UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS QUÍMICAS

Departamento de Bioquímica y Biología Molecular

TESIS DOCTORAL

Desarrollo de herramientas moleculares para la producción de policétidos y péptidos no ribosomales

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Carlos del Cerro Sánchez

Directores

José Luis García López
Beatriz Galán Sicilia

Madrid, 2017

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS QUÍMICAS

DEPARTAMENTO DE BIOQUÍMICA Y BIOLOGÍA MOLECULAR

TESIS DOCTORAL

DESARROLLO DE HERRAMIENTAS MOLECULARES PARA LA PRODUCCIÓN DE POLICÉTIDOS Y PÉPTIDOS NO RIBOSOMALES.

CARLOS DEL CERRO SÁNCHEZ

DIRECTORES:

Dr. JOSÉ LUIS GARCÍA LÓPEZ **Dra. BEATRIZ GALÁN SICILIA**

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

CENTRO DE INVESTIGACIONES BIOLÓGICAS

MADRID, 2015

El trabajo descrito en esta Tesis Doctoral se ha llevado a cabo en el Departamento de Biología Medioambiental del Centro de Investigaciones Biológicas del Consejo Superior de Investigaciones Científicas (CIB-CSIC) (Madrid, España). La investigación ha sido financiada por el proyecto Biokétido IPT-2011-0752-900000 (2011-2014) dentro del programa INNPACTO del Ministerio de Economía y Competitividad.

Agradecimientos:

En primer lugar quiero agradecer especialmente a mis directores de Tesis José Luis García y Beatriz Galán por toda la ayuda prestada para realizar este trabajo y por sus valiosas discusiones. También agradezco la estrecha colaboración con el departamento de I+D de Microbiología de Pharmamar al completo y en especial a su director Fernando de la Calle. Agradezco también a todos mis compañeros de los grupos de Biotecnología Mediambiental, Microbiología Medioambiental y Biotecnología de Biopolímeros, la ayuda técnica y emocional que me han prestado para poder llevar a cabo este trabajo. Finalmente me gustaría dar las gracias a mis padres, a mis abuelos y mi hermano, y por supuesto a Isabel por todo el apoyo que siempre me han prestado y sin el que este trabajo no habría sido posible.

Índice:

Abreviaturas

°C	Grado centígrado
A	Adenina o dominio de adenilación
aa	Aminoácidos
AdoMet	S-adenosil metionina
ACP	Proteína trasportadora de acilos (*Acyl Carrier Protein*)
AMP	Monofosfato de adenosina
Ap^R	Resistencia a ampicilina
APS	5' Fosfosulfato de adenosina
AT	Dominio acil-transferasa
ATP	Trifosfato de adenosina
BAC	Cromosoma artificial bacteriano
bp	Pares de bases
BSA	Seroalbúmina bovina
C	Citosina o dominio de condensación
COM	Dominios mediadores de comunicación
Cy	Dominio de condensación de ciclación
cDNA	DNA complementario
Cm^R	Resistencia a cloranfenicol
CoA	Coenzima A
CRISPR	Regiones con repeticiones cortas palindrómicas aglomeradas regularmente interespaciadas
CTAB	Bromuro de hexadeciltrimetilamonio
Da	Dalton
DAD	Detector de diodos (*Diode Array Detector*)
DH	Domino dehidratasa
DMSO	Dimetilsulfóxido
DNA	Ácido desoxirribonucleico
dNTP	Desoxinucleótido trifosfato
DO_{600}	Densidad óptica medida a 600 nm
DOC	Ácido deoxicólico
DTT	Ditiotreitol
EDTA	Ácido etilendiaminotetracético
ER	Dominio enoil-reductasa
FACS	Separación de células activada por fluorescencia
FAS	Sintasa de ácidos grasos
G	Guanina
Gm^R	Resistencia a gentamicina
h	Hora
HMM	Modelos escondidos de Markov
HMW	Alto peso molecular
HPLC	Cromatografía líquida de alta eficiencia
HPLC-MS	HPLC acoplado a espectrometría de masas
IPTG	Isopropil-ß-D-tiogalatopiranósido
ISFET	Transistor de efecto de campo sensible a iones (*Ion Sensitive Field Effect Transitor*)
Km^R	Resistencia a kanamicina
KR	Dominio ceto-reductasa
KS	Dominio cetosintasa
LB	Medio de cultivo Lysogeny broth
min	Minuto
mg	Miligramo
mL	Mililitro
mM	Milimolar
MT	Dominio metil-transferasa
NAD	Nicotinamida-adenina-dinucleótido
$NADH^+H^+$	Nicotinamida-adenina-dinucleótido reducido
NADP	Fosfato de nicotinamida-adenina-dinucleótido
$NADPH^+H^+$	Fosfato de nicotinamida-adenina-dinucleótido reducido

NCBI	National Center for Biotechnology Information
ng	Nanogramos
nm	Nanómetro
NRP	Péptido no ribosomal
NRPS	Sintasa de péptidos no ribosomales
nt	Nucleótido(s)
ORF	Pauta abierta de lectura (*Open Reading Frame*)
OTU	Unidad Taxonómica Operacional (*Operational Taxonomic Unit*)
p/v	Relación peso-volumen
PCP	Proteína transportadora de pétidos (*Peptidil Carrier Protein)*
PCR	Reacción de amplificación en cadena con DNA Polimerasa termorresistente
PE	Extremos pareados
PK	Policétido
PKS	Policétido sintasa
PMSF	Fluoruro de fenilmetilsulfonilo
PP	Sitio de unión a fosfopanteteina
PPi	Pirofosfato
RDP	*Ribosomal Database Project*
RNA	Ácido ribonucleico
mRNA	RNA mensajero
PPTasa	Fosfopanteteinil transferasa
rpm	Revoluciones por minuto
rRNA	RNA ribosomico
RT-PCR	Reacción de retrotranscripción acoplada a PCR
SAM	S-adenosil metionina
SDS	Dodecilsulfato sódico
Sm^R	Resistencia a estreptomicina
T	Timina
TAE	Tampón Tris-Acetato-EDTA
TBE	Tampón Tris-Borato-EDTA
TE	Tampón Tris-EDTA o dominio tioesterasa
Tris	Tri(hidroximetil)aminometano
U	Unidad de actividad enzimática
UTP	Uridina-5'-trifosfato
UV	Ultravioleta
V	Voltio
µg	Microgramos
µF	Microfaradio
µM	Micromolar
Ω	Ohmio

ABREVIATURAS PARA AMINOÁCIDOS:

Ala (A): Alanina	Gly (G): Glicina	Pro (P): Prolina
Arg (R): Arginina	His (H): Histidina	Ser (S): Serina
Asn (N): Asparragina	Ile (I): Isoleucina	Thr (T): Treonina
Asp (D): Aspártico	Leu (L): Leucina	Trp (W): Triptófano
Cys (C): Cisteína	Lys (K): Lisina	Tyr (Y): Tirosina
Gln (Q): Glutamina	Met (M): Metionina	Val (V): Valina
Glu (E): Glutámico	Phe (F): Fenilalanina	

I. Summary

Summary

Introduction:

Marine environments are a huge source of natural compounds. Many of these molecules are synthesized as a defense mechanism by symbiotic microorganisms found in marine organisms such as sponges, tunicates and polychaetes. Among the wide diversity of natural products, the presence of molecules produced by bacterial biosynthetic clusters that encode polyketide synthases (PKS) and/or non-ribosomal peptide synthases (NRPS) is remarkable. The modular organization of these gene clusters, which sometimes cover large genome regions, allows the synthesis of compounds that can present a huge chemical variability.

Depending on whether it concerns a culturable microorganism or a full microbiome, there are some differences in the tools used to identify PKSs and/or NPRSs gene clusters. If the producer microorganism has been isolated and cultured, some genomic tools such as massive sequencing procedures (including subsequent bioinformatics analysis) or genomic libraries cloned in high capacity vectors can be used. However, if the DNA sequence of interest belongs to a non-culturable microorganism, different metagenomics tools can be used. Among these techniques it is possible to find the generation of metagenomic libraries from environmental DNA and processes using massive sequencing that include analysis performed with specialized bioinformatics tools.

Objetives:

At the beginning of this PhD. Thesis we proposed the following objectives:

1. To develop genomic tools to optimize the production of molecules synthesized by genetic clusters enconding PKSs and/or NRPSs in culturable microorganisms.

2. To develop bioinformatics tools in order to identify sequences belonging to genetic clusters encoding PKSs or NRPSs in microbiomes.

Results:

Using genetic tools for the identification, characterization and heterologous expression of the didemnins biosynthetic gene cluster contained in *Tistrella mobilis* MES-10-09-028.

To address the first objective of this Thesis, the culturable symbiont *T. mobilis* MES-10-09-028 was used as a proof of concept. First, massive sequencing procedures were used to obtain the *T. mobilis* MES-10-09-028 complete genome sequence. Further analysis allowed the didemnin biosynthetic gene cluster (*ddn* cluster) identification. The construction of the KR3 and DidA *T. mobilis* mutants, which are affected in the initial stages of the biosynthetic pathway of didemnins, showed that *ddn* cluster is responsible for the synthesis of didemnins, since both mutations block the didemnin production in MES-10-09-028 strain. Alternative intermediates from the synthesis were not detected in these mutants.

To achieve the heterologous production of didemnins, a BACs library using genomic DNA from MES-10-09-028 strain was constructed in *Escherichia coli* and a clone that contains all PKSs and NRPSs genes from *ddn* cluster was identified. In addition to this, the *ddn* cluster was efficiently expressed in the heterologous host.

Using metagenomic tools for the identification of secondary metabolites biosynthetic clusters in non-culturable strains.

To address the second objective of this Thesis, three different marine sponge samples were used, *i.e.*, *Polymastia littoralis,* PMLT01 and *Lithoplocamia lithistoides*, in which antitumoral compounds were detected. Once microbiome DNA was isolated, environmental sequences were obtained using massive sequencing procedures. Taxonomic and functional information was obtained using bioinformatics tools with the metagenomic sequences. It was demonstrated that some symbionts are majority and therefore have to play an important metabolic role in sponges' development. *In silico* tools were applied on metagenomic sequences for the detection of regions that belong to gene clusters enconding PKSs and/or NRPSs. As a result, biosynthetic clusters that can be related with antitumoral compounds production were detected.

