utilizan el algoritmo denominado OLC, el cual es apropiado para secuencias mayores a 100 pb. Siendo Newbler desarrollado en particular para trabajar sobre secuencias de tecnología 454 y distribuido por 454 Life Sciences. No obstante, y a pesar de los preceptos que permiten intuir qué herramientas son mejores que otras en determinada situación, es difícil predecir cuál de ellas rendirá finalmente el mejor resultado (Ekblom *et al*., 2014).

De acuerdo a las métricas de ensamblaje, el genoma de *Beggiatoa* sp. HS

ensamblado a nivel de *contigs,* obtenido a través de Celera, resultó ser el de mejor calidad. Logrando un genoma de 6,25 Mb, 806 *contigs*, N50 de 29,3 kb, tamaño máximo de *contigs* de 113 kb, tamaño medio de *contigs* de 7,6 kb y una profundidad de ensamblaje de 32x. En términos generales, la profundidad depende del contenido de GC del genoma, el objetivo del estudio y la plataforma de secuenciación utilizada. Como regla general, se considera que una profundidad de al menos 20x es necesaria en ensamblaje *de novo* de un nuevo organismo (Chiu y Miller, 2016). Por otra parte, los ensamblajes rendidos por Newbler y MIRA presentaron una profundidad de 15x y 11x, respectivamente.

Tanto Celera como Newbler tienen la capacidad de ordenar y unir *contigs* dentro de *scaffolds*, valiéndose de la información de distancia contenida en las secuencias *mate-pair*. A este nivel de ensamblaje, Celera fue nuevamente el que presentó mejores resultados globales, generando un genoma de menor tamaño (6,2 Mb versus 6,3 Mb de Newbler), un N50 (67 kb versus 48 kb de Newbler) y tamaño máximo de *scaffold* superiores a los rendidos por Newbler (289 kb versus 195 kb de Newbler). Sin embargo, Newbler fue levemente superior en tamaño medio de *scaffolds* (12 kb versus 11 kb de Celera) y rindió un número inferior de *scaffolds* (487 vs 560 *scaffolds* de Celera).

En busca de mejorar las métricas de ensamblaje de ambos genomas se siguió la estrategia de conciliar los ensamblajes de Celera, Newbler y MIRA, con el software CISA. El genoma de *Beggiatoa* sp. HS conseguido tras la conciliación con CISA presentó métricas superiores al ensamblaje a nivel de *contigs* obtenido con Celera. No obstante, el genoma resultante tras ordenar los “super” *contigs* generados con CISA dentro de *scaffolds* utilizando la herramienta de *scaffolding* SSPACE, presentó métricas inferiores al genoma a nivel de *scaffolds* generado con Celera, considerando el N50 y tamaño máximo de *scaffold* (Además de aumentar el contenido de GC de 39,8% a 43,7%). Este resultado mostró menor efectividad de SSPACE para unir y ordenar los *contigs* de CISA dentro de *scaffolds* (Unió 22 *contigs*

dentro de *scaffolds*) que Celera, el cual fue más efectivo al ordenar y unir contigs dentro de *scaffolds*. Probablemente, debido a que el flujo de trabajo de SSPACE solo considera a la herramienta Bowtie, altamente recomendado para secuencias <75 pb y la opción “bwasw” del mapeador BWA, apropiado para secuencias provenientes de plataformas Illumina <100 pb (No siendo posible utilizar la opción “mem”; enfocadas en lecturas >100 pb y hasta 1000 pb), para la etapa de alineamiento entre los contigs usados como entradas y las bibliotecas mate pair utilizadas. En consecuencia, y a pesar de la reducción del número de *contigs* y *scaffolds* realizado por CISA y SSPACE, se descartó el genoma generado por esta vía, en beneficio del genoma generado con Celera al nivel de *scaffolds*, el cual fue seleccionado para la etapa de anotación. Por otra parte, no existe un gran número de programas enfocados en hacer *scaffolding* con secuencias *mate-pair* provenientes de tecnologías 454, probablemente debido al retiro de estas plataformas del mercado activo, lo que implica una baja progresiva en su utilización.

Actualmente existen solo algunos borradores de genomas y un par de genomas finiquitados cercanos taxonómicamente a *Beggiatoa* sp. HS (Tabla 5). El tamaño de genoma de *Beggiatoa* sp. HS obtenido está en medio de los tamaños informados en estos genomas. Siendo los de de *C*. Isobeggiatoa de 7,6 Mb (Mußmann *et al*. 2007), *C*. Maribeggiatoa de 4,7 Mb (Mcagregor *et al*. 2013) y *Thioploca ingrica* de 4,8 Mb (Kojima *et al.*, 2014). Más abajo está *Beggiatoa leptomitiformis* de 4,2 Mb (Fomenkov *et al*., 2015). De los genomas existentes; *C*. Isobeggiatoa, *C*. Maribeggiatoa y *C*. Parabeggiatoa son marinos, pero a diferencia de *Beggiatoa* sp. HS poseen enormes vacuolas para almacenar nitrato, mientras que

Tabla 5. Genomas relacionados con *Beggiatoa* sp. HS obtenidos en proyectos anteriores.

Taxón	tamaño	GC (%)	contigs/ scaffolds	ensamblador	publicación	tecnología
Beggiatoa HS sp	6,2	39,8	807(ctg) 560(scf)	Celera	-	454 GS-FLX
C. Maribeggiatoa	4,7	NI	822 (ctg)	Celera	MacGregor *et al*. (2013)	454 GSFLX titanium
C. Isobeggiatoa	7,6	39	5.619 (ctg)	Newbler	Mußmann *et al*. (2007)	454
C. Parabeggiatoa	1,3	43	1.091 (ctg)	Newbler	Mußmann *et al*. (2007)	454
*Thioploca ingrica**	4,8	41	38 (scf)	Platanus	Kojima *et al*. (2014)	Illumina MiSeq
*Beggiatoa leptomitiformis**	4,2	NI	(ctg ?) 1 (scf)	HGAP3 + Quiver	Fomenkov *et al*. (2015)	PacBio RSII

* Genoma finiquitado.
NI = No informa dato.
scf = Número de *scaffolds*.
ctg = Número de *contigs.*

Fuente: Elaboración propia.

Beggiatoa leptomitiformis, Beggiatoa alba (no mostrado en Tabla 5) y *Thioploca ingrica* son de agua dulce y no poseen vacuolas. Por otra parte, el contenido de GC (39,8%) *Beggiatoa* sp. HS, es bastante cercano a los genomas relacionados, el que varía entre el 39 y el 43% en contenido de GC (Tabla 5). Se estima que el contenido de GC en genomas bacterianos variaría entre el 15% y ~85% (Land *et al*., 2015), genomas que además tienden a ser más grandes y tener un mayor contenido de GC en hábitats complejos (Karpinets *et al.,* 2012), lo cual se corresponde con el gran tamaño de genoma de Beggiatoa sp. HS.

La variación de GC se atribuye generalmente a las diferencias en el patrón de mutación entre bacterias y el ambiente en el que se desarrollan, y a pesar que dentro de un mismo grupo pueden existir variaciones, es altamente esperable que un grupo taxonómicamente cercano coincida en el contenido de GC (Hildebrand *et al*., 2010; Lassalle *et al*., 2015).

Por su parte, el borrador del genoma de *Leptospira* sp. de mayor calidad, considerando el menor número de *contigs* (1.057), menor tamaño de genoma (6,8 Mb), mayor N50 (18 kb), tamaño máximo de *contig* (88 kb) y cobertura promedio (23.5x), fue generado por Celera. La alta fragmentación de los genomas entregados por Newbler (1.199 *contigs* >500 pb) y MIRA (1.741 *contigs* > 500 pb) coincide con el gran tamaño de genoma generado, 7,9 y 9,1 Mb (*contigs* >500 pb) respectivamente. Además, ambas herramientas de ensamblaje generaron un alto número de *contigs* y baja cobertura, en ambos casos <11X. Por último, el contenido en GC de los ensamblajes hecho con las tres Celera, Newble y MIRA fue el mismo; ~32% de GC.

Tras la conciliación de los ensamblajes con CISA, se generó un genoma con mejores métricas y se redujo el número de *contigs* de 1.057, en el ensamblaje por Celera, a 257 *contigs*. Aumentó ostensiblemente el N50 y el tamaño máximo y medio de *contigs*, además aumentó levemente el tamaño del genoma de 6,82 a 6,89 Mb. El aumento en el tamaño del genoma tras la conciliación de ensamblajes, se podría haber producido debido a la extensión de regiones superpuestas entre *contigs*. Resultados similares en contigüidad de genoma se ha documentado en otros proyectos genómicos tras utilizar CISA, como en la generación del borrador del genoma de la cyanobacteria *Cyanobacterium Aphanizomenon* (Šulčius, *et al*., 2015) y en la generación de cuatro cepas de *Leptospira interrogans*; Aceguá, RCA, Prea y Capivara (Kremer *et al*., 2016). A diferencia de *Beggiatoa* sp. HS, *Leptospira* sp. no cuenta con lecturas de tipo *mate-pair*, las que posibilitan ordenar *contigs* dentro de *scaffolds*. En consecuencia, el genoma obtenido con CISA a nivel de *contigs*, resultó de mayor calidad que los realizados de forma individual con Celera, Newbler y MIRA

y no modificó el contenido de GC (32,3%), siendo considerado como el de mejor calidad y seleccionado para la etapa de anotación.

Por otra parte, el género Leptospira contiene cepas serológicamente clasificados en más de 250 serovares patógenos, intermedios y saprófitos clasificados en 22 especies diferentes (Adler *et al*., 2014; Kremer *et al*., 2016). Según la literatura revisada en el presente trabajo de tesis, no existiría hasta el momento algún registro previo de alguna especie de leptospira marina de vida libre. No obstante, existen algunos genomas de especies de *Leptospira* disponibles, tomados desde otras fuentes (Tabla 6). El tamaño del genoma de *Leptospira* sp. obtenido está por sobre el rango de tamaños observados en otros genomas de Spirochaeta y Leptospira. En la Tabla 5 se observa que las cuatro cepas de *Leptospira interrogans* poseen un rango de tamaño de genoma entre 4,43 y 4,68 Mb, cercanos a una especie de Spirochaeta denominada *S. smaragdinae* de 4,65 Mb (Mavromatis *et al*., 2010). Por otra parte, el contenido de GC del genoma de *Leptospira* sp. (32,3%) es bastante cercano al reportado en genomas de distintas cepas de *Leptospira interrogans* (~34% de GC). Algo importante de remarcar es que en casi todos los proyectos listados en la Tabla 6, se utilizaron distintas aproximaciones de secuenciación, mezclando tecnologías 454 de Roche y Ion Torrent con Illumina, buscando probablemente mayor contigüidad y cobertura en los ensamblajes resultantes, algo común hoy en día.

Por otra parte, la alta fragmentación de los genomas obtenidos tras el ensamblaje *de novo* puede explicarse en parte debido a la amplificación por MDA del ADN desde un solo filamento bacteriano previo a la secuenciación. Tal como advierte Kojima *et al*. (2014); la amplificación MDA puede ser eficaz para hacer frente a la diversidad genética entre organismos morfológicamente indistinguibles que habitan en el mismo sedimento, pero existe el riesgo de generar secuencias quiméricas durante el proceso de amplificación.

Tabla 6. Genomas relacionados con *Leptospira* sp. obtenidos en proyectos anteriores.

Taxón	tamaño	GC (%)	contigs/ scaffolds	ensamblador	publicación	tecnología
L. HS sp.	6,8	32,3	257 (ctg)	Celera - Newbler - MIRA - CISA	--	454 GS-FLX
L. santarosai	3,93	41,8	111 (ctg)	MAQ (mapeo)	Chou *et al*. (2012)	Illumina
S. smaragdinae *	4,65	48,9	58 (ctg) 1 (scf)	Newbler + hred/Phrap/ Consed	Mavromatis et al. (2010)	Illumina 454 Roche
L. interrogans Acegua	4,68	35	158 (scf)	A5 – SGA – Ray CISA	Kremer *et al*. (2016)	Illumina MiSeq
L. interrogans RCA	4,43	35	89 (scf)	A5 – SGA – Ray MIRA – Newbler - SPAdes – CISA	Kremer *et al*. (2016)	Illumina MiSeq Ion Torrent
L. interrogans Prea	4,44	35	106 (scf)	A5 – SGA – Ray MIRA – Newbler - SPAdes – CISA	Kremer *et al*. (2016)	Illumina MiSeq Ion Torrent
L. interrogans Capivara	4,51	34	160 (scf)	MIRA – Newbler SPAdes – CISA	Kremer *et al*. (2016)	Ion Torrent
L. interrogans PigK151*	4,45	34	(ctg?) 1 (scf)	Roche gsAssembler version 2.8 MIRA 3.4	Alt *et al*. (2015)	Illumina HiSeq 454 FLX-Titanium

* Genoma finiquitado.
NI = No informa dato.
ctg = Número de *contigs.*
scf = Número de scaffolds.

Fuente: Elaboración propia.

En consecuencia es presumible que a causa de la presencia de tales

secuencias quiméricas y/o otras dificultades (Ej. secuencia cortas y regiones repetidas en los genomas) el ensamblaje de los genomas arroje un gran número de *contigs* como los ensamblajes ensayados con Newbler y MIRA de *Beggiatoa* sp. HS y *Leptospira* sp. y aún mayor, como el caso de *C*. Isobeggiatoa (5.619 *contigs*), detallado en la Tabla 5.

3. Anotación de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp.

3.1 *Anotación general y filogenia de los genomas de Beggiatoa sp. HS y Leptospira sp.*

Tras la anotación del *draft* del genoma de *Beggiatoa* HS sp. se identificaron 5.220 genes candidatos, 5 ARNr, 43 ARNt y un ARNtm. Del número total de genes candidatos identificados, solo 3.061 de ellos tienen función conocida. Esto resulta similar al número total de genes identificados en el genoma de *Thioploca ingrica* (*genoma* = 4,8 Mb) (Kojima *et al*., 2014), en el cual se identificaron 3.964 genes. Por otro lado, el número de genes aumenta de forma importante en el genoma de; *C*. Isobeggiatoa (genoma = 7,6 Mb) (Mußmann *et al*. 2007), en el cual se identificaron 6.686. El número de genes predichos en un genoma determinado, pareciera estár relacionado con el tamaño del genoma, ya que a mayor tamaño, mayor es el número de genes predichos. No obstante, sólo una fracción es identificada con una función conocida, quedando el resto como codificantes de proteínas hipotéticas. En efecto, del total de genes identificados en *Beggiatoa* sp. HS, el 58% tiene una función conocida. En consecuencia, en los genomas más fragmentados y con mayor tamaño, se tiende a sobreestimar el número de genes identificados.

En el borrador del genoma de *Leptospira* sp. se detectaron 7.151 genes, tres ARNr; un 16S y dos 5S, 35 ARNt y un ARNtm. La tendencia establecida anteriormente se repite, coincidiendo un alto número de genes, con el tamaño de

genoma generado (*Leptospira* sp. = 6,8 Mb). Sin embargo, solo 3.785 genes poseen función conocida, lo que representa el 52% de los genes totales. A diferencia del genoma de *Beggiatoa* sp. HS, no fue posible detectar el gen 23S de ARNr, probablemente debido a los defectos del ensamblaje. En los genomas relacionados de cepas de *Leptospira interrogans*; Acengua, RCA, Prea y Capivara, se reporta la presencia de un número de genes que va desde 3.591 (RCA) a los 4.146 (Capivara) (Kremer *et al*., 2016). Una cantidad bastante menor que los genes predichos en *Leptospira* sp. Sin embargo, los ARNt identificados van entre los 33 y 37, lo que es muy similar a *Leptospira* sp. Por otra parte, en el genoma de *Leptospira interrogans* PigK151, se identificaron 3,486 genes (genoma = 4.4 Mb), 37 ARNt, una copia de 5S ARNr y dos copias del gen 16S de ARNr (Alt *et al*., 2015).

Tanto en el genoma de *Beggiatoa* sp. HS y *Leptospira* sp. se identificó sólo un gen 16S de ARNr, lo que sugiere ausencia de contaminación con material genético de otros organismos bacterianos. El gen de 16S de ARNr es el marcador molecular más utilizado en la identificación de bacterias, ya que posee una funcionalidad constante y por lo tanto se asume como un cronómetro molecular válido, esencial para inferir relaciones filogenéticas precisas entre organismos (Srinivasan *et al*., 2015). El árbol filogenético construido para *Beggiatoa* sp. HS utilizando el gen 16S de ARNr (Figura 19) confirmó que pertenece a la familia *Beggiatoacea*, la cual agrupa a grandes bacterias sulfuro-oxidantes, ubicándose en un multi-clado conformado por cepas de *Beggiatoa* spp. no cultivadas (Grunke *et al*., 2012), *Beggiatoa* sp. MS-81-6 (Hinck *et al*., 2007) y *Beggiatoa* sp. Arauama II (de Albuquerque *et al*., 2010) con una probabilidad del 88%, todas correspondientes a cepas de tipo litotrofas. Dentro del grupo de bacterias sulfuro-oxidantes, hay bacterias que almacenan altas concentraciones de nitrato dentro de vacuolas, utilizándolo para oxidar sulfuro, mientras aquellas como *Beggiatoa* sp. HS, *Beggiatoa* sp. MS-81-6 y *Beggiatoa* sp. Arauama II carecen de vacuolas, y pueden captar nitrato desde el ambiente o utilizar oxígeno para oxidar sulfuro.

El árbol filogenético de *Leptospira* sp., fue construido en base al gen de 16S de ARNr y el método Bayesiano (Figura 20), mostro que efectivamente pertenece a la familia Spirochaetaceae. De esta forma, el árbol filogenético indico que con un 60% probabilidad la bacteria está relacionada con *Leptospira interrogans* y con 100% con el género *Leptospira*. Formando parte de un multi-clado junto a *L*. weilii, *L*. santarosai, *L*. borgpetersenii y *L*. kirschneri. Todas estas especies son conocidas por causar la grave enfermedad de leptospirosis en animales y infecciones zoonóticas en seres humanos (Kremer *et al*, 2016).

3.2 *Anotación funcional del genoma de Beggiatoa sp. HS y Leptospira sp.*

La anotación funcional del borrador del genoma de *Beggiatoa* sp. HS de acuerdo a la clasificación de los genes candidatos dentro de familias de proteínas funcionalmente relacionadas (Figura 16), mostró que existen genes relacionados al metabolismo de sulfuro (39 genes) y nitrógeno (69 genes), así como también al metabolismo del fósforo (20 genes). La oxidación de sulfuro, acumulado de forma intracelular, a través de nitrato y/o oxígeno, es parte central del metabolismo de bacterias sulfuro oxidantes de la familia Beggiatoacea, consideradas como quimiolitotróficas (Kamp *et al*. 2006; Nelson *et al*., 1983). Además, es conocida la capacidad de acumulación de polifosfatos y su liberación al ambiente por parte de este tipo de bacterias (Brock *et al*., 2012). En consecuencia, *Beggiatoa* sp. HS puede jugar un importante rol en los ciclos biogeoquímicos del nitrógeno, azufre y fósforo, entre otros, considerando los enormes mantos que pueden formar sobre los sedimentos (Gallardo *et al.,* 2013). Por otra parte, solo 5 secuencias aparecen relacionadas con metabolismo secundario, las que son predichas en primera instancia como hormonas vegetales.

En el genoma de *Leptospira* sp. parece menos importante el metabolismo del sulfuro (9 secuencias) y del nitrógeno (41 secuencias), a diferencia de *Beggiatoa* sp.

HS. Sin embargo, existe un gran número de genes relacionados con el metabolismo de carbohidratos (255), a diferencia de *Beggiatoa* sp. HS en cual solo se identificaron 144 genes dentro de ese grupo de proteínas. Otra característica importante en *Leptospira* sp. es la presencia de genes relacionados con movilidad y quimiotaxis (76 secuencias), concordante con la frenética movilidad observada *in vivo* y con la capacidad de infectar y moverse entre tejidos y fluidos densos como la sangre (Lehmann *et al*., 2013). Por otro lado, ambos genomas presentan niveles similares de genes implicados en virulencia y toxinas.

3.3. *Anotación funcional y estructural de los CGBs identificados en los genomas de Beggiatoa sp. HS y Leptospira sp.*

La identificación y caracterización de los CGBs en los genomas de *Beggiatoa* sp. HS y *Leptospira* sp. se llevó a cabo utilizando de forma concertada la herramienta de anotación de CGBs antiSMASH (Utilizando como entrada la anotación de RAST con formato genbank), junto al anotador Prokka, la base de datos no redundante de BLASTP y Pfam, además de la visualización con el programa artemis. Esto permitió evaluar los genes predichos por antiSMASH como parte de los CGBs, y realizar un consenso entre todas las herramientas para cada gen y verificar que las predicciones de genes coincidieran por más de una sola herramienta.