Conclusions:

1. *T. mobilis* MES-10-09-028 strain can be modified by genetic engineering tools, making it an excellent candidate for its use as a bacterial chassis for the production of PKSs and/or NRPSs related molecules.
2. It has been shown that *ddn* cluster is responsible for the synthesis of didemnins and it has been heterologously expressed in *E. coli*.
3. Analyzed marine sponges contain a few very major symbionts, which allows obtaining their complete genomic sequences using relatively low-depth massive sequencing procedures.
4. Analysis of gene sequences encoding PKS and/or NRPS clusters present in sponge metagenomes shows that antitumoral molecules detected are not always produced by the most abundant symbionts.

II. Resumen

Resumen:

Introducción:

Los ambientes marinos son una gran fuente de compuestos naturales. Muchas de estas moléculas son sintetizadas como mecanismo de defensa por microorganismos simbiontes de organismos marinos tales como esponjas, tunicados o poliquetos. Entre la gran variedad de productos naturales, cabe destacar la presencia de moléculas producidas por clústeres biosintéticos bacterianos que codifican policétido sintasas (PKS) y/o sintasas de péptidos no ribosomales (NRPS). La organización modular de este tipo de clústeres génicos, que en ocasiones pueden abarcar grandes regiones del genoma, permite la síntesis de compuestos con una gran variabilidad química.

Existen diferencias en las herramientas utilizadas para identificar clústeres génicos que codifican PKSs y NRPSs, dependiendo de que se trate de un microorganismo cultivable o de un microbioma completo que incluye el microorganismo productor. En el caso de que se haya conseguido aislar y cultivar el microorganismo productor se pueden utilizar herramientas genómicas, como procedimientos de secuenciación masiva y su posterior análisis bioinformático, o bien la construcción de librerías genómicas en vectores de clonación de gran capacidad. Sin embargo, si se pretende identificar la secuencia de DNA de un clúster productor que pertenece a un microorganismo no cultivable, se pueden utilizar diferentes herramientas metagenómicas. Entre estas técnicas se encuentran la generación de librerías metagenómicas a partir de DNA metagenómico obtenido del microbioma y procesos de secuenciación masiva para su posterior análisis mediante herramientas bioinformáticas especializadas.

Objetivos:

Al comienzo de esta Tesis Doctoral se propusieron los siguientes objetivos:

1. Desarrollar herramientas genómicas para optimizar la producción de moléculas sintetizadas por clústeres génicos que codifican PKSs y/o NRPSs en microorganismos cultivables.

2. Desarrollar herramientas bioinfórmaticas para identificar secuencias pertenecientes a clústeres génicos que codifican PKSs y/o NRPSs en microbiomas.

Resultados:

Utilización de herramientas genómicas para la identificación, caracterización y expresión heteróloga del clúster de síntesis de didemninas contenido en *Tistrella mobilis* MES-10-09-028:

Para abordar el primer objetivo de la Tesis se utilizó como prueba de concepto la bacteria simbionte cultivable *Tistrella mobilis* MES-10-09-028 productora de didemninas. En primer lugar se llevó a cabo la secuenciación del genoma de la bacteria mediante técnicas de secuenciación masiva. Su análisis posterior permitió la identificación de la secuencia del clúster génico responsable de la síntesis de didemninas (clúster *ddn*). Mediante la construcción de los mutantes KR3 y DidA de *T. mobilis*, que están afectados en las etapas iniciales de la ruta biosintética de las didemninas, se demostró que el clúster

ddn es el responsable de la síntesis de didemninas, ya que ambas mutaciones anulan la capacidad de producir estas moléculas en la cepa MES-10-09-028. En estos mutantes tampoco se detectó la producción de otros intermediarios alternativos.

Para abordar la producción heteróloga de didemninas se construyó una genoteca de BACs en *E. coli* con el DNA genómico de la cepa MES-10-09-028 y se identificó un clon que contenía el clúster *ddn* completo de genes que codifican PKSs y NRPSs para la síntesis de didemnina. Además, se comprobó que el clúster *ddn* se expresa de manera eficiente en el hospedador heterólogo.

Utilización de herramientas metagenómicas para la identificación de secuencias de clústeres de síntesis de metabolitos secundarios en cepas no cultivables

Para abordar el segundo objetivo de la Tesis se utilizaron muestras de 3 esponjas marinas, *i.e., Polymastia littoralis*, PMLT01 y *Lithoplocamia lithistoides*, en las cuales se había detectado la presencia de compuestos antitumorales. Una vez aislado el DNA de la microbiota de estas esponjas, se obtuvo su secuencia metagenomica mediante secuenciación masiva. Aplicando herramientas bioinformáticas en las secuencias del metagenoma se obtuvo información sobre los microbiomas a nivel funcional y taxonómico. Se ha demostrado en los casos analizados que algunos simbiontes son mayoritarios y por lo tanto han de desempeñar un papel metabólico importante en el desarrollo de las esponjas. Sobre las secuencias metagenómicas se aplicaron herramientas *in silico* para la detección de regiones pertenecientes a clústeres génicos que codifican PKSs y/o NRPSs. De este modo se han localizado clústeres de biosíntesis, que podrían ser el responsable de la producción de compuestos antitumorales.

Conclusiones:

1. La cepa *T. mobilis* MES-10-09-028 puede ser modificada mediante herramientas de ingeniería genética lo que la convierte en un candidato excelente para ser utilizada como chasis para la producción de moléculas sintetizadas por PKSs y/o NRPSs.

2. Se ha demostrado que el clúster *ddn* es el responsable de la síntesis de didemninas y se ha expresado heterólogamente en *E. coli*.

3. Algunos microbiomas de las esponjas marinas analizadas contienen unos pocos simbiontes muy mayoritarios, lo que permite obtener sus genomas completos utilizando profundidades de secuenciación masiva relativamente bajas.

4. El análisis de las secuencias de clústeres génicos que codifican PKSs y/o NRPSs presentes en los metagenomas de esponjas demuestra que las moléculas antitumorales detectadas en las mismas no siempre son producidas por un simbionte mayoritario.

III. Introducción

1. Los ambientes marinos como fuente de nuevos compuestos bioactivos

Los océanos, el lugar de origen de la vida en la Tierra, constituyen aproximadamente tres cuartas partes de la superficie del planeta, y sin embargo son uno de los recursos biológicos más infrautilizados (Revisado en Kim y Jayachandran, 2015). Estos ecosistemas marinos contienen una gran diversidad taxonómica y biológica de macroorganismos y microorganismos. En concreto, más de 40000 especies distintas están presentes en los ambientes marinos, las cuales se clasifican como microorganismos, componentes de las praderas marinas, algas, corales y animales (Nybakken, 2005). Los ambientes marinos son muy diversos pudiéndose dar condiciones extremas y/o cambiantes de presión, salinidad, temperatura o calidad del agua. Los organismos marinos poseen la capacidad de producir metabolitos no convencionales denominados productos naturales, que se sintetizan debido a sus peculiares características fisiológicas que han adquirido a lo largo del tiempo, y que les permiten sobrevivir en estos ambientes (Irigoien *et al*., 2004; Steele, 1985). Estas circunstancias especiales, han favorecido la competencia entre las especies que coexisten en estos ambientes, y de este modo, en una gran multitud de casos, se ha desarrollado una enorme diversidad de compuestos naturales utilizados en principio como mecanismos que aumentan su supervivencia.

Los productos naturales y sus derivados representan más del 50% de todas las drogas de uso clínico a nivel mundial (Revisado en Kim y Senthilkumar, 2015). Estos compuestos, que se originan como consecuencia de la adaptación ambiental generalmente suelen formar parte del mal denominado "metabolismo secundario" e involucran compuestos pertenecientes a muy distintas clases químicas como son alcaloides, terpenoides, esteroides, azúcares, policétidos, péptidos, etc. En el caso de los ambientes marinos, es común encontrar derivados con halógenos como cloro o bromo, posiblemente, debido su gran disponibilidad en el agua del mar (Firn y Jones, 2003).

2. Simbiosis, relaciones entre microorganismos y organismos marinos y la producción de metabolitos secundarios

El término simbiosis (del griego: *syn*, "juntos" y *biosis*, "vivir") hace referencia a la interacción estrecha y normalmente de larga duración entre dos o más especies biológicas diferentes. Dentro de las relaciones posibles, la endosimbiosis es aquella que se da cuando un organismo vive en el interior del cuerpo o de las células de otro organismo. Este tipo de relación entre especies está presente en el caso de multitud de organismos marinos como por ejemplo los corales, tunicados, briozoos, poliquetos o esponjas marinas, las cuales son conocidas por poseer en muchos casos comunidades microbianas simbiontes de alta complejidad. En esta Tesis nos hemos centrado en el estudio de los microorganismos simbiontes de esponjas marinas, las cuales son una de las fuentes más ricas e interesantes de compuestos bioactivos, habiéndose aislado más fármacos que de ningún otro organismo marino (Blunt *et al*., 2011).

2.1. | Las esponjas marinas

Las esponjas marinas pertenecen al phylum *Porifera* y son los animales metazoos existentes más antiguos, ya que se han encontrado restos fósiles que se han datado en casi 630 millones de años de antigüedad (Maloof *et al*., 2010). Las esponjas se encuentran globalmente distribuidas y juegan un papel importante en el ciclo marino de nutrientes, ya que son unos de los miembros más importantes de las comunidades bénticas. En concreto, los microorganismos endosimbiontes nitrificantes son los responsables de fijar el nitrógeno inorgánico disuelto, generando a su vez altas concentraciones de nitratos en los fondos oceánicos. Estos microorganismos también juegan un papel relevante como fuentes de carbono orgánico particulado, carbono orgánico disuelto y nitrógeno orgánico disuelto (Diaz y Ward, 1997).

La base de datos mundial de esponjas marinas (The World Porifera Database) (Van Soest *et al*., 2015) contiene actualmente más de 8500 especies distintas, que se distribuyen a su vez en cuatro clases diferentes: *Calcarea, Hexactinellida, Demospongiae* y *Homoscleromorpha*, siendo *Demospongiae* la más numerosa, ya que contiene más de un 80% de las especies.

2.1.1. | Anatomía y fisiología de las esponjas marinas

Las esponjas marinas muestran una gran diversidad de morfologías diferentes, como por ejemplo incrustadas, ramificadas o de tipo barril, sin embargo, todas las esponjas filtradoras comparten características fisiológicas básicas.