3.3.1 *Anotación funcional y estructural de los CGBs identificados en el genoma de Beggiatoa sp. HS*

En el borrador del genoma de *Beggiatoa* sp. HS se identificaron 3 CGBs candidatos (Figura 21), dos de ellos con arquitectura relacionada a biosíntesis de compuestos de tipo terpeno (CGB tipo Terpeno = 21.082 pb) (Figura 23) y un CGB con una arquitectura indeterminada (CGB tipo indeterminado = 41.958 pb) (Figura

22). Por su parte, en el borrador del genoma de *Leptospira* sp. se identificaron cinco CGBs candidatos (Figura 24), dos con una arquitectura tipo PKS no canónica (PKS indeterminado), (Figura 27-29), un CGB con la arquitectura de un PKS tipo I no convencional; PKS tipo indeterminado-I (Figura 28), un CGB con una arquitectura central relacionada a biosíntesis de compuestos de tipo homoserina lactona (Figura 25) y uno con la arquitectura de un PKS tipo III, comúnmente relacionados con síntesis de Chalconas (Figura 26). En consecuencia, en ambos genomas se identificaron CGBs candidatos, con probable implicancia en síntesis de compuestos naturales. No obstante, el número de CGBs identificados fue reducido en comparación con microorganismos conocidos por su producción de MSs. Por ejemplo, en el borrador del genoma del micro hongo patógeno *Aspergillus ustus* publicado recientemente, se identificaron 52 CGBs candidatos (Pi *et al*., 2015), muy por encima del número identificado en en *Beggiatoa* sp. HS y *Leptospira* sp. Al respecto, es bien conocida también la capacidad de grupos de bacterias como los actinomicetos (Filo: Actinobacteria), los cuales producen gran variedad de compuestos naturales (Nett *et a*l., 2009), los que podrían ser más numerosos de los conocidos hasta hoy, de acuerdo a modelos matemáticos propuestos recientemente (Cimermancic *et al*., 2014). Sin embargo, un número reducido de CGBs, pudiera de igual forma albergar el potencial de generar algún tipo de producto natural novedoso, con aplicaciones prometedoras.

Aunque en bacterias sulfuro-oxidantes de la familia *Beggiatoacea* no se ha investigado profundamente la presencia de CGBs, existen algunos antecedentes previos, por ejemplo; en el genoma de "*Candidatus"* Parabeggiatoa sp. se identificaron varios genes de tipo NRPS y PKS, presumiblemente provenientes desde cianobacterias (Musmann *et al*., 2007). Además, Li *et al*. (2013) reportaron el hallazgo de un nuevo compuesto macrólido, denominado Macplocimine A, aislado desde la megabacteria *C*. marithioploca, recolectada desde sedimentos de Chile central.

Los terpenos abarcan una amplia gama de compuestos complejos que pueden actuar como toxinas, sustancias repelentes o atrayentes (Gershenzon y Dudareva, 2007). Se conocen alrededor de 50.000 metabolitos terpenoides en cerca de 400 familias estructurales distintas, sobre todo, aislados desde plantas y sólo unos pocos procedentes de procariotas. Sin embargo, Yamada y colaboradores (2015) sugieren que la capacidad de biosíntesis de terpenos estaría ampliamente distribuida en bacterias. Tomando como base, la identificación de 262 secuencias putativas que codifican para sintasas de terpeno.

Los 2 CGBs candidatos relacionados con biosíntesis de terpeno identificados en *Beggiatoa* sp. HS son idénticos entre ellos y su tamaño es la mitad del CGB tipo otro (tamaño = 41.958 pb) identificado en el mismo genoma. La duplicidad del CGB tipo terpeno es llamativa, no descartandose algún evento de duplicidad de *contigs*, no detectada y removida previamente. Sin embargo, existen casos como el de la bacteria infecciosa *Orientia tsutsugamushi*, en la que se reportan hasta 20 clúster de genes repetidos, identificados como efectores de proteína (Toft y Andersson, 2010). Estos CGBs tipo terpeno están caracterizados por la presencia del gen *ctrb* que codifica para la enzima Escualeno/Fitoeno Sintasa, la cual contiene el dominio activo "firma" de este CGB (pHMM nombre = phytoene_synth). Contiguo a *ctrb* se encuentran numerosos genes reguladores, entre ellos el gen *luxR,* conocido por su rol de regulador transcripcional*, y Crp* el cual ha sido vinculado previamente a la regulaciòn del metabolismo secundario en bacterias (Gao *et al*., 2012) y puede actuar como represor o activador de la transcripción. Rol que ha sido reportado en la expresión del antibiótico Stambomicina A-D en la bacteria *S. ambofaciens* (Laureti *et al*., 2011). Por otra parte, a través de la base de datos Uniprot (ID = D5KXJ0) se identifica a fitoeno Sintasa como parte de la vía de biosíntesis de fitoeno, que a su vez es parte de la biosíntesis de carotenoides. El gen *ctrB* que codificaría para fitoeno Sintasa en *Beggiatoa* sp. HS, posee alta identidad con genes de *Thioploca ingrica*, *Beggiatoa alba*, *Beggiatoa leptimitiformis* y *Candidatus* Parabeggiatoa sp. Sugiriendo que tal vez la capacidad de biosíntesis de terpenos es una característica

común de este grupo. Por otra parte, el CGB tipo terpeno identificado, es similar en un 14% con un clúster identificado previamente en la bacteria metano-oxidante *Methylobacter tundripaludum* (Wartiainen *et al*., 2006).

El CGB candidato tipo indeterminado (Figura 22), identificado en el genoma de *Beggiatoa* sp. HS contiene un gen (gen *pkj*) que codifica para una enzima con dominios A (pHMM = AMP-binding) y CP (proteína portadora; pHMM = PP-binding), los que sirven como identificadores para el CGB. En la primera etapa de la ruta biosintética, el dominio A cataliza una adenilación dependiente de ATP de un aminoácido. El monómero se transfiere entonces a un CP, al cual post-traduccionalmente se le ha dotado con un brazo de fosfopanteteína, llamado dominio de tiolación (T) o proteína portadora de peptidil (PCP) en los NRPS y proteína portadora de acilo (ACP) en los ensamblajes multienzimáticos de tipo PKS (Millano *et al*., 2013). La detección del tándem de dominios A-CP (CP: tiolación) dentro de una secuencia, es utilizada como identificador para clasificar un CGB candidato como indeterminado por antiSMASH. Esto se debe a la ausencia del dominio de condensación C en el módulo. De hecho, la regla fundamental para determinar un "verdadero" módulo NRPS es la detección de la combinación de módulos A-C. Ya que existen numerosas enzimas con la estructura de dominios A-CP (CP: tiolación) que no son "verdaderos" NRPSs, tales como aquellas que están dedicadas a la síntesis de aminoácidos no proteinogénicos (Comunicación personal con el Dr. Marnix Medema, Wageningen UR, Holanda). Además, se identifica la presencia de un dominio AT (gen *plsC*), el cual forma normalmente parte esencial de los CGB tipo PKS, y son responsables de cargar un ACP con un acil-coenzima-A específico. A pesar que la secuencia aminoacídica codificada por el gen *pksj* presenta cierta homología con la enzima envuelta en la vía de síntesis del antibiótico Bacillaene, un inhibidor de la síntesis de proteínas en procariotas, producida por *Bacillus subtilis* (Patel *et al*., 1995). No obstante la ausencia de un dominio de condensación, este CGB puede estar relacionado con síntesis de algun MS no canónico.

3.3.2 *Anotación funcional y estructural de los CGBs identificados en el genoma de Leptospira sp.*

Los CGBs candidatos identificados en el borrador del genoma de *Leptospira* sp. son superiores en número y diversidad que los CGBs candidatos identificados en el borrador del genoma de *Beggiatoa* sp. HS. El primero de los 5 CGBs identificados tiene una arquitectura de tipo homoserina lactona (Figura 25). La identificador de este cúster corresponde al dominio biosintético codificado en el gen *tag-15* (Nombre de gen indefinido), el que codifica para un dominio *Autoinducer Synthetase* (pHMM = Autoind_synth) o Acyl-homoserina Lactona. La arquitectura de este CGB luce sencilla en relación a otros CGBs, como; PKS tipo Otro (*Contig* 185) Figura 27) y PKS tipo otro-I (*Contig* 187) (Figura 28), ya que solo está compuesto por una enzima biosintética (Autoind_synth), y tres genes reguladores (Crp, *kinE* y *rpfC*). El gen *Crp* codifica para *Cyclic AMP receptor protein*, el cual y como se mencionó en la descripción del CGB tipo terpeno identificado en *Beggiatoa* sp. HS, es un regulador transcripcional que controla diversos procesos celulares en bacterias, incluyendo síntesis de MSs. Este efecto regulatorio fue demostrado por Gao y colegas (2012), los que evidenciaron que *Crp* actúa como un regulador clave del metabolismo secundario y producción de antibióticos en la bacteria *Streptomyces coelicolor,* demostrando que puede coordinar y ejercer como precursor del flujo del metabolismo primario al secundario. Los productos derivados de homoserina lactona son conocidos por estar implicados en el proceso de *quorum sensing*, el cual actúa en bacterias evaluando la densidad local de la población, a través de pequeñas moléculas y péptidos (Waters y Bassler, 2005), y controla la expresión génica en respuesta al crecimiento de densidad celular. Este CGB presenta poca similitud con algún otro clúster conocido.

Los CGBs candidatos identificados en los *contigs* 185 (Figura 27) y 200 (Figura 29) en *Leptospira* sp. no clasifican dentro de ninguna categoría conocida (tipo indeterminado). Sin embargo, el gen *pksJ* (4.136 pb) del CGB candidato tipo

indeterminado (tamaño = 28.694 pb) identificado en el *contig* 185, codifica para una enzima multidominio, muy similar a un módulo PKS tipo I, contando con los tres dominios esenciales KS, AT y CP, excepto por el tipo de dominio PKS inusual hglE, el cual es un tipo similar a dominios KS de enzimas PKSs (hgiE: *Heterocyst glycolipid tipe lE)* y es el que le da el carácter o identificador al CGB. La base de datos Uniprot identifica al gen *pksJ* (ID = B2HIL7) como homóloga de una enzima tipo PKS (en *Mycobacterium marinum*) relacionada con la biosintesis de acidos grasos. Contiguo al gen *pksJ* se ubica el gen *pksL*, el cual codifica para un dominio opcional KR y es sindicado como un intermediario de la síntesis de un tipo de antibiótico denominado Bacillaene producido por una cepa de *Bacillus subtilis*. Adicionalmente, se encuentra el dominio opcional DH (gen *tag-3*) y un par de enzimas (genes *fcl* y *galE*) relacionada a epimerización. *hglE* es una versión inusual de genes identificados en cianobacterias como *Nostoc punctiforme* (Campell *et al*., 1997). En esta cianobacteria se determinó que *hglE* está involucrada en la síntesis de glicolípidos de heterocistos (proceso diferenciador en la cual se genera un ambiente intracelular anóxico y se genera una capa externa aislante de polisacáridos y glicolípidos). De acuerdo a la búsqueda BLAST del CGB, se determina una similaridad de un 22% con varias cepas de M*icrocystis aeruginosa* y un 28% con un CGB de *Nostoc* sp. PCC (Heterocyst_glycolipids biosynthetic BGC0000869_c1). Estas observaciones sugieren la posibilidad de que este CGB haya sido incorporado desde cianobacterias por transferencia horizontal. Sin embargo, no se podría descartar la posibilidad de que el CGB esté implicado en la síntesis de algún tipo de antibiótico similar Bacillaene, o a otro tipo de compuesto indeterminado.

Por otro lado, el tipo de enzimas y distribución del CGB tipo indeterminado identificado en el *contig* 185 no posee una estructura tipo PKS modular canónica. Estando conformado por una enzima multidominio (gen *pksJ*) y tres dominios PKS codificados en enzimas diferentes, de tipo no modular (*pksL*, *tag-3* y *pksE*). Una situación similar a lo que ocurre en los CGBs PKS tipo II, los que comúnmente se organizan conteniendo dominios catalíticos en enzimas independientes. Lo que

confiere al presente CGB una arquitectura híbrida, una especie de estado transitorio entre PKS I y PKS II. A este respecto, Wang *et al*. (2014) sugieren que las enzimas biosintéticas no modulares, no canónicos, no son tan extrañas en bacterias, ya que Proteobacteria, Actinobacteria, Firmicutes, y Cyanobacterias contienen este tipo de arreglos de genes y son susceptibles de producir una amplia variedad de productos naturales de tipo PKS y NRPS.

El segundo CGB PKS tipo otro; identificado en el *contig* 200 (Figura 29), al igual que el CGB anterior, posee un dominio activo de tipo inusual y similar a PKS; hglD, junto a un dominio KS, codificados en el gen *ppsA*, el cual fue utilizado por antiSMASH como firma. Sin embargo, la anotación realizada con prokka y la base de datos Uniprot, *ppsA* y *ppsC*, sugieren que estos estarían ligados a la biosíntesis de lípidos._Por otra parte,_los tres genes contiguos; *pksN*, *eryA* y *pksJ*, presentan alta homología con enzimas que toman parte en biosíntesis de antibióticos. *pksN* (Uniprot ID = O31782) codifica un dominio KS y *pksJ* (Uniprot ID = P40806) un dominio DH, los que aparecen vinculados a la biosíntesis del antibiótico *Bacillaene*. Por su parte el gen *eryA* (Uniprot ID = Q03132) codifica para un dominio AT homólogo de un intermediario de la biosíntesis del antibiótico eritromicina (Khosla *et al*., 2007). Junto a todos los dominios putativos de tipo PKS aparecen los genes *tag-3* y *tag-4,* las que codifican para dominios de tipo Lantipéptido (adh_short). Los Lantipéptidos pertenecen a una familia de péptidos policíclicos que se caracterizan por la presencia de los aminoácidos tioéter lantionina y metillantionina (Piper *et al*., 2009). Estos compuestos están ampliamente distribuidos en especies taxonómicamente distantes y agrupan desde compuestos antimicrobianos a antialodínicos (Zhang *et al*., 2015). Este CGB candidato al igual que el anterior CGB tipo Otro posee una arquitectura no convencional. Excepto por la secuencia *ppsC*, la cual aparenta codificar para una enzima tipo PKS I multidominio, todos los demás dominios son codificados de forma individual, característica fundamental de las PKS tipo II. Toda esta complejidad, abre la posibilidad de síntesis desde glicolípidos hasta algún tipo de antibiótico, como por ejemplo alguno similar a *bacillaene,* eritromicina u otros tan variados como

lantibióticos. Es muy sugerente que en un tramo de ~10 Kb se encuentren contiguos numerosos genes biosintéticos relacionados con vías de biosíntesis de MSs, y es altamente probable que aquel tramo esté implicado en la síntesis de algu MS indeterminado. Por otra parte, a pesar que no se identifican genes reguladores, estos podrían están alejados del clúster y ejercer acción reguladora de igual forma, si el clúster fuera funcional.

Los CGBs candidatos PKS tipo III (Figura 26) y PKS tipo indeterminado-I (Figura 28) identificados en los *contigs* 172 y 187, respectivamente, están dentro de dos de las tres categorías canónicas de PKSs. No obstante, el CGB PKS tipo indeterminado-I, posee un componente no canónico con la presencia de un inusual dominio PKS (hglE). EL CGB candidato PKS tipo III de ~30 kb de extensión posee un gen (*tag-8*) que codifica para los dominios C y N de Chalcona sintasa (CHS), utilizado como identificador. Las CHSs son una superfamilia de enzimas PKS tipo III presentes en plantas y bacterias que forman homodímeros (Austin y Noel, 2003). Su único sitio activo en cada monómero cataliza la iniciación, extensión y reacción de ciclación de forma iterativa para formar productos de policétidos. A pesar de su simplicidad estructural, las PKSs tipo III producen una amplia gama de compuestos tales como chalconas, acridina, floroglucinoles, estilbenos y lípidos de resorcinol (Yu *et al*., 2012). En plantas las CHSs son enzimas muy importantes, las cuales catalizan la primera etapa de la biosíntesis de flavonoides, los cuales toman parte en defensa antimicrobiana, pigmentación, fotoprotección UV y fertilidad del polen. Por otra parte, los dominios CHS identificados en el CGB tipo III de *Leptospira* sp., sería parte de la familia Chalcona Sintasa/Estilbeno Sintasas (CHS/STS), las cuales presentan mucha más divergencia entre ellas que las presentes en plantas y son consideradas como “similares” a CHSs. Contrariamente a las CHSs previamente identificadas en plantas, las CHS/STS prefieren moléculas de partida distintas, además difieren en el número de adiciones de acetilo que catalizan, y su mecanismo de terminación de cadena, incluyendo los patrones alternativos de ciclación intramolecular (Austin y Noel, 2003). Desde que la primera PKS tipo III en bacteria se descubrió en 1999 (Funa *et al.*,

1999), se han caracterizado cinco grupos de PKS tipo III en bacterias hasta el 2012, en base a las estructuras de los productos que generan (Yu *et al*., 2012). Entre ellos los tipos *RppA* en *Streptomyces griseus*, el tipo *PhD* identificado en *Pseudomonas fluorescens, el cual genera un producto de* floroglucinol a partir de tres unidades de malonil-CoA y el tipo *alkylpyrones synthases*, el cual incluye a germicidina sintasa, relacionado con la construcción de varios alquilpironas con propiedades antifúngicas en *S. coelicolor* (Claydon *et al*., 1987). Por otra parte, asociado a este clúster se identificó una secuencia que codifica para un regulador de respuesta (*Response regulador PleD*), el cual es un regulador global. Normalmente este tipo de genes se encuentran fuera de los CGBs y pueden influir de forma pleiotrópica en la producción de MSs (Park y Choi, 2015). Además, se identificó la presencia de dos genes que codifican para un dominio AT y A. No obstante, estos no serían requeridos, ya que cada sitio activo de la enzima CHS/STS posee la capacidad de catalizar por sí mismo la reacción de iniciación, elongación y ciclización de un producto de policétido de forma iterativa (Yu *et al*., 2012).

El último CGB candidato fue identificado en el *contig* 187, y corresponde a un CGB PKS tipo indeterminado-I (Figura 28). Este codifica para tres enzimas *(pksj_1*, *pks_2* y *ppsE*) multidominio tipo PKS I, las que se organizan en módulos individuales. Las enzimas PKS tipo I generalmente consisten en grandes enzimas multidominio, organizadas en módulos (Centeno-Leija *et al*., 2016). Cada módulo es responsable de un solo ciclo de alargamiento de la cadena de policétidos, y el número de módulos con frecuencia se correlaciona con el número de ciclos de extensión. Un ejemplo prototípico de esta subfamilia de enzimas es la 6-desoxieritronolida B sintasa (DEBS), que participan en la biosíntesis de eritromicina (Khosla *et al*. 2007). En las enzimas PKS tipo I cada módulo contiene dominios CP, KS y AT que extienden la secuencia lineal de un intermediario por dos átomos de carbono. El AT carga el CP con un bloque de construcción de un acil-CoA específico, y el KS corriente abajo cataliza la formación del enlace carbono carbono entre el intermedio y el acil-ACP (Dutta *et al*,. 2014). A pesar que el CGB identificado estaría formado por enzimas