Atendiendo a las características anatómicas de las esponjas marinas se pueden distinguir:

- **Pinacodermo:** también llamado ectosoma, consiste en la epidermis de la esponja y está compuesta por pinacocitos. El pinacodermo está atravesado por múltiples poros dermales, los cuales están tapizados por un tipo celular llamado porocito. En algunas especies se puede observar una capa de colágeno llamada cutícula, que sustituye al pinacodermo.
- **Coanodermo:** se trata del conjunto de células flageladas que tapizan la superficie interna de la esponja. La cavidad central principal se denomina espongocele o atrio. Las células flageladas se denominan coanocitos, los cuales organizados en coanosomas, consiguen crear, gracias al movimiento de sus flagelos, una corriente de agua que fluye desde fuera de la esponja a través de los poros de pinacodermo (ostia), por el sistema acuífero de la esponja, para acabar saliendo a través del ósculo.
- **Mesohilo o mesoglea:** entre las dos capas anteriormente descritas se encuentra un espacio llamado mesohilo. En esta zona se pueden encontrar, fibras de soporte, espículas del esqueleto y varios tipos de células ameboides relacionados con la digestión, la secreción del esqueleto, la producción de gametos y el transporte de nutrientes y desechos.
- **Esqueleto:** los sistemas esqueléticos de las esponjas están compuestos de espículas que pueden ser de tipo calcáreas, compuestas de carbonato cálcico

($CaCO_3$); silíceas, compuestas de dióxido de silicio (SiO_2); o espongina, una proteína colagenosa. Estos tipos de sistemas esqueléticos se encuentran distribuidos entre el conjunto de las esponjas y tradicionalmente han sido utilizados en los procesos de clasificación taxonómica.

Dentro de los tipos celulares más importantes que se pueden encontrar en las esponjas se distinguen:

- **Pinacocitos:** forman parte de la capa externa de la esponja y poseen funciones de protección y fagocitosis.
- **Porocitos:** poseen una forma cilíndrica y forman el sistema de acuífero de poros a través del cuerpo de la esponja. Solo se encuentran en esponjas calcáreas.
- **Coanocitos:** generan con sus flagelos corrientes de agua a través de la esponja. Su largo flagelo central está rodeado de microvellosidades que se conectan entre sí formando un retículo. El agua que fluye cargada de partículas atraviesa las microvellosidades, donde queda retenido el alimento que posteriormente será fagocitado.
- **Colenocitos y Lofocitos:** son células secretoras de colágeno (fibras de espongina) que se encuentran en el mesohilo.
- **Espongiocitos:** se encargan de la producción de espículas.
- **Miocitos:** son células contráctiles situadas alrededor de los canales principales y el ósculo.
- **Arqueocitos:** pueden transformarse en cualquier otro tipo celular. Tienen un papel importante en la digestión ya que viajan por el mesohilo fagocitando alimento. También conforman el sistema de transporte de las esponjas. Dada su totipotencia, son claves para la reproducción de la esponja, pudiéndose diferenciar en los llamados oocitos o gémulas, utilizados para la reproducción sexual o asexual respectivamente.

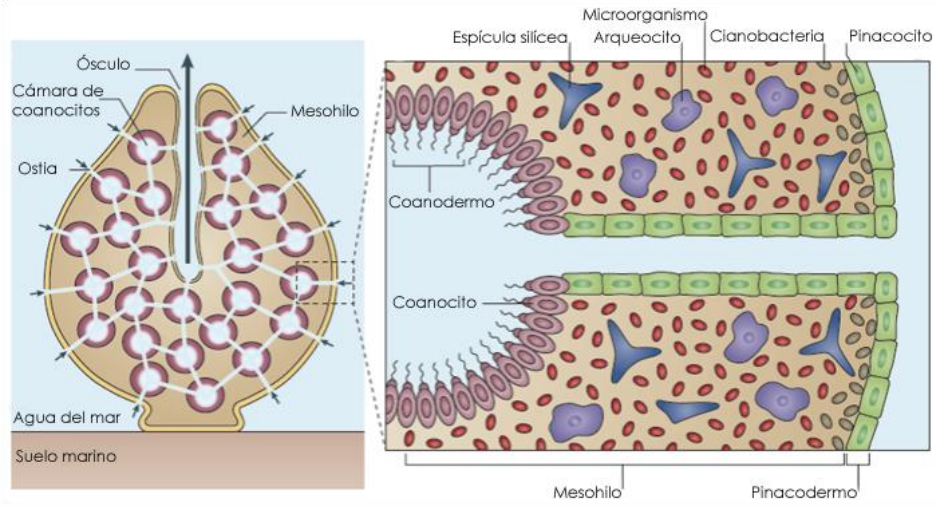

Figura 1 | Esquema del cuerpo de las esponjas marinas. Vista general esquemática de una demosponja y ampliación de la estructura celular del tejido. Modificado de Hentschel *et al*., 2012.

El sistema acuífero de las esponjas hace las veces de sistema circulatorio, digestivo y excretor. La mayoría de los ejemplares adultos son animales sésiles que se alimentan de filtrar bacterias, eucariotas microscópicos y materia particulada del ambiente marino.

Estos nutrientes son bombeados hasta el interior de sus cuerpos donde se distribuyen por su sistema de canales. El oxígeno necesario se obtiene mediante difusión y los nutrientes se asimilan y digieren mediante fagocitosis, principalmente en la zona del mesohilo. Los residuos metabólicos son eliminados generalmente por la corriente constante de agua generada a través del cuerpo (Fig. 1).

2.2. | Microorganismos asociados a las esponjas marinas

Aunque las esponjas llevan varios cientos de millones de años en el planeta, las bacterias ya existían en el mar cuando aparecieron las esponjas, lo que significa que las esponjas evolucionaron en un ambiente plagado de potenciales parásitos y patógenos. La estrategia del cuerpo simple se ha mantenido durante toda su evolución y ha permanecido esencialmente sin cambios, aunque su falta de complejidad morfológica contradice la existencia de un complejo arsenal de defensas ante microorganismos invasores. Además las esponjas son capaces de alojar una comunidad de microorganismos densa, enorme y diversa que pueden llegar a constituir hasta el 35-40% de toda la biomasa llegándose a encontrar hasta en densidades que exceden 10^9 células microbianas por centímetro cúbico de tejido de la esponja, superándose en 3 o 4 órdenes de magnitud su número en el agua. La mayoría de los microorganismos asociados a las esponjas habitan el tejido del mesohilo, que corresponde a la zona mayoritaria del cuerpo de la esponja. Sin embargo, también existen evidencias de que algunos simbiontes pueden encontrarse de forma intracelular (Vacelet y Donadey, 1977). El mesohilo es una zona muy interesante, ya que por una parte, es el lugar donde son digeridos algunos microorganismos, y por otro lado, en muchas esponjas, es el lugar donde residen las comunidades de microorganismos simbiontes, los cuales se escapan a la digestión por parte del hospedador. Algunos estudios han demostrado que simbiontes ingeridos por la esponja pasan a través de todo el sistema intactos mientras que bacterias no simbiontes son consumidas en el mismo proceso (Lee *et al*., 2001; Wehrl *et al*., 2007). Esto implica que la esponja posee mecanismos para reconocer e ignorar aquellos microorganismos simbiontes específicos, quizás en algunos casos debido a que los microorganismos son capaces de protegerse ante su detección (Lee *et al*., 2001). Además se conoce que las esponjas puede secretar compuestos antimicrobianos (Blunt *et al*., 2011) y poseen un sistema inmune innato desarrollado (Wiens *et al*., 2007; Gauthier *et al*., 2010). Los microorganismos que establecen una relación de simbiosis con las esponjas disfrutan de beneficios entre los que se incluyen el acceso a una fuente estable de nutrientes, resultado del sistema de alimentación mediante filtración de la esponja, así como acceso a una fuente de nitrógeno, el cual se excreta en forma de amoniaco como producto final del metabolismo del hospedador.

La estabilidad y la transmisión de las comunidades microbianas simbiontes en las esponjas marinas a lo largo de sucesivas generaciones se postula como fruto de la combinación de dos procesos diferentes (Fig. 2) (Taylor *et al*., 2007; Schmitt *et al*., 2008 Webster y Taylor, 2012). El primero de ellos consistiría en una transmisión vertical o maternal, ya que se ha demostrado la transmisión de microorganismos en huevos y larvas (On *et al*., 2009). El hecho de que los individuos adultos y en fase reproductiva posean composiciones solapantes de las comunidades microbianas, concuerda con que gran parte o la mayoría de los microorganismos asociados se transmitan verticalmente (Webster *et*

al., 2010). Esto hace posible que la asociación entre microorganismos y esponja pueda mantenerse durante tiempos largos permitiendo que se den procesos de co-evolución. Sin embargo, la aparente falta de co-especiación observada entre simbionte y hospedador (Revisado en Taylor *et al*., 2007), además de las observaciones relacionadas con bacterias asociadas a esponjas en el agua del mar (Webster *et al*., 2010), sugieren la existencia de un segundo proceso que consistiría en el reclutamiento efectivo de microorganismos del agua que rodea a la esponja a lo largo de su ciclo vital.

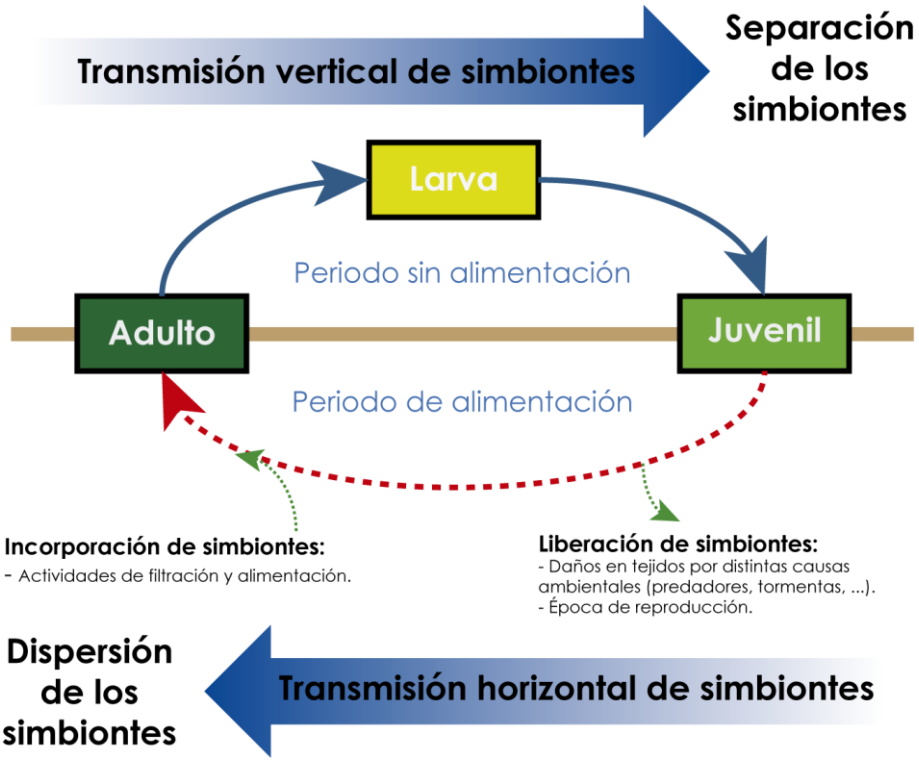

Figura 2 | Modelo propuesto de la evolución de las comunidades microbianas en esponjas marinas. Las fases sin alimentación y sin reproducción se representan mediante una línea continua mientras que estadios vitales de alimentación se señalan con una línea discontinua. Modificado de Schmitt *et al*. (2008).