multidominios, este difiere un tanto de los modelos arquetípicos. En particular el gen *pks_1* codifica para los tres dominios esenciales en un módulo PKS tipo I; KS, AT y CP. No obstante, posee un subdominio adicional e inusual denominado hgID, similar al dominio KS de las PKSs. Por otra parte, el gen contiguo *pksj_2* codifica solo para dos dominios opcionales KR y DH, estando ausente los dominios KS, AT y CP, y el gen *ppsE* codifica para los tres dominios esenciales, pero el dominio KS es reemplazado por el dominio inusual hglE similar a PKS (KS). En consecuencia, el presente CGB posee una arquitectura general tipo PKS I, pero la presencia de los dominios inusuales hglE y hglD, le agregan además el carácter de indeterminado. Por otro parte, no se logra identificar algún dominio Tioesterasa (TE), el que debiese estar en el módulo final, ejecutando la liberación del producto de policétido, a través de hidrólisis o ciclación (Dutta *et al*., 2014). La ausencia del dominio TE, pudiera deberse a que esta coincidiendo en ubicación con algún otro dominio, en calidad de subdominio, siendo pasado por alto por las herramientas de anotación, o debido a las fallas de ensamblaje, lo que pudo truncar alguna de las secuencias que originalmente lo codificaba. Por otro lado, el producto del gen *pksE*, contiguo al gen *ppsE*, es controversial. El anotador Prokka lo reconoce como un gen que codifica para una enzima PKS (Uniprot ID = O34787). No obstante, Blastp y Pfam, indican un dominio activo relacionado con a una proteína portadora de acilo S-malonyltransferase FABD. La ontología génica de *pksE* presente en Uniprot, coincide con la identificación funcional anterior, pero también lo ubica como un intermediario importante en la síntesis del antibiótico Bacillaene (41% de identidad con *pksE* de *Bacillus subtili*), actuando probablemente como aciltransferasa. Sumado a los genes biosintéticos, se identificaron dos factores de transcripción; *ylaC* y *tag-21*. Estos factores de iniciación sigma promueven la unión de la ARN polimerasa a sitios de iniciación específicos y luego son liberadas. En medio de estos dos factores se identificó al activador transcripcional *Crp*, el cual ha sido relacionado con regulación del metabolismo secundario y de la producción de antibióticos. De esta forma, se podría estimar que este conjunto de genes, junto al transportador *btuB*, tendrían el potencial de participar en la síntesis de algún metabolito, probablemente de tipo

antibiótico. Y a pesar de solo contar con tres módulos individuales, se conoce la capacidad de generar productos naturales con un reducido número de módulos no iterativos, como el conocido fármaco anti-colesterol lovastatina, producto natural sintetizado por dos módulos PKS no iterativos en *Aspergillus terreus* (Kennedy *et al*., 1999). Por otro lado, este CGB resultó ser similar en un 12% con un CGB identificado en las Proteobacterias *Hyphomonas* sp. y *Hyphomonas jannaschiana*, aislada desde una afloración termal del Océano Pacífico (Jannasch y Wirsen, 1984).

Adicionalmente, antiSMASH rinde una predicción de la estructura central del hipotético producto natural que generaría este CGB (Ver recuadro del extremo superior derecho en la Figura 28). Este modelo considera las modificaciones catalizadas por los dominios opcionales de ceto reductasa, deshidratasa y enoil reductasa que influyen en el estado redox de los grupos ceto en policétidos (Weber *et al*., 2015). La estructura predicha corresponde a una cetona con cadena lateral y su predicción se realizó en base al sitio activo del módulo AT, utilizando 24 aminoácidos firmas y el método propuesto por Minowa *et al*., (2007).

Un aspecto importante a considerar, es que el sistema enzimático codificado en los CGBs tipo PKS I que se relacionan con biosíntesis de productos de policétidos y el sistema enzimático responsable de la síntesis de ácidos grasos (FAS) comparten muchas similitudes, incluyendo la utilización común de precursores, similar química, estructuras y diseño arquitectónico general (Smith y Tsai, 2007). Por lo tanto, no se puede descartar que este arreglo de genes pudiese estar envuelto también en la generación *de novo* de ácidos grasos.

Reportes de la presencia de CGBs implicados en biosíntesis de MSs en bacterias del género *Leptospira* son muy escasos en la literatura, lo que convierte a los CGBs identificados en *Leptospira* sp. en esta tesis, en alguno de los pocos reportes existentes hasta el dia de hoy.

V. CONCLUSIONES

A partir de los resultados de esta tesis es posible concluir que:

La reconstrucción de los genomas de *Beggiatoa* sp. HS y *Leptospira* sp. a través de ensamblaje *de novo*, utilizando las herramientas Celera, Newbler y MIRA, generaron ensamblajes fragmentados. Siendo el de mejor calidad en ambos casos el que generó Celera. La estrategia de conciliación de ensamblajes con la herramienta CISA reduce el número de *contigs* de forma significativa en ambos casos. Sin embargo, el genoma final de *Beggiatoa* sp. HS resultó con mejores métricas que el ensamblaje de *Leptospira* sp. debido a la ventaja que le brindan las lecturas *mate-pair.*

La anotación de ambos genomas permitió identificar un solo gen de 16S de ARNr en cada genoma. La filogenia construida en base a ellos ubico a las bacterias cercanas a los género Leptospira y Beggiatoa. El hallazgo de Leptospira en ambiente marino es inédito, además de presentar un genoma más complejo que otras.

La anotación funcional de genes en *Beggiatoa* sp. HS es concordante con observaciones previas en genomas relacionados, en cuanto a la presencia de genes implicados en el metabolismo del sulfuro, nitrógeno y fósforo. Mientras que *Leptospira* sp. posee más genes implicados en el metabolismo de carbohidratos, y un importante parte relacionada con movimiento y quimiotaxis, algo común en Leptospira.

La anotación funcional y estructural de CGBs candidatos implicados en biosíntesis de MSs a través de herramientas bioinformáticas, resultó en la identificación de 3 CGBs candidatos en *Beggiatoa* sp. HS y 5 en *Leptospira* sp. Los dos CGBs candidatos tipo terpeno identificados en *Beggiatoa* sp. HS podrían generar MSs de tipo antibiótico y el CGB indeterminado, estaría probablemente relacionado

con síntesis de aminoácidos no proteinogénicos. Por otra parte, de los 5 CGBs identificados en *Leptospira* sp. el CGB PKS tipo III, es altamente probable que esté implicado en síntesis de algún compuesto de tipo antibiótico. Mientras, que el CGB PKS tipo indeterminado-I posee arquitectura similar a PKSs tipo I canónica, implicados en entre otros, en la síntesis de antibióticos. Por otro lado, el CGB tipo homoserina lactona, estaría implicado en el mecanismo de *Quorum sensing* y los CGBs tipo PKS indeterminado de los *contigs* 185 y 200, son de proyección incierta.

En líneas generales se puede concluir que tanto *Beggiatoa* sp. HS y *Leptospira* sp. albergan CGBs con el potencial para generar algún MSs de tipo de proyección farmacológica..

VI. PROYECCIONES

Los genomas *Beggiatoa* sp. HS y *Leptospira s*p. generados tienen el estatus de *drafts* o borradores de genomas. Sin embargo, re-secuenciando el ADN de ambos organismos con una estrategia de secuenciación de tipo Illumina, y agregandola a la ya obtenida, se lograría mayor profundidad de ensamblaje, disminuyendo la fragmentación de los genomas, lo que junto al diseño de partidores en los extremos de los *gaps* remanentes, permitirá obtener genomas finiquitados, sin fragmentación y llevar los errores al mínimo.

Por otra parte, una caracterización más profunda de los CGBs más prometedores, podría abrir la posibilidad de explorar su posible activación por medio de técnicas de ingeniería en laboratorio.

La anotación de ambos genomas abre la posibilidad de situarse sobre objetivos distintos a los tratados en esta tesis, tales como identificación y caracterización de proteínas transportadoras, con alguna aplicación en diseño de fármacos. O explorar las capacidades de resistencia a fármacos, entre otras muchas posibilidades que brinda la exploración genómica.

VII. REFERENCIAS

Abt, Birte, Han, Cliff, Scheuner, Carmen, Lu, Megan, Lapidus, Alla, Nolan, Matt, Cheng, Jan-Fang. (2012). Complete genome sequence of the termite hindgut bacterium Spirochaeta coccoides type strain (SPN1 T), reclassification in the genus Sphaerochaeta as Sphaerochaeta coccoides comb. nov. and emendations of the family Spirochaetaceae and the genus Sphaerochaeta. Standards in genomic sciences, **6**(2), 194.

Adler, Ben. (2014). Leptospira and leptospirosis (Vol. **387**): Springer.

Alt, David P., Wilson-Welder, Jennifer H, Bayles, Darrell O., Cameron, Caroline, Adler, Ben, Bulach, Dieter M., Darby, Alistair C. (2015). Complete genome sequence of Leptospira interrogans serovar Bratislava, strain PigK151. Genome announcements, 3 (3), e00678-00615.

Austin, Michael B., & Noel, Joseph P. (2003). The chalcone synthase superfamily of type III polyketide synthases. Natural product reports, **20**(1), 79-110.

Aziz, Ramy K., Bartels, Daniela, Best, Aaron A., DeJongh, Matthew, Disz, Terrence, Edwards, Robert A., Kubal, Michael. (2008). The RAST Server: rapid annotations using subsystems technology. BMC genomics, **9**(1), 1.

Baas-Becking, LGM. (1925). Studies on the sulphur bacteria. Annals of Botany, **39**(155), 613-650.

Banskota, Arjun H., McAlpine, James B., Sørensen, Dan, Ibrahim, Ashraf, Aouidate, Mustapha, Piraee, Mahmood, Zazopoulos, Emmanuel. (2006). Genomic analyses lead to novel secondary metabolites. Journal of Antibiotics, **59**(9), 533.

Beuf, Kristof D., Schrijver, Joachim D., Thas, Olivier, Criekinge, Wim V, Irizarry, Rafael A., & Clement, Lieven. (2012). Improved base-calling and quality scores for 454 sequencing based on a Hurdle Poisson model. BMC bioinformatics, **13**(1), 303.

Boetzer, Marten, Henkel, Christiaan V., Jansen, Hans J., Butler, Derek, & Pirovano, Walter. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, **27**(4), 578-579.

Britstein, Maya, Devescovi, Giulia, Handley, Kim M., Malik, Assaf, Haber, Markus, Saurav, Kumar, Gilbert, Jack A. (2016). A new N-Acyl homoserine lactone

synthase in an uncultured symbiont of the Red Sea sponge Theonella swinhoei. Applied and environmental microbiology, **82**(4), 1274-1285.

Brock, Jörg, Rhiel, Erhard, Beutler, Martin, Salman, Verena, & Schulz-Vogt, Heide N. (2012). Unusual polyphosphate inclusions observed in a marine Beggiatoa strain. Antonie van Leeuwenhoek, **101**(2), 347-357.

Campbell, A., Malcolm, Heyer, Laurie J.A., & Laurie, J., Heyer. (2007). Discovering genomics, proteomics, and bioinformatics.

Campbell, Elsie L., Cohen, Michael F., & Meeks, John C. (1997). A polyketide-synthase-like gene is involved in the synthesis of heterocyst glycolipids in Nostoc punctiforme strain ATCC 29133. Archives of microbiology, **167**(4), 251-258.

Centeno-Leija, Sara, Guzmán-Trampe, Silvia, Rodríguez-Peña, Karol, Bautista-Tovar, Diana, Espinosa, Allan, Trenado, Miriam, & Sánchez, Sergio. (2016). Different Approaches for Searching New Microbial Compounds with Anti-infective Activity New Weapons to Control Bacterial Growth (pp. 395º-431): Springer.

Cimermancic, Peter, Medema, Marnix, Claesen, Jan, Kurita, Kenji, Brown, Laura C Wieland, Mavrommatis, Konstantinos, Clardy, Jon. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell, **158**(2), 412-421.

Claydon, N., Allan, M., Hanson, J.R., & Avent, A.G. (1987). Antifungal alkyl pyrones of Trichoderma harzianum. Transactions of the British Mycological Society, **88**(4), 503-513.

Compeau, Phillip E.C., Pevzner, Pavel A., & Tesler, Glenn. (2011). How to apply de Bruijn graphs to genome assembly. Nature biotechnology, **29**(11), 987-991.

Chevreux, Bastien, Wetter, Thomas, & Suhai, Sándor. (1999). Genome sequence assembly using trace signals and additional sequence information. Paper presented at the German conference on bioinformatics.

Chiu, Charles, & Miller, Steve. (2016). Next-Generation Sequencing. Persing D, Tenover F, Hayden R, Ieven M, Miller, 461.

de Albuquerque, Julia Peixoto, Keim, Carolina Neumann, & Lins, Ulysses. (2010). Comparative analysis of Beggiatoa from hypersaline and marine environments. Micron, **41**(5), 507-517.

Debnath, Mousumi, Paul, A.K., & Bisen, P.S. (2007). Natural bioactive compounds and biotechnological potential of marine bacteria. Current pharmaceutical biotechnology, **8**(5), 253-260.

Dutta, Somnath, Whicher, Jonathan R., Hansen, Douglas A., Hale, Wendi A., Chemler, Joseph A., Congdon, Grady R., Smith, Janet L. (2014). Structure of a modular polyketide synthase. Nature, **510**(7506), 512.

Edgar, Robert C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, **32**(5), 1792-1797.

Ekblom, Robert, & Wolf, Jochen B.W. (2014). A field guide to whole-genome sequencing, assembly and annotation. Evolutionary applications, **7**(9), 1026-1042.

Ewing, Brent, & Green, Phil. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research, **8**(3), 186-194.

Fedorova, Natalie D., Moktali, Venkatesh, & Medema, Marnix H. (2012). Bioinformatics approaches and software for detection of secondary metabolic gene clusters. Fungal secondary metabolism: methods and protocols, 23-45.

Fei, Peng, Chuan-xi, Wang, Yang, Xie, Hong-lei, Jiang, Lu-jie, Chen, Uribe, Paulina, Yun-yang, Lian. (2013). A new 20-membered macrolide produced by a marine-derived Micromonospora strain. Natural product research, **27**(15), 1366-1371.

Fernandes, Prabhavathi. (2015). The global challenge of new classes of antibacterial agents: an industry perspective. Current opinion in pharmacology, 24, 7-11.

Finn, Robert D., Tate, John, Mistry, Jaina, Coggill, Penny C., Sammut, Stephen John, Hotz, Hans-Rudolf, Sonnhammer, Erik LL. (2008). The Pfam protein families database. Nucleic acids research, **36**(suppl 1), D281-D288.

Fomenkov, Alexey, Vincze, Tamas, Grabovich, Margarita Y, Dubinina, Galina, Orlova, Maria, Belousova, Elena, & Roberts, Richard J. (2015). Complete genome sequence of the freshwater colorless sulfur bacterium Beggiatoa leptomitiformis neotype strain D-402T. Genome announcements, **3**(6).

Fuellgrabe, Marc W., Herrmann, Dietrich, Knecht, Henrik, Kuenzel, Sven, Kneba, Michael, Pott, Christiane, & Brüggemann, Monika. (2015). High-throughput, amplicon-based sequencing of the CREBBP gene as a tool to develop a universal platform-independent assay. PloS one, **10**(6), e0129195.

Funa, Nobutaka, Ohnishi, Yasuo, Fujii, Isao, Shibuya, Masaaki, Ebizuka, Yutaka, & Horinouchi, Sueharu. (1999). A new pathway for polyketide synthesis in microorganisms. Nature, **400**(6747), 897-899.

Gallardo, V.A., Fonseca, A., Musleh, S.S., & Espinoza, C. (2013). Extrapolations of Standing-Stocks of Big Bacteria in Humboldt Eastern Boundary Current Ecosystem (HEBCE). Oceanography: Open Access, 2013.

Gallardo, Victor A. (1977). Large benthic microbial communities in sulphide biota under Peru–Chile subsurface countercurrent. Nature, **268**(5618), 331-332.

Gallardo, Víctor Ariel, Espinoza, Carola, Fonseca, Alexis, & Musleh, Selim. (2013). Las grandes bacterias del Sulfureto de Humboldt. Gayana (Concepción), **77**(2), 136-170.

Gao, Chan, Mulder, David, Yin, Charles, & Elliot, Marie A. (2012). Crp is a global regulator of antibiotic production in Streptomyces. MBio, **3**(6), e00407-00412.

Gärtner, Andrea, Ohlendorf, Birgit, Schulz, Dirk, Zinecker, Heidi, Wiese, Jutta, & Imhoff, Johannes F. (2011). Levantilides A and B, 20-membered macrolides from a Micromonospora strain isolated from the mediterranean deep sea sediment. Marine drugs, **9**(1), 98-108.

Gershenzon, Jonathan, & Dudareva, Natalia. (2007). The function of terpene natural products in the natural world. Nature chemical biology, **3**(7), 408-414.

Gerwick, William H., & Fenner, Amanda M. (2013). Drug discovery from marine microbes. Microbial ecology, **65**(4), 800-806.

Gilles, André, Meglécz, Emese, Pech, Nicolas, Ferreira, Stéphanie, Malausa, Thibaut, & Martin, Jean-François. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC genomics, **12**(1), 245.

Giske, Christian G., Monnet, Dominique L., Cars, Otto, & Carmeli, Yehuda. (2008). Clinical and economic impact of common multidrug-resistant gram-negative bacilli. Antimicrobial agents and chemotherapy, **52**(3), 813-821.

Goodwin, Sara, McPherson, John D., & McCombie, Richard W. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, **17**(6), 333-351.

Grada, Ayman, & Weinbrecht, Kate. (2013). Next-generation sequencing:

methodology and application. Journal of Investigative Dermatology, **133**(8), 1-4.

Grünke, S., Lichtschlag, Anna, Beer, D de, Felden, J., Salman, V., Ramette, A., Boetius, A. (2012). Mats of psychrophilic thiotrophic bacteria associated with cold seeps of the Barents Sea. Biogeosciences, **9**(8), 2947-2960.

Gulder, Tobias A.M., & Moore, Bradley S. (2009). Chasing the treasures of the sea-bacterial marine natural products. Current opinion in microbiology, **12**(3), 252-260.

Hildebrand, Falk, Meyer, Axel, & Eyre-Walker, Adam. (2010). Evidence of selection upon genomic GC-content in bacteria. PLoS Genet, **6**(9), e1001107.

Hinck, Susanne, Neu, Thomas R., Lavik, Gaute, Mussmann, Marc, De Beer, Dirk, & Jonkers, Henk M. (2007). Physiological adaptation of a nitrate-storing Beggiatoa sp. to diel cycling in a phototrophic hypersaline mat. Applied and environmental microbiology, **73**(21), 7013-7022.

Huelsenbeck, John P., & Ronquist, Fredrik. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics, **17**(8), 754-755.

Ichikawa, Natsuko, Sasagawa, Machi, Yamamoto, Mika, Komaki, Hisayuki, Yoshida, Yumi, Yamazaki, Shuji, & Fujita, Nobuyuki. (2013). DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. Nucleic acids research, **41**(D1), D408-D414.

Imhoff, Johannes F., Labes, Antje, & Wiese, Jutta. (2011). Bio-mining the microbial treasures of the ocean: New natural products. Biotechnology Advances, **29**(5), 468-482.

Ishoey, Thomas, Woyke, Tanja, Stepanauskas, Ramunas, Novotny, Mark, & Lasken, Roger S. (2008). Genomic sequencing of single microbial cells from environmental samples. Current opinion in microbiology, **11**(3), 198-204.

Jannasch, Holger W., & Wirsen, C.O. (1984). Chemosynthetic microbial mats of deep-sea hydrothermal vents. Microbial Mats: Stromatolites, 121-131.

Jenke-Kodama, Holger, Müller, Rolf, & Dittmann, Elke. (2008). Evolutionary mechanisms underlying secondary metabolite diversity Natural Compounds as Drugs **Volume I** (pp. 119-140): Springer.

Kamp, Anja, Stief, Peter, & Schulz-Vogt, Heide N. (2006). Anaerobic sulfide oxidation with nitrate by a freshwater Beggiatoa enrichment culture. Applied and environmental microbiology, **72**(7), 4755-4760.

Kanehisa, Minoru, & Goto, Susumu. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, **28**(1), 27-30.

Kang, Han-Young. (2012). Total Synthesis of Pikromycin and Related Macrolide Antibiotics. Rapid Communication in Photoscience, **1**(2), 59-59.

Karpinets, Tatiana V., Park, Byung H., & Uberbacher, Edward C. (2012). Analyzing large biological datasets with association networks. Nucleic acids research, **40**(17), e131-e131.

Keller, Simone, Schadt, Heiko S., Ortel, Ingo, & Süssmuth, Roderich D. (2007). Action of atrop-Abyssomicin C as an Inhibitor of 4-Amino-4-deoxychorismate Synthase PabB. Angewandte Chemie International Edition, **46**(43), 8284-8286.

Khosla, Chaitan, Tang, Yinyan, Chen, Alice Y, Schnarr, Nathan A, & Cane, David E. (2007). Structure and mechanism of the 6-deoxyerythronolide B synthase. Annu. Rev. Biochem., **76**, 195-221.

Kojima, Hisaya, Ogura, Yoshitoshi, Yamamoto, Nozomi, Togashi, Tomoaki, Mori, Hiroshi, Watanabe, Tomohiro, Fukui, Manabu. (2015). Ecophysiology of Thioploca ingrica as revealed by the complete genome sequence supplemented with proteomic evidence. The ISME journal, **9**(5), 1166-1176.