2.2.1. | Composición y diversidad de la comunidad microbiana simbionte

Como se ha visto hasta ahora, las relaciones entre los microorganismos simbiontes y la esponja hospedadora son de una gran complejidad. Un factor que se debe tener en cuenta a la hora de intentar entender este tipo de sistemas es la existencia de una gran diversidad microbiana asociada en el interior del hospedador. En concreto, las esponjas marinas son hospedadoras de diversos tipos de microorganismos: eucariotas (Baker *et al*., 2009; Cerrano *et al*., 2004), arqueas (Margot *et al*., 2002; Webster *et al*., 2004) y bacterias (Taylor *et al*., 2007). También se ha detectado la presencia de virus y bacteriófagos en los tejidos de la esponja (Lohr *et al*., 2005; Harrington *et al*., 2012).

2.2.1.1. | Bacterias asociadas a esponjas marinas

Las poblaciones bacterianas asociadas a las esponjas marinas han sido estudiadas desde dos puntos de vista distintos. Para obtener información del conjunto de las bacterias asociadas, por un lado, se ha estudiado la diversidad de las bacterias aisladas cultivables y por otro, la de las bacterias que forman el microbioma total utilizando herramientas como

la generación de genotecas de amplicones del 16S rRNA o la secuenciación masiva del metagenoma (ver apartado 4.2.1.2. de la Introducción).

Existen numerosos estudios que analizan con detalle las características de la presencia de bacterias en distintas tipos de esponjas marinas en condiciones diferentes (revisado en Taylor *et al*.2007). En trabajos recientes (Hentschel *et al*., 2012), se menciona la presencia de hasta 28 phyla bacterianos en esponjas marinas utilizando técnicas convencionales como el cultivo de aislados y la construcción de genotecas de 16S rRNA, de los cuales 18 se han descrito formalmente y 10 son phyla candidatos. Además, se han detectado otros phyla adicionales utilizando la secuencia de 16S rRNA obtenida mediante técnicas de secuenciación masiva. Con todos los datos obtenidos hasta la fecha utilizando las distintas aproximaciones experimentales en general se puede concluir que los phyla más abundantes asociados a las esponjas incluyen las Proteobacteria (especialmente las clases Alpha-, Gamma- y Delta-Proteobacteria), Chloroflexi, Actinobacteria, Cyanobacteria, Bacteroidetes, Firmicutes, Planctomycetes, Acidobacteria, Nitrospirae y el phylum candidato Poribacteria (Webster y Taylor, 2012) (Fig. 3).

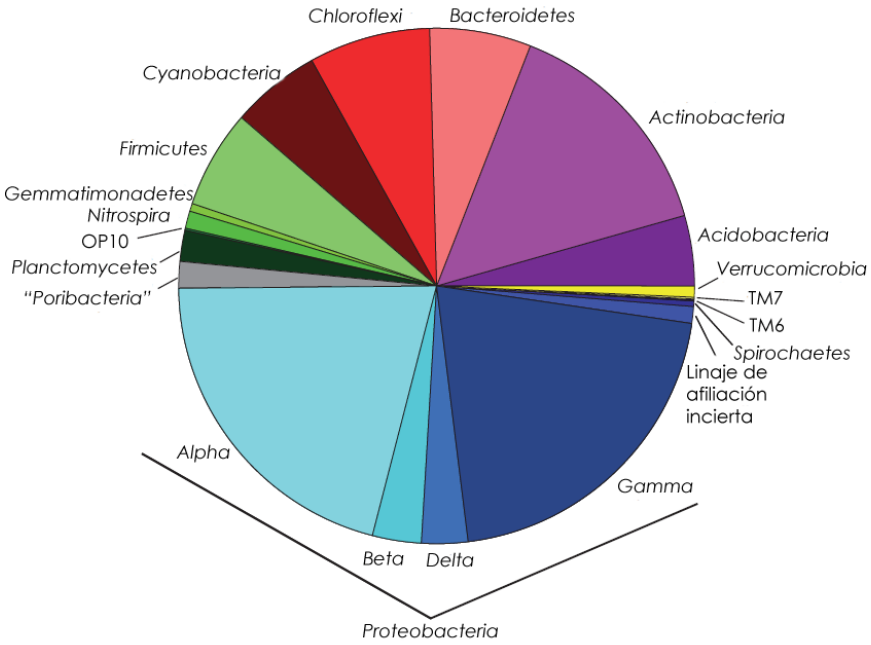

Figura 3 | Distribución filogenética de bacterias asociadas a esponjas. La distribución fue calculada por Webster *et al*. (2010) utilizando 11 284 secuencias de GenBank (versión septiembre de 2010) utilizando las cadenas de búsqueda (sponge* or porifer*) y (16S or ssu* or rRNA*) y no (18S* or lsu* or large subunit or mitochondri* or 23S* or 5S* or 5.8S* or 28S* or crab* or alga* or mussel* or bivalv* or crustacea*). Las secuencias se asignaron taxonómicamente utilizando la base de datos SILVA (Quast *et al*., 2013) (versión 100). Aquellos phyla con menos de 10 secuencias fueron excluidos de la representación. Tampoco se incluyeron secuencias pertenecientes a Crenarchaeota y Euryarchaeota. Modificado de Webster y Taylor (2012).

Al comparar el nivel de diversidad de las bacterias asociadas a esponjas marinas con otros ejemplos de sistemas de simbiosis con un animal hospedador, se puede decir que las comunidades en las esponjas son al menos tan diversas como otros ejemplos parecidos registrados como el caso de los corales (Cárdenas *et al*., 2012). En el caso muy estudiado del microbioma humano se puede decir que existen niveles similares de diversidad a nivel de especie pero que, sin embargo, a nivel de phylum, las esponjas marinas muestran más variabilidad. La diversidad bacteriana encontrada es a su vez

dependiente de la esponja siendo una de las comunidades más diversas a nivel de género, familia, orden y clase, la de la esponja *Rhopaloides odorabile* (Webster *et al*., 2010), para la cual se han descrito aproximadamente 3000 unidades taxonómicas operacionales (OTUs), con un 95% de identidad, utilizando métodos de secuenciación masiva.

2.2.1.2. | Arqueas asociadas a esponjas marinas

Las arqueas pueden encontrarse habitualmente en las comunidades microbianas de la esponja en una proporción relevante. El primer caso detectado de una arquea simbionte es el de *Cenarchaeum symbiosum* en los tejidos de la esponja *Axinella mexicana* por Preston *et al*., 1996. A partir de ahí se han ido sucediendo ejemplos de la presencia de arqueas en esponjas marinas, y hasta han llegado a observarse fenómenos de transferencia vertical de estos microorganismos (Steger *et al*., 2008). Este tipo de fenómenos sugiere que al igual que en el caso de las bacterias, existe una relación muy estrecha entre algunas arqueas y las esponjas marinas. Un ejemplo en el que se utilizaron herramientas de secuenciación masiva llegó a estimar una proporción de entre el 4% y el 28% de arqueas en muestras de distintas esponjas del Mar Rojo (Lee *et al*., 2011).

2.2.1.3. | Otros microorganismos asociados

A parte de las bacterias y las arqueas, existen otros microrganismos que se pueden encontrar asociados a esponjas como por ejemplo virus, hongos y otros eucariotas como diatomeas y dinoflagelados. Sin embargo, las relaciones que involucran este conjunto de microorganismos han sido menos estudiadas y requieren un esfuerzo de investigación adicional.

En el caso de los virus asociados a las esponjas existen muy pocos estudios lo que sorprende considerando la gran cantidad de agua filtrada por las esponjas y la alta densidad de bacterias disponibles para los ataques de fagos. Estudios metagenómicos del microbioma revelan una gran abundancia de regiones con repeticiones cortas palindrómicas aglomeradas regularmente interespaciadas (CRISPR) entre las secuencias (Thomas *et al*., 2010), las cuales se asocian con la presencia de fagos en el hospedador. Aunque los virus son conocidos como patógenos de muchos organismos marinos, su papel como simbiontes en esponjas no ha recibido demasiada atención, por lo que su estudio requiere de un mayor esfuerzo para tratar de comprender la importancia fundamental de estas relaciones en la ecología y la evolución de la simbiosis en la esponja.

Por otro lado, aunque se trata de un campo poco explorado, se conocen casos de simbiosis entre hongos y esponjas marinas. Muchos de los hongos aislados de esponjas están muy relacionados con especies terrestres, encontrándose especies pertenecientes a los géneros *Penicillium* y *Aspergillus*. De hecho existen evidencias de algunos compuestos de interés producidos por hongos asociados a las esponjas, algunos con actividad anticancerígena, lo que ha motivado que se pongan en marcha estudios taxonómicos que involucran la subunidad 18s del rRNA para comprender el papel de los hongos en las esponjas. En estos estudios se han identificado 32 géneros de los phyla Acomycota (conteniendo 22 órdenes distintos), Basidiomycota y Zygomycota (revisado en Dobson *et al*., 2015). Sin embargo, pese al interés reciente, las relaciones entre esponjas y hongos son

las menos estudiadas quedando aún mucho camino por recorrer a la hora de entender este tipo de sistemas.

Además de los hongos, existen otros microorganismos eucariotas que pueden mantener relaciones estrechas de parasitismo o simbiosis con las esponjas hospedadoras. Ejemplos de esto son organismos como los dinoflagelados (Garson *et al*., 1998; Scalera-Liaci *et al*., 1999) y las diatomeas (Gaino *et al*., 1994) los cuales han sido observados en presencia de esponjas. Microalgas endosimbióticas, principalmente las pertenecientes al género *Zoochlorella* también han sido detectadas (Frost *et al*., 1997).