Kong, De-Xin, Jiang, Ying-Ying, & Zhang, Hong-Yu. (2010). Marine natural products as sources of novel scaffolds: Achievement and concern. Drug discovery today, **15**(21), 884-886.

Kremer, Frederico S., Eslabão, Marcus R., Jorge, Sérgio, Oliveira, Natasha R., Labonde, Julia, Santos, Monize N.P., Forster, Karine M. (2016). Draft genome of the Leptospira interrogans strains, Acegua, RCA, Prea, and Capivara, obtained from wildlife maintenance hosts and infected domestic animals. Memórias do Instituto Oswaldo Cruz, **111**(4), 280-283.

Krzywinski, Martin, Schein, Jacqueline, Birol, Inanc, Connors, Joseph, Gascoyne, Randy, Horsman, Doug, Marra, Marco A. (2009). Circos: an information aesthetic for comparative genomics. Genome research, **19**(9), 1639-1645.

Land, Miriam, Hauser, Loren, Jun, Se-Ran, Nookaew, Intawat, Leuze, Michael R, Ahn,

Tae-Hyuk, Wassenaar, Trudy. (2015). Insights from 20 years of bacterial genome sequencing. Functional & integrative genomics, **15**(2), 141-161.

Lassalle, Florent, Périan, Séverine, Bataillon, Thomas, Nesme, Xavier, Duret, Laurent, & Daubin, Vincent. (2015). GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet, **11**(2), e1004941.

Laureti, Luisa, Song, Lijiang, Huang, Sheng, Corre, Christophe, Leblond, Pierre, Challis, Gregory L, & Aigle, Bertrand. (2011). Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in Streptomyces ambofaciens. Proceedings of the National Academy of Sciences, **108**(15), 6258-6263.

Lee, Keyong-Ho, Kim, K.W., & Rhee, Ki-Hyeong. (2010). Identification of Streptomyces sp. KH29, which produces an antibiotic substance processing an inhibitory activity against multidrug-resistant Acinetobacter baumannii. Journal of microbiology and biotechnology, **20**(12), 1672-1676.

Lehmann, Jason S., Fouts, Derrick E., Haft, Daniel H., Cannella, Anthony P., Ricaldi, Jessica N., Brinkac, Lauren, Sutton, Granger. (2013). Pathogenomic inference of virulence-associated genes in Leptospira interrogans. PLoS Negl Trop Dis, **7**(10), e2468.

Li, Xiang, Vanner, Stephanie, Wang, Wenliang, Li, Yongchang, Gallardo, Victor Ariel, & Magarvey, Nathan A. (2013). Macplocimine A, a new 18-membered macrolide isolated from the filamentous sulfur bacteria Thioploca sp. The Journal of antibiotics, **66**(7), 443-446.

Loman, Nicholas J., Misra, Raju V., Dallman, Timothy J., Constantinidou, Chrystala, Gharbia, Saheer E., Wain, John, & Pallen, Mark J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology, **30**(5), 434-439.

MacGregor, Barbara J., Biddle, Jennifer F., Harbort, Christopher, Matthysse, Ann G., & Teske, Andreas. (2013). Sulfide oxidation, nitrate respiration, carbon acquisition, and electron transport pathways suggested by the draft genome of a single orange Guaymas Basin Beggiatoa (Cand. Maribeggiatoa) sp. filament. Marine genomics, 11, 53-65.

Machado, Henrique, Sonnenschein, Eva C., Melchiorsen, Jette, & Gram, Lone.

(2015). Genome mining reveals unlocked bioactive potential of marine Gram-negative bacteria. BMC genomics, **16**(1), 1.

Mavromatis, Konstantinos, Yasawong, Montri, Chertkov, Olga, Lapidus, Alla, Lucas, Susan, Nolan, Matt, Pitluck, Sam. (2010). Complete genome sequence of Spirochaeta smaragdinae type strain (SEBR 4228 T). Standards in genomic sciences, **3**(2), 136.

Mayer, Alejandro M.S., Glaser, Keith B., Cuevas, Carmen, Jacobs, Robert S., Kem, William, Little, R., Daniel, Shuster, Dale E. (2010). The odyssey of marine pharmaceuticals: a current pipeline perspective. Trends in pharmacological sciences, **31**(6), 255-265.

McGinnis, Scott, & Madden, Thomas L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic acids research, **32**(suppl 2), W20-W25.

Medema, Marnix H., Trefzer, Axel, Kovalchuk, Andriy, van den Berg, Marco, Müller, Ulrike, Heijne, Wilbert, Nierman, William C. (2010). The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. Genome Biology and Evolution, **2**(10), 212-224.

Medema, Marnix H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic acids research, **39**(suppl 2), W339-W346.

Milano, Teresa, Paiardini, Alessandro, Grgurina, Ingeborg, & Pascarella, Stefano. (2013). Type I pyridoxal 5′-phosphate dependent enzymatic domains embedded within multimodular nonribosomal peptide synthetase and polyketide synthase assembly lines. BMC structural biology, **13**(1), 1.

Miller, Jason R., Koren, Sergey, & Sutton, Granger. (2010). Assembly algorithms for next-generation sequencing data. Genomics, **95**(6), 315-327.

Minowa, Yohsuke, Araki, Michihiro, & Kanehisa, Minoru. (2007). Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. Journal of molecular biology, **368**(5), 1500-1517.

Molinski, Tadeusz F., Dalisay, Doralyn S., Lievens, Sarah L., & Saludes, Jonel P.

(2009). Drug development from marine natural products. Nature reviews Drug discovery, **8**(1), 69-85.

Montaser, Rana, & Luesch, Hendrik. (2011). Marine natural products: a new wave of drugs? Future medicinal chemistry, **3**(12), 1475-1489.

Munro, Murray H.G., Blunt, John W., Dumdei, Eric J., Hickford, Sarah J.H., Lill, Rachel E., Li, Shangxiao, Duckworth, Alan R. (1999). The discovery and development of marine compounds with pharmaceutical potential. Journal of Biotechnology, 70(1), 15-25.

Mußmann, Marc, Hu, Fen Z., Richter, Michael, de Beer, Dirk, Preisler, André, Jørgensen, Bo B, Koopman, Werner JH. (2007). Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. PLoS Biol, 5(9), e230.

Myers, Eugene W., Sutton, Granger G., Delcher, Art L., Dew, Ian M., Fasulo, Dan P., Flanigan, Michael J., Remington, Karin A. (2000). A whole-genome assembly of Drosophila. Science, **287**(5461), 2196-2204.

Nelson, Douglas C., & Jannasch, Holger W. (1983). Chemoautotrophic growth of a marine Beggiatoa in sulfide-gradient cultures. Archives of Microbiology, **136**(4), 262-269.

Nett, Markus, Ikeda, Haruo, & Moore, Bradley S. (2009). Genomic basis for natural product biosynthetic diversity in the actinomycetes. Natural product reports, **26**(11), 1362-1384.

Newman, David J., & Cragg, Gordon M. (2012). Natural products as sources of new drugs over the 30 years from 1981 to 2010. Journal of natural products, **75**(3), 311-335.

Newman, David J., & Cragg, Gordon M. (2016). Natural products as sources of new drugs from 1981 to 2014. Journal of natural products, **79**(3), 629-661.

Park, Ji-Min, & Choi, Sun-Uk. (2015). Identification of a novel unpaired histidine sensor kinase affecting secondary metabolism and morphological differentiation in Streptomyces acidiscabies ATCC 49003. Folia microbiologica, **60**(4), 279-287.

Park, Sung Ryeol, Yoo, Young Ji, Ban, Yeon-Hee, & Yoon, Yeo Joon. (2010). Biosynthesis of rapamycin and its regulation: past achievements and recent

progress. The Journal of antibiotics, **63**(8), 434-441.

Patel, Pramathesh S., HuANG, Stella, Fisher, Susan, Pirnik, Dolores, Aklonis, Carol, Dean, Loretta, Mayerl, Friedrich. (1995). Bacillaene, a novel inhibitor of procaryotic protein synthesis produced by Bacillus subtilis: production, taxonomy, isolation, physico-chemical characterization and biological activity. The Journal of antibiotics, **48**(9), 997-1003.

Pi, Borui, Yu, Dongliang, Dai, Fangwei, Song, Xiaoming, Zhu, Congyi, Li, Hongye, & Yu, Yunsong. (2015). A genomics based discovery of secondary metabolite biosynthetic gene clusters in Aspergillus ustus. PloS one, **10**(2), e0116089.

Piper, Clare, Cotter, Paul D., Ross, Paul R., & Hill, Colin. (2009). Discovery of medically significant lantibiotics. Current drug discovery technologies, **6**(1), 1-18.

Rahman, Hafizur, Austin, Brian, Mitchell, Wilfrid J., Morris, Peter C., Jamieson, Derek J., Adams, David R., Schweizer, Michael. (2010). Novel anti-infective compounds from marine bacteria. Marine drugs, **8**(3), 498-518.

Raza, Khalid, & Ahmad, Sabahuddin. (2016). Principle, analysis, application and challenges of next-generation sequencing: a review. arXiv preprint arXiv:1606.05254.

Reed, Katherine A., Manam, Rama Rao, Mitchell, Scott S., Xu, Jianlin, Teisan, Sy, Chao, Ta-Hsiang, Potts, Barbara CM. (2007). Salinosporamides DJ from the marine actinomycete Salinispora tropica, bromosalinosporamide, and thioester derivatives are potent inhibitors of the 20S proteasome. Journal of natural products, **70**(2), 269-276.

Romano, G., Costantini, M., Sansone, C., Lauritano, C., Ruocco, N., & Ianora, A. (2016). Marine microorganisms as a promising and sustainable source of bioactive molecules. Marine environmental research. In Press, Corrected Proof.

Ronaghi, Mostafa. (2001). Pyrosequencing sheds light on DNA sequencing. Genome research, **11**(1), 3-11.

Rutherford, Kim, Parkhill, Julian, Crook, James, Horsnell, Terry, Rice, Peter, Rajandream, Marie-Adèle, & Barrell, Bart. (2000). Artemis: sequence visualization and annotation. Bioinformatics, **16**(10), 944-945.

Salman, Verena, Amann, Rudolf, Girnth, Anne-Christin, Polerecky, Lubos, Bailey, Jake V., Høgslund, Signe, Schulz-Vogt, Heide N. (2011). A single-cell sequencing approach to the classification of large, vacuolated sulfur bacteria. Systematic and Applied Microbiology, **34**(4), 243-259.