2.2.2. | Especificidad en los microorganismos asociados a esponjas

De todo el conjunto de microorganismos asociados, existen algunos que tienden a encontrarse en las esponjas de forma más común que en otros ambientes. El grado en el cual estos microorganismos parecen ser específicos para los hospedadores varía mucho entre los phyla. Los mayores niveles de abundancia podrían indicar si este tipo de microorganismos pueden representar la fracción de verdaderos simbiontes frente a aquellos que forman parte de la esponja en el momento de la recolección debido a su forma de alimentarse. Tal y como se observa en la Figura 4 existen algunos phyla en los que hay unos pocos individuos específicos como por ejemplo Bacteroidetes y Firmicutes y otros que presentan más especificidad como por ejemplo Poribacteria y Chloroflexi (Simister *et al*., 2012; Hentschel *et al*., 2012). De hecho, el phylum candidato Poribacteria resulta muy interesante en este sentido ya que sus miembros se han encontrado casi exclusivamente en presencia de esponjas marinas (Siegl *et al*., 2011).

Las arqueas, mayoritariamente del phylum Thaumarchaeota, suelen mostrar más especificidad en sus asociaciones con esponjas (Pape *et al*., 2006; Radax *et al*., 2012). Los hongos poseen un nivel de especificidad menor y otros eucariotas como dinoflagelados y diatomeas están presentes pero no parecen mostrar especificidad.

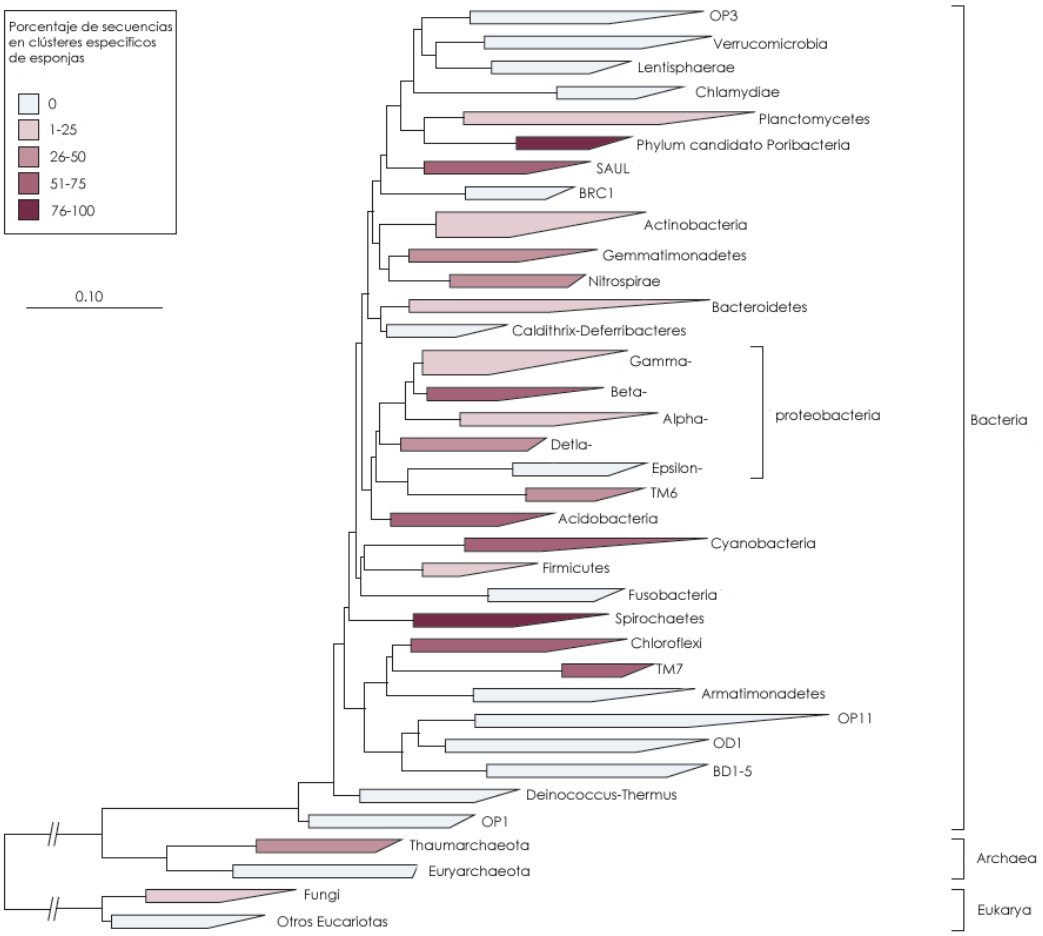

Figura 4 | Diversidad y especificidad de los microorganismos asociados a esponjas marinas. Análisis filogenético basado en clústeres de secuencias de genes 16S rRNA en el cual se señalan los representantes de todos los phyla microbianos que se han descrito asociados a esponjas según el meta-análisis correspondiente en el estudio de Simister *et al*. (2012). Modificado de Hentschel *et al*. (2012).

Esta diversidad en los niveles de especificidad sugiere la existencia de relaciones especializadas que en ocasiones podrían derivar en mecanismos de adaptación como la síntesis de nuevos productos naturales.

2.2.3. | Productos naturales aislados de esponja

Una de las características más interesantes de las esponjas marinas es la capacidad de producir un rango muy diverso de productos naturales. Como ya se ha comentado, las esponjas se encuentran entre las fuentes más ricas de metabolitos secundarios biológicamente activos y de ellas se han aislado más fármacos que de ningún otro organismo marino (Blunt *et al*., 2011). La causa de esta enorme fuente de compuestos de interés reside en el modo de vida cooperativo que llevan a cabo las esponjas y los microorganismos simbiontes. En concreto, se postula que muchas de estas sustancias podrían actuar como agentes protectores frente a predadores y otros microorganismos competidores, y este hecho puede haber contribuido al éxito evolutivo y ecológico de las esponjas marinas.

Además de todos los indicios ecológicos, la observación de la estructura de los compuestos aislados de esponjas sugiere que muchos de éstos puedan ser producidos por los

microorganismos simbiontes, y muy probablemente por bacterias. En concreto muchas de estas moléculas poseen estructuras de compuestos producidos por rutas de biosíntesis microbianas, como aquellos que involucran policétidos complejos o péptidos no ribosomales. El posible origen bacteriano de este tipo de compuestos abre un abanico de posibilidades para utilizar herramientas microbiológicas y genéticas a la hora de crear sistemas biotecnológicos de producción sostenibles.

3. El metabolismo secundario: policétido sintasas y sintasas de péptidos no ribosomales

3.1. | El metabolismo secundario microbiano

Los metabolitos secundarios son compuestos orgánicos con estructuras químicas muy variadas que pueden llegar a ser muy complejas. Aunque su función no está directamente relacionada con el crecimiento, su déficit se ha asociado a efectos negativos en la supervivencia o en la capacidad reproductiva. Este tipo de moléculas son producidas por una gran variedad de organismos, tales como plantas, hongos, bacterias, etc. Históricamente, las investigaciones se centraron en aquellos metabolitos secundarios producidos por plantas, debido a la abundancia del material y a la facilidad con la que este podía ser recolectado. Sin embargo, ya en el siglo XX, la producción de metabolitos secundarios en microorganismos comenzó a ser reconocida y estudiada.

Debido a que los metabolitos secundarios son compuestos que en principio carecen de funciones metabólicas esenciales, su síntesis puede ser explicada según distintas teorías.. Una de ellas sostiene que la mayoría de los metabolitos secundarios no jugarían papel alguno en la adaptación y/o la competencia del organismo productor y sólo algunos tendrían propiedades relevantes para el microorganismo. Sin embargo, otros autores apoyan la teoría de que todo metabolito secundario que se sintetiza debería poseer algún tipo función biológica que incremente (ahora o en algún estadio previo evolutivo) la adaptación del organismo.

Estos compuestos generados, que pueden actuar de modo muy diverso en el microrganismo productor, suponen un gran arsenal de moléculas con actividades de interés biotecnológico, entre las que se incluyen, antibióticos, pigmentos, toxinas, efectores de competición ecológica y simbiosis, feromonas, inhibidores enzimáticos, agentes inmunomoduladores, antagonistas y agonistas de receptores, pesticidas, promotores de crecimiento de animales y plantas, y agentes antitumorales, entre otros muchos.

3.2. | Clasificación de metabolitos secundarios

Generalmente, los metabolitos secundarios se pueden clasificar atendiendo al modo como son sintetizados. En muchos casos, estas moléculas se sintetizan utilizando compuestos que provienen del metabolismo primario, por lo tanto, algunas de las categorías son definidas por el metabolito primario utilizado.

De este modo entre los metabolitos secundarios podemos encontrar moléculas de bajo peso molecular (<500) (e.g., alcaloides, terpenos, glicósidos, compuestos aromáticos, etc.), moléculas de peso molecular algo mayor producidas por grandes estructuras moleculares a modo de factoría celular como los policétidos, productos de sintasas de ácidos grasos, péptidos no ribosomales y los híbridos de estos tres tipos de moléculas, y por último, moléculas mucho más grandes de tipo polimérico (e.g., péptidos ribosomales, polisacáridos, etc.).

En relación con el trabajo desarrollado en esta Tesis Doctoral, cabe destacar los policétidos, sintetizados por las llamadas policétido sintasas (PKS) y los péptidos no ribosomales, sintetizados a su vez por sintasas de péptidos no ribosomales (NRPS). Los metabolitos secundarios producidos por este tipo de enzimas se sintetizan forma modular pudiendo alcanzar una diversidad y complejidad estructural enorme que los hace muy interesantes a la hora de estudiar nuevas moléculas con potencial actividad farmacológica. En muchas ocasiones los genes que codifican policétidos sintasas y sintasas de péptidos no ribosomales aparecen organizados en clústeres génicos que pueden codificar uno de los tipos enzimáticos o ambos de forma híbrida. A partir de ahora todos estos tipos de clústeres génicos se denominarán como clústeres PKS/NRPS.

3.2.1. | Policétido sintasas

Los policétidos constituyen una gran familia de productos naturales que se pueden encontrar en bacterias, hongos y plantas. Estas moléculas son sintetizadas desde precursores acil-CoA mediante enzimas llamadas policétido sintasas (PKS) y pueden formar compuestos de una gran complejidad estructural normalmente con una potente actividad biológica.

Las PKS son complejos multienzimáticos que sintetizan compuestos naturales denominados policétidos. Se encuentran estructural y funcionalmente muy relacionadas con las sintasas de ácidos grasos (FAS) y al igual que estas catalizan la condensación de metabolitos primarios activados (e.g., acetil-CoA y malonil-CoA) para formar polímeros β-cetoacetil que se unen covalentemente a la enzima mediante enlaces tioester. Esta actividad β-ceto sintasa permite la incorporación secuencial de unidades acetato de dos carbonos en una cadena multi-modular desde la cabeza a la cola de la molécula. En la síntesis de ácidos grasos, esta condensación continúa con un proceso de β-cetorreducción, una deshidratación y una enoil-reducción, resultando un ácido graso reducido (saturado). Sin embargo, en el caso de las policétido síntasas estos pasos pueden ser omitidos o modificados en mayor o menor grado, generándose así una molécula estructuralmente más variable (Fujii *et al*., 2001). De hecho, algunas PKS explotan la reactividad de los policetos intermediarios para facilitar la ciclación intramolecular y la reorganización de los enlaces π, para generar una colección muy diversa de productos sustituidos monocíclicos y policíclicos desde un simple acetilo (Austin y Noel, 2003).

3.2.1.1. | Organización en módulos y dominios en las PKSs

Las PKSs tienen una organización modular y cada uno de los módulos que conforman las unidades de síntesis está formado por una sucesión de distintos dominios

de función definida separados por conectores muy pequeños. Algunos de estos dominios son esenciales y necesarios para la actividad básica de la enzima (Fig. 5), y otros sólo aparecen en ciertas ocasiones en función del tipo de modificaciones adicionales que requiera la síntesis de un compuesto dado. Los dominios esenciales son:

- **Dominio cetosintasa (KS):** su función es la de catalizar una condensación Claisen descarboxilativa entre el policétido que está siendo sintetizado y una nueva unidad que extiende la cadena. Este dominio generan enlaces C-C y cada ciclo catalítico resulta en la adición de dos carbonos a la cadena creciente si se utiliza una unidad malonil o bien tres carbonos si se utiliza una unidad metil-malonil (revisado en Fischbach y Walsh, 2006).
- **Dominio acil-transferasa (AT):** es el encargado de seleccionar y cargar el grupo acilo que extiende la molécula en el dominio ACP durante el ciclo de elongación. En ocasiones, en los sistemas modulares de PKS tipo I, en lugar de un dominio AT dedicado exclusivamente a cada módulo, existe la posibilidad de encontrar enzimas AT disociadas actuando en *trans*. En estos casos, estos sistemas toman el nombre de trans-AT PKSs. El primer sistema trans-AT PKS investigado en detalle fue la síntesis de pederina (Piel, 2002). Sin embargo, este tipo de arquitectura ha ido encontrándose con una frecuencia cada vez mayor, y gracias a esto se han identificado muchos compuestos con diversas actividades farmacológicas producidos por este tipo de clústeres (revisado en Piel, 2010)
- **Dominio portador (*Acyl Carrier Protein*) (ACP):** se trata de un dominio pequeño (80-100 residuos) no catalítico que se encarga de sostener, mediante un enlace tioéster, la cadena creciente policetídica así como la nueva unidad encargada de extender dicha cadena. Para cumplir esta función, este dominio requiere una modificación post-traduccional previa llevada a cabo por una enzima fosfopanteteinil transferasa (PPTasa) que transfiere la porción 4'-fosfopanteteina del CoA a un residuo de serina conservado del dominio portador, transformando dicho dominio de la forma apo inactiva a la forma holo activa (Beld *et al*., 2014).

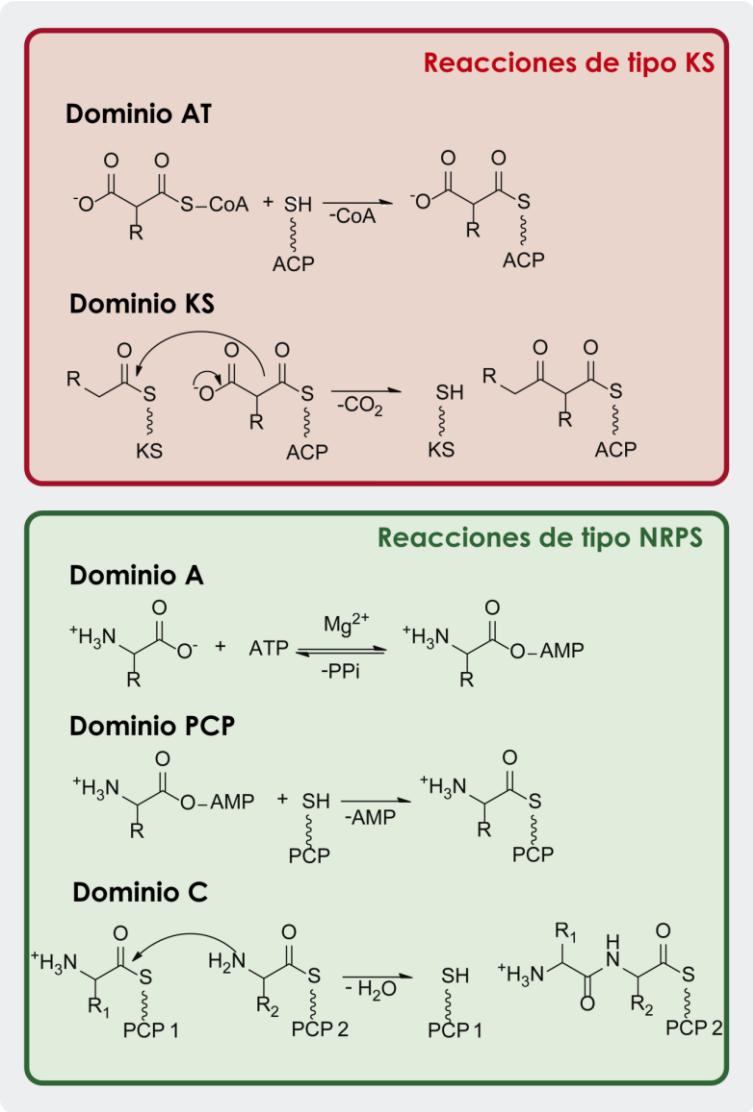

Figura 5 | Reacciones catalizadas por los dominios esenciales de los clústeres PKS y NRPS. Se representan para cada caso aquellas funciones asociadas a un módulo mínimo de síntesis.

Por otro lado, además de estos tres dominios esenciales, existen otros cuya presencia realiza distintas modificaciones sobre la molécula que está siendo sintetizada. Los más relevantes son los siguientes:

- **Dominio tioesterasa (TE):** está involucrado en la liberación de la molécula que está siendo sintetizada del complejo multi-proteico. En concreto, es capaz de romper el enlace tioester del dominio ACP catalizando la ciclación o la hidrólisis del sustrato acilo. Se pueden encontrar dos tipos de dominios TE con distintas estructuras. Los de tipo I actúan en *cis* y suelen encontrarse en el último módulo de síntesis de los clústeres de PKS tipo I. Los de tipo II que son los responsables de la liberación hidrolítica de aquellas cadenas cargadas que se están sintetizando de forma aberrante o se han quedado estancadas (Du y Lou, 2010).
- **Dominio ceto-reductasa (KR):** tanto en PKS como en FAS estos dominios se encargan de reducir de forma estereoespecífica el grupo carbonilo de los intermediarios β-cetoacil-ACP. Además, estos dominios pertenecen a la familia

de las deshidrogenasas/reductasas de cadena corta (SDR), las cuales son conocidas por ser oxidorreductasas dependientes de $NAD(P)^+$/NAD(P)H (Reid *et al*., 2003).

- **Dominio dehidratasa (DH):** se encargan de catalizar reacciones reversibles de deshidratación en los intermediarios β-hidroxi-acil unidos a los dominios ACP. El resultado de esta reacción genera un intermediario acil-ACP αβ-insaturado ya sea en configuración *cis* o *trans*.
- **Dominio enoil-reductasa (ER):** catalizan la reducción de un intermediario enoil-ACP a un acil-ACP α-β-saturado.
- **Dominio metil-transferasa (MT):** catalizan la transferencia de un grupo metilo desde una adenosilmetionina (SAM o AdoMet) a un átomo de carbono, nitrógeno u oxígeno del residuo que está siendo introducido en la cadena y dependiendo de la posición del residuo que se metilada, estos dominios se pueden dividir en C-MT, N-MT y O-MT, respectivamente. Generalmente estos dominios poseen una estructura que consiste en dos subdominios, de los cuales el primero contiene el sitio de unión del donador del grupo metilo y el segundo posee el sitio de unión del sustrato aceptor (Miller *et al*., 2003; Martin y McMillan, 2002).

3.2.1.2. | Mecanismo enzimático de síntesis en las PKSs

En la Figura 6, se esquematiza el mecanismo de un prototipo de PKS dimodular a modo de ejemplo que contiene los tres dominios esenciales (KS, AT y ACP) junto con un dominio KR en el segundo módulo. El primer módulo juega el papel iniciador o de carga y contiene únicamente un dominio AT seguido de un ACP. El segundo módulo contiene los dominios KS, AT, KR y ACP. La síntesis en este ejemplo comienza con la selección del primer sustrato mediante el dominio AT inicial. Este sustrato se transfiere al grupo 4'-fosfopanteteinil del correspondiente dominio ACP activado previamente a la forma *holo*. Este sustrato es recolocado a continuación en el sitio activo del residuo de cisteína del dominio KS. El derivado malonil que a su vez ha sido cargado en segundo dominio ACP adyacente es descarboxilado, lo cual proporciona el nucleófilo necesario para la condensación de Claisen con el grupo cétido unido al dominio KS. De este modo, la cadena inicial siempre se transfiere a la molécula extendedora. Finalmente, una vez que la cadena se ha sintetizado a través de todos los pasos de elongación correspondientes, esta suele liberarse mediante un dominio TE terminal, el cual puede ciclar la molécula final resultante.