Schatz, Michael C., Phillippy, Adam M., Sommer, Daniel D., Delcher, Arthur L, Puiu, Daniela, Narzisi, Giuseppe, Pop, Mihai. (2013). Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. Briefings in bioinformatics, **14**(2), 213-224.

Schneemann, Imke, Kajahn, Inga, Ohlendorf, Birgit, Zinecker, Heidi, Erhard, Arlette, Nagel, Kerstin, Imhoff, Johannes F. (2010). Mayamycin, a cytotoxic polyketide from a Streptomyces strain isolated from the marine sponge Halichondria panicea. Journal of natural products, **73**(7), 1309-1312.

Schöner, Tim A., Gassel, Sören, Osawa, Ayako, Tobias, Nicholas J., Okuno, Yukari, Sakakibara, Yui, Bode, Helge B. (2016). Aryl Polyenes, a Highly Abundant Class of Bacterial Natural Products, Are Functionally Related to Antioxidative Carotenoids. ChemBioChem, **17**(3), 247-253.

Schopf, J. William. (2006). Fossil evidence of Archaean life. Philosophical Transactions of the Royal Society of London B: Biological Sciences, **361**(1470), 869-885.

Schopf, J. William, Kudryavtsev, Anatoliy B., Walter, Malcolm R., Van Kranendonk, Martin J., Williford, Kenneth H., Kozdon, Reinhard, Flannery, David T. (2015). Sulfur-cycling fossil bacteria from the 1.8-Ga Duck Creek Formation provide promising evidence of evolution's null hypothesis. Proceedings of the National Academy of Sciences, **112**(7), 2087-2092.

Schulz, Heide N., & Jørgensen, Bo Barker. (2001). Big bacteria. Annual Reviews in Microbiology, **55**(1), 105-137.

Shen, Ben. (2015). A New Golden Age of natural products drug discovery. Cell, **163**(6), 1297-1300.

Shivani, Y, Subhash, Y, Tushar, L, Sasikala, Ch, & Ramana, Ch V. (2015). Spirochaeta lutea sp. nov., isolated from marine habitats and emended description of the genus Spirochaeta. Systematic and applied microbiology, **38**(2), 110-114.

Shoaib, Muhammad, Baconnais, Sonia., Mechold, Undine, Le Cam, Eric, Lipinski,

Marc, & Ogryzko, Vasily. (2008). Multiple displacement amplification for complex mixtures of DNA fragments. BMC genomics, **9**(1), 415.

Smith, Stuart, & Tsai, Shiou-Chuan. (2007). The type I fatty acid and polyketide synthases: a tale of two megasynthases. Natural product reports, **24**(5), 1041-1072.

Srinivasan, Ramya, Karaoz, Ulas, Volegova, Marina, MacKichan, Joanna, Kato-Maeda, Midori, Miller, Steve, Lynch, Susan V. (2015). Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. PloS one, **10**(2), e0117617.

Stachelhaus, Torsten, Mootz, Henning D., & Marahiel, Mohamed A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chemistry & biology, **6**(8), 493-505.

Šulčius, Sigitas, Pilkaitytė, Renata, Mazur-Marzec, Hanna, Kasperovičienė, Jūratė, Ezhova, Elena, Błaszczyk, Agata, & Paškauskas, Ričardas. (2015). Increased risk of exposure to microcystins in the scum of the filamentous cyanobacterium Aphanizomenon flos-aquae accumulated on the western shoreline of the Curonian Lagoon. Marine pollution bulletin, **99**(1), 264-270.

Tambadou, Fatoumata, Lanneluc, Isabelle, Sablé, Sophie, Klein, Géraldine L, Doghri, Ibtissem, Sopéna, Valérie, Chevrot, Romain. (2014). Novel nonribosomal peptide synthetase (NRPS) genes sequenced from intertidal mudflat bacteria. FEMS microbiology letters, **357**(2), 123-130.

Toft, C., & Andersson, S. G. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. Nature Reviews Genetics, **11**(7), 465-475.

Udwary, Daniel W., Zeigler, Lisa, Asolkar, Ratnakar N., Singan, Vasanth, Lapidus, Alla, Fenical, William, Moore, Bradley S. (2007). Genome sequencing reveals complex secondary metabolome in the marine actinomycete Salinispora tropica. Proceedings of the National Academy of Sciences, **104**(25), 10376-10381.

Van Duin, David, & Paterson, David L. (2016). Multidrug-Resistant Bacteria in the Community: Trends and Lessons Learned. Infectious disease clinics of North America, **30**(2), 377-390.

Wang, Hao, Fewer, David P, Holm, Liisa, Rouhiainen, Leo, & Sivonen, Kaarina. (2014). Atlas of nonribosomal peptide and polyketide biosynthetic pathways

reveals common occurrence of nonmodular enzymes. Proceedings of the National Academy of Sciences, **111**(25), 9259-9264.

Wartiainen, Ingvild, Hestnes, Anne Grethe, McDonald, Ian R., & Svenning, Mette M. (2006). Methylobacter tundripaludum sp. nov., a methane-oxidizing bacterium from Arctic wetland soil on the Svalbard islands, Norway (78 N). International Journal of Systematic and Evolutionary Microbiology, **56**(1), 109-113.

Waters, Christopher M., & Bassler, Bonnie L. (2005). Quorum sensing: cell-to-cell communication in bacteria. Annu. Rev. Cell Dev. Biol., 21, 319-346.

Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D.H., & Wohlleben, W. (2009). CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. Journal of biotechnology, **140**(1), 13-17.

Weber, Tilmann, Blin, Kai, Duddela, Srikanth, Krug, Daniel, Kim, Hyun Uk, Bruccoleri, Robert, Wohlleben, Wolfgang. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic acids research, **43**(W1), W237-W243.

Weber, Tilmann, & Kim, Hyun U.k. (2016). The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. Synthetic and Systems Biotechnology, **1**(2), 69-79.

Williams, Philip G. (2009). Panning for chemical gold: marine bacteria as a source of new therapeutics. Trends in biotechnology, **27**(1), 45-52.

Wink, Michael. (2003). Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. Phytochemistry, **64**(1), 3-19.

Winter, Jaclyn M., Behnken, Swantje, & Hertweck, Christian. (2011). Genomics-inspired discovery of natural products. Current opinion in chemical biology, **15**(1), 22-31.

Xiong, Zhi-Qiang, Wang, Jian-Feng, Hao, Yu-You, & Wang, Yong. (2013). Recent advances in the discovery and development of marine microbial natural products. Marine drugs, **11**(3), 700-717.

Yamada, Yuuki, Arima, Shiho, Nagamitsu, Tohru, Johmoto, Kohei, Uekusa, Hidehiro, Eguchi, Tadashi, Ikeda, Haruo. (2015). Novel terpenes generated by heterologous expression of bacterial terpene synthase genes in an engineered

Streptomyces host. The Journal of antibiotics, 68(6), 385-394.

Yu, Dayu, Xu, Fuchao, Zeng, Jia, & Zhan, Jixun. (2012). Type III polyketide synthases in natural product biosynthesis. Iubmb Life, **64**(4), 285-295.

Zazopoulos, Emmanuel, Huang, Kexue, Staffa, Alfredo, Liu, Wen, Bachmann, Brian O., Nonaka, Koichi, Farnet, Chris M. (2003). A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nature biotechnology, **21**(2), 187-190.

Zerbino, Daniel R., & Birney, Ewan. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research, **18**(5), 821-829.

Zerbino, Daniel R., McEwen, Gayle K., Margulies, Elliott H., & Birney, Ewan. (2009). Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. PloS one, **4(**12), e8407.

Zhang, Songhe, Gu, Ju, Wang, Chao, Wang, Peifang, Jiao, Shaojun, He, ZhenLi, & Han, Bing. (2015). Characterization of antibiotics and antibiotic resistance genes on an ecological farm system. Journal of Chemistry, **2015**(2015)8, Article ID.

Ziemert, Nadine, Podell, Sheila, Penn, Kevin, Badger, Jonathan H, Allen, Eric, & Jensen, Paul R. (2012). The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One, **7**(3), e34064.

Zotchev, Sergey B., Sekurova, Olga N., & Katz, Leonard. (2012). Genome-based bioprospecting of microbes for new therapeutics. Current opinion in biotechnology, **23**(6), 941-947.

ANEXOS

Anexo 1. Archivo de configuración para ejecutar el algoritmo runCA de Celera.

Spec file_Beggiatoa HS sp.

overlapper = mer # Desarrollado por JCVI para secuencias 454.
merSize = 14 # Default = 22 / Establece el K, largo de cada K-mer. Establece el largo de cada semilla y extensión (Algoritmo *seed & extended*). Es equivalente a “size” de BLAST # Afecta el mer overlapper y el meryl seed finder.

ERROR RATE
utgErrorRate = 0.15 # Un solapamiento se utiliza si está por debajo de la "tasa de error’ o del umbral 'límite de error'.
cnsErrorRate = 0.14 # El consenso espera encontrar alineaciones debajo de este nivel, pero no cumplir estrictamente la misma.
cgwErrorRate = 0.14 # Tasa de error para scaffolder.
frgCorrThreads = 8 # Hilos a utilizar en corrección de errores.
merOverlapperThreads = 8 # número de subprocesos informáticos para su uso. En general, el número de CPUs del anfitrión.
frgCorrThreads = 8 # El número de hilos a utilizar para la corrección de errores de fragmentos.
createACE = 1 # Crea archivos de salida en formato ACE, estos sirven como entrada al *app* ContigScape.

Anexo 2. Achivo de configuración para ejecutar el programa MIRA.

Archivo de configuración para ensamblar genoma de *Beggiatoa* sp. HS
project = Beggiatoa # Nombre del ensamblaje
job = genome,denovo,accurate # Define tipo de data, tipo de reconstrucción y método.
parameters = -GE:not=8 # utilizamos 8 threads en paralel # Definir data; tipo 454 single-end.
readgroup = Unpaired454 # Nombre asignado a la data
data = Unpaired_3.fastq # Define la biblioteca single-end
technology = 454 # Plataforma de secuenciación

Definir data tipo 454 mate-pair
readgroup = PairedReads454 # Nombre asignado a la data
autopairing # Define por sí mismo la dirección de las secuencias
data = Beggiatoa_V1.fastq Beggiatoa_V2.fastq # bibliotecas mate-pair a cargar
technology = 454 # Plataforma de secuenciación de las biblioteca