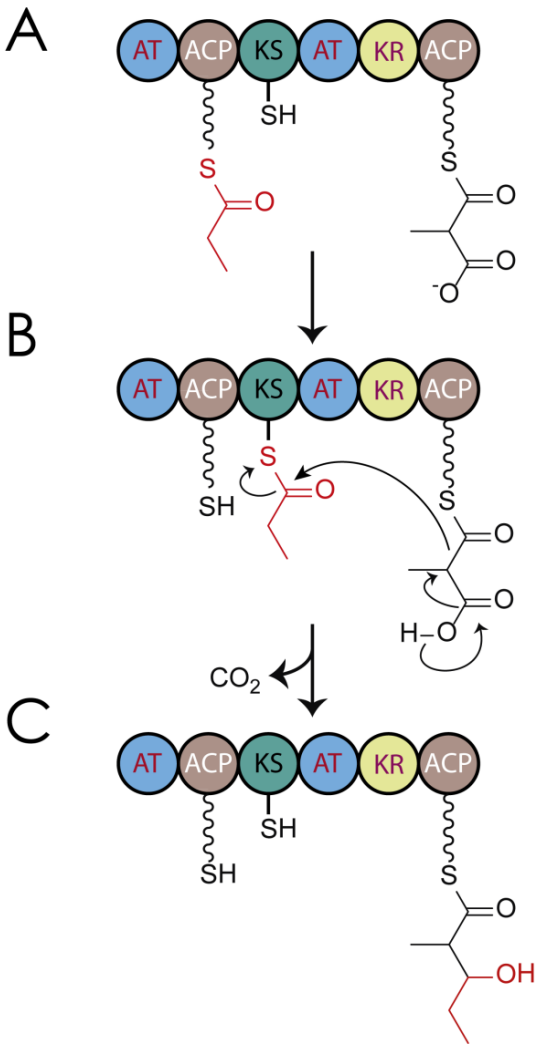

Figura 6 | Esquema del funcionamiento de una PKS bimodular. (A) Fase de carga de las moléculas en los dominios ACP por la acción de los dominios AT. **(B)** Traslocación del residuo propionil (en rojo) al dominio KS. El residuo metil-malonil (en negro) se descarboxila para ceder el paso al nucleófilo para permitir la posterior condensación con el residuo propionil unido al dominio KS. **(C)** La molécula resultante queda unida al dominio ACP del segundo módulo, posibilitando su elongación en pasos sucesivos. La presencia del dominio KR permite la reducción del β-carbonilo a un grupo hidroxilo. Modificado de Marahiel y Essen (2009).

3.2.1.3. | Clasificación de las PKS

Todas las PKS, igual que sus ancestros las FAS, poseen la actividad β-ceto sintasa . En algunos sistemas biosintéticos, los llamados dominios portadores (ACP), se encuentran en proteínas distintas a las que contienen el dominio catalíticos. En estos casos, estas proteínas actúan en *trans* para acabar formando complejos funcionales para las elongaciones de la cadena de acilos. Sin embargo, existen otros sistemas en los que los dominios portadores se encuentran en la misma proteína que los dominios catalíticos fusionados en *cis,* formando módulos y haciendo posible la generación de una línea de ensamblaje multi-modular. Aquellos sistemas PKS que poseen dominios conectados en *cis* se les denominan PKS de tipo I, mientras que aquellos que poseen dominios en *trans* se les llama de tipo II (Fig. 7). Por lo tanto, las PKS de tipo I consisten generalmente en grandes poliproteinas multidominio, las cuales pueden llegar a formar grandes complejos biosintéticos que actúan de forma iterativa o modular. En la síntesis de las PKS de tipo I iterativas, el proceso reutiliza dominios de forma cíclica, mientras que en las PKS de tipo

modular, la síntesis sigue una determinada secuencia de módulos separados en la cual no se repite el mismo dominio, con la excepción de los dominios trans-AT (ver apartado de Dominios acil-transferasa (AT) en la Introducción (3.2.1.1)). También, el grupo de las PKS de tipo I iterativas se puede subdividir a su vez en: i) PKS no reductoras (NR-PKS), cuyos productos son considerados verdaderos policétidos; ii) PKS parcialmente reductoras (PR-PKS); iii) PKS completamente reductoras (FR-PKS), las cuales sintetizan derivados de los ácidos grasos.

Sin embargo, a diferencia de las PKS de tipo I, las PKS de tipo II dirigen el flujo de carbono supuestamente a través de complejos enzimáticos que consisten en proteínas aisladas, como en el caso de las FAS de tipo II encontradas en bacterias y plantas. Estas enzimas, aunque se encuentran separadas y no siempre actúan de forma lineal en la síntesis, se piensa que en algunos casos pueden llegar a formar complejos similares a las PKS de tipo I. La PKS tipo II mínima incluye los dominios KS_{α} y KS_{β}, así como un dominio ACP y solo el dominio KS_{α} contribuye como sitio activo de la condesación. De este modo ocurren iteraciones durante un número determinado de extensiones de la cadena para acabar construyendo una cadena de policetonas. También existe un tipo III de PKS (Fig. 7), el cual no utiliza dominios ACP y de este modo incorpora malonil-CoA en lugar de especies malonil-*S*-panteteinil-T como sustrato (Austin y Noel, 2003).

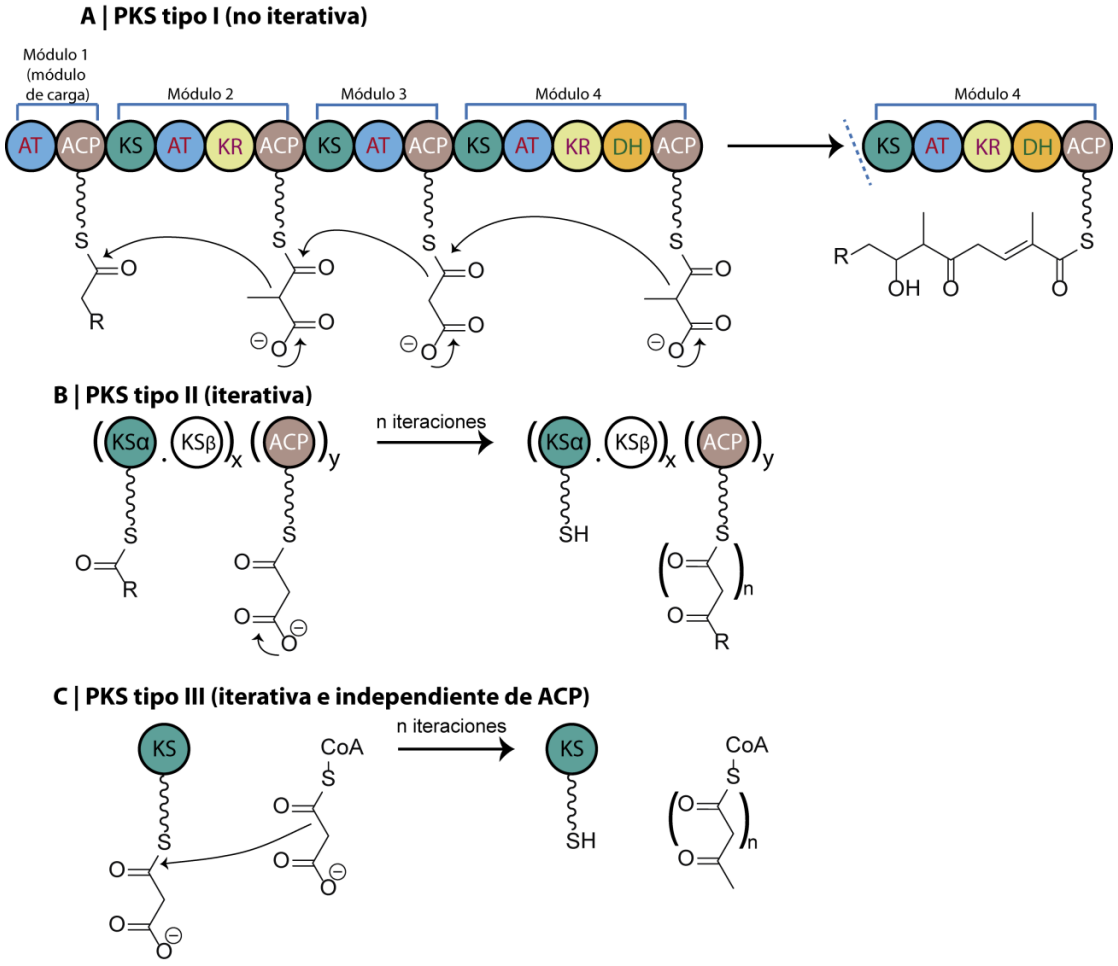

Figura 7 | Esquema del funcionamiento de los tres tipos de PKS. A | Funcionamiento de las PKSs de tipo I. **B |** Funcionamiento de las PKSs tipo 2 poseen funciones catalíticas de forma discreta. **C |** Funcionamiento **de l**as PKSs tipo III. Modificado de Shen (2003).

3.2.2. | Sintasas de péptidos no ribosomales

Las sintasas de péptidos no ribosomales (NRPS) son biocatalizadores multimodulares que poseen la capacidad de sintetizar péptidos macrocíclicos complejos sin utilizar un molde derivado de los ácidos nucleicos. Además de los 20 aminoácidos proteinogénicos, los péptidos no ribosomales pueden estar formados por un gran número de aminoácidos no proteinogénicos, los cuales suelen ser esenciales para su actividad biológica. Como sucede en el caso de las PKS de tipo I, los módulos NRPS son estructuras repetitivas que forman parte de megasintasas, y cada uno de ellos es responsable de la incorporación de un precursor monomérico a la cadena final. En el caso de las NRPS, estos precursores suelen ser productos peptídicos. También, de forma comparable al caso de las PKS, el orden y la especificidad de cada uno de los módulos de síntesis en un clúster biosintético, suele reflejar el producto peptídico no ribosomal resultante (revisado en Marahiel y Essen, 2009).

3.2.2.1. | Organización en módulos y dominios en las NRPS

Las NRPS se organizan en dominios que suelen ser ubicuos en cada uno de los módulos de síntesis. Existen tres dominios esenciales que catalizan en conjunto las reacciones básicas para la elongación de los intermediarios peptídicos. Estos dominios son:

- **Dominio de adenilación (A):** es el responsable de la selección de los aminoácidos (o de sus derivados) que se van a incorporar al producto. Los dominios A activan el sustrato acido-amino, o carboxi, como un amino-acil adenilado consumiendo ATP en el proceso (revisado en Finking y Marahiel, 2004). Analizando la secuencia proteica del dominio se puede observar la presencia de determinados aminoácidos que participan activamente en la selección de la molécula que va a ser incorporada y activada. De este modo, conociendo este código no ribosomal se pueden realizar predicciones sobre cuál será el posible sustrato que se incorporará a la molécula (Stachelhaus *et al*., 1999; Challis *et al*., 2000)
- **Dominio portadores (*Peptidil Carrier Protein*) (PCP):** muy similar a los dominios ACP de las PKS y como estos pertenecen a la superfamilia CP (Carrier Protein). Su función es la de aceptar el derivado aminoacídico unido covalentemente al cofactor 4'-fosfopanteteinil como un tioéster. Como sucede en los dominios ACP, los PCP deben ser activados pasando de su forma *apo* a la forma *holo* por las enzimas fosfopanteteinil transferasas.
- **Dominio de condensación (C):** es el responsable de la formación del enlace peptídico entre el sustrato y la cadena creciente. En concreto, este dominio cataliza el ataque nucleofílico del grupo amino (o imino, o hidroxil) del aminoácido activado uniéndolo a la molécula creciente.

Del mismo modo que en las PKS, existen otros dominios accesorios que no siempre aparecen en los módulos. Estos dominios son los MT y KR, los cuales poseen actividades similares a las descritas para los clústeres PKS. Además, se han observado otros tipos de dominios como son por ejemplo variantes del dominio de condensación, que poseen

capacidad de ciclar (Cy) o dominios de epimerización (E). También cabe desatacar la importancia de los dominios TE, los cuales igual que ocurre en los sistemas PKS, liberan la molécula del complejo macroproteico y en algunos casos la circularizan.

3.2.2.2. | Mecanismo de síntesis en las NRPSs

Tal y como se muestra en el esquema de la figura 8, el mecanismo de síntesis de un módulo prototípico individual de una NRPS comienza con la selección y la activación de la molécula que se va a incorporar mediante el dominio A, proceso que conlleva gasto de ATP. A continuación esta molécula se une al dominio PCP de forma covalente. Seguidamente, el domino C realiza la unión de las moléculas contenidas en los dos dominios PCP contenidos cada uno en uno de los dos módulos consecutivos. De este modo el aminoácido electrófilo (aceptor), entra en la posición 1 del dominio C, y el aminoácido nucleófilo (donador) se acomoda en la posición 2. A continuación se forma un enlace peptídico, cuya molécula resultante volverá a actuar como electrófila en la reacción catalizada por el dominio C del siguiente módulo (revisado en Finking y Marahiel, 2004). Finalmente el dominio TE suele encargarse de liberar la molécula del complejo de proteínas.

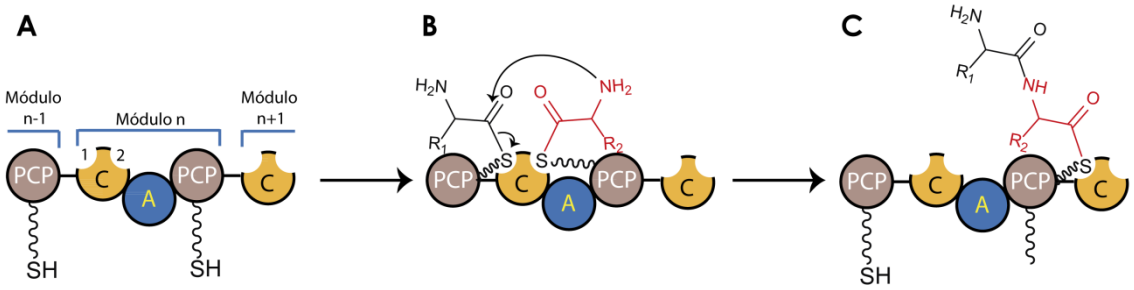

Figura 8 | Funcionamiento de un módulo prototipo de una NRPS. A | Representación mínima de un módulo NPRS con un dominio C y un dominio PCP. **B |** Incorporación de la nueva molécula (rojo) a la cadena que proviene de hipotéticos módulos anteriores (negro). **C |** El dominio PCP cede la cadena para que ocurra el siguiente paso de condensación. Modificado de Finking y Marahiel (2004).

3.2.3. | Moléculas híbridas PKS-NRPS

Debido a que poseen maquinarias de ensamblaje modulares similares, existen multitud de casos en los que las líneas de ensamblaje son híbridas y contienen módulos de PKS y NRPS (Tang *et al*., 2004; Piel, 2002; Du *et al*., 2013). La proporción de módulos de PKS y NRPS es variable, siendo en algunos casos la presencia de PKS mayoritaria (Wu et al., 2000) mientras que en otros lo es la de NRPS (Du *et al*., 2000). Uno de los detalles más interesantes de este tipo de clústeres es el hecho de que cada vez que la línea de ensamblaje cambia de PKS a NRPS o viceversa, la cadena creciente debe ir cambiando de PKS a NRPS o de NRPS a PKS, produciéndose estos saltos mediante reacciones de condensación distintas.

3.2.4. | Genes accesorios en clústeres PKS/NRPS

Además del núcleo donde se encuentran los genes de tipo PKS y/o NRPS, este tipo de clústeres en general están acompañados por otros genes accesorios que poseen funciones relacionadas con la biosíntesis de la molécula en cuestión. Estos genes se

pueden dividir según los distintos cometidos que puedan presentar en el funcionamiento de la síntesis del clúster. En concreto, se pueden distinguir principalmente:

- **Genes reguladores:** como por ejemplo activadores o represores de algunos de los promotores de los genes implicados en la síntesis.
- **Genes de transporte y/o resistencia:** existen bombas de extrusión que se encargan de expulsar la molécula al exterior celular, como en el caso de la streptolidigina (Olano *et al*., 2009; Olano, 2011). En concreto, la toxicidad de la molécula producida puede ser una de las causas que provoquen este tipo de fenómenos.
- **Genes de síntesis de precursores:** como por ejemplo, la síntesis de ácido ciclopentano 1,2-dicarboxílico como unidad iniciadora en el clúster de biosíntesis de borrelidina (Olano *et al*., 2004; Olano, 2011).
- **Genes involucrados en el proceso de modificaciones específicas "tailoring"**: que se encargarían de modificar la molécula resultante una vez esta ha sido sintetizada por el núcleo PKS/NRPS. En este proceso, las moléculas son "decoradas" con distintas modificaciones químicas que amplían aún más el abanico de la diversidad de estructuras que se puede desarrollar con este tipo de clústeres. Estas modificaciones en la mayoría de los casos resultan cruciales para la actividad específica del producto natural que está siendo sintetizado y pueden ser de tipo aminotransferasa, halogenasa, aciltransferasa, oxirreductasa, carbamoiltransferasa, metiltransferasa, ciclasas o glicosiltransferasas, entre otras. Resultan especialmente interesantes las incorporaciones de azúcares, ya que estos pueden ser de una naturaleza muy variada y pueden incorporarse de formas muy diversas siendo un objetivo importante su estudio a la hora de realizar síntesis de nuevas moléculas derivadas (Olano *et al*., 2010).

3.2.5. | Implicaciones de la biosíntesis de clústeres PKS/NRPS de gran tamaño

Los clústeres biosintéticos que contienen PKSs y NRPSs suelen tener un gran tamaño llegando a exceder en ocasiones las 100 kb (Weber *et al*., 2008). Así mismo, las pautas abiertas de lectura (ORFs) pueden contener varios módulos de síntesis, lo cual implica un gran tamaño de la proteína resultante. Un ejemplo es el caso del clúster de PKS lineal correspondiente a la molécula ECO-02301, que contiene 26 módulos de extensión sin contar el módulo iniciador. En total consta de 122 dominios repartidos en únicamente 9 proteínas que pueden alcanzar una masa molecular de 4,7 MDa (McAlpine *et al*., 2005). En el caso de las NRPS un ejemplo es el clúster de biosíntesis de la molécula de syringopeptina, responsable del ensamblaje de 21 residuos aminoacídicos que está formado por 68 dominios contenidos en tan solo 3 proteínas, generando una línea de ensamblaje de una masa molecular de 2,7 MDa (Scholz-Schroeder *et al*., 2003).

Teniendo en cuenta el gran tamaño que puede llegar a alcanzar este tipo de clústeres, no resulta difícil imaginar que las enormes proteínas multifuncionales que se generan pueden tener problemas en el plegamiento además de resultar más propensas a

fenómenos como la proteolisis o la inactivación por mutaciones. De este modo, la estrategia común de organizar los módulos en distintas subunidades proteicas podría evitar parcialmente este problema a la vez que facilita la evolución de los componentes de la línea de ensamblaje, ya sea mediante eliminaciones, adiciones o sustituciones. Sin embargo, el hecho de dividir el núcleo PKS/NRPS del clúster en varias proteínas distintas puede generar nuevos problemas. Una de estas restricciones puede resultar de la dificultad que supone el plegamiento entre dominios separados en proteínas distintas. Por tanto se hace absolutamente necesario que exista una alta afinidad entre los distintos componentes proteicos para que estos puedan interaccionar de manera correcta (revisado en Fischbach y Walsh, 2006). Este hecho se resuelve con la existencia de ciertas regiones de comunicación entre distintas proteínas. Estos dominios llamados COM (*Communication-Mediating Domains*) en el caso de las NRPS están involucrados en la interacción selectiva entre los integrantes del complejo multienzimático (Hahn y Stachelhaus, 2004; Hahn y Stachelhaus, 2006). En el caso de las PKS, estos elementos de unión fueron también descritos en la línea de ensamblaje de la molécula de DEBS, la cual consta de 3 subunidades (Kumar *et al*., 2003).

Desde el punto de vista metabólico, los organismos productores invierten una gran cantidad de recursos y energía para construir estas enormes estructuras moleculares. En el caso de las estructuras peptídicas resulta evidente que desde el punto de vista energético la síntesis usando moldes de RNA es más eficiente que utilizando sistemas NRPS, sin embargo, la síntesis de tipo PKS/NRPS ofrecería a cambio una mayor variabilidad. Debido a este gran coste, la importancia de las funciones de las PKS/NRPS debería garantizar el elevado gasto energético y de recursos que supone su síntesis.

3.3. | El ejemplo de la didemnina B y sus derivados

La didemnina B constituye un ejemplo de molécula de interés producida por un clúster PKS/NRPS cuyo estudio ha sido abordado en esta Tesis Doctoral (Fig. 9). Los primeros miembros de la familia de las didemninas descubiertos fueron la didemnina A, B y C, las cuales se aislaron por primera vez en 1978 de un tunicado del mar Caribe llamado *Tridemnum solidum* (familia Didemnidae) (Rinehart *et al*., 1981). Estos desipéptidos de origen marino tienen actividad antitumoral, antiviral e inmunosupresora a una concentración del orden de nano y femtomolar, por lo que la didemnina B fue el primer producto marino en alcanzar la fase II de ensayos clínicos en Estados Unidos. Además, presentan un mecanismo de acción por el cual la actividad antitumoral radica en la inducción de la apoptosis celular al unirse al factor de elongación EF-1α, inhibiendo de este modo la síntesis de proteínas mediante la competición con la unión del GTP. Sin embargo, debido a la toxicidad del compuesto, los ensayos clínicos terminaron en 1990 (revisado en Lee *et al*., 2012).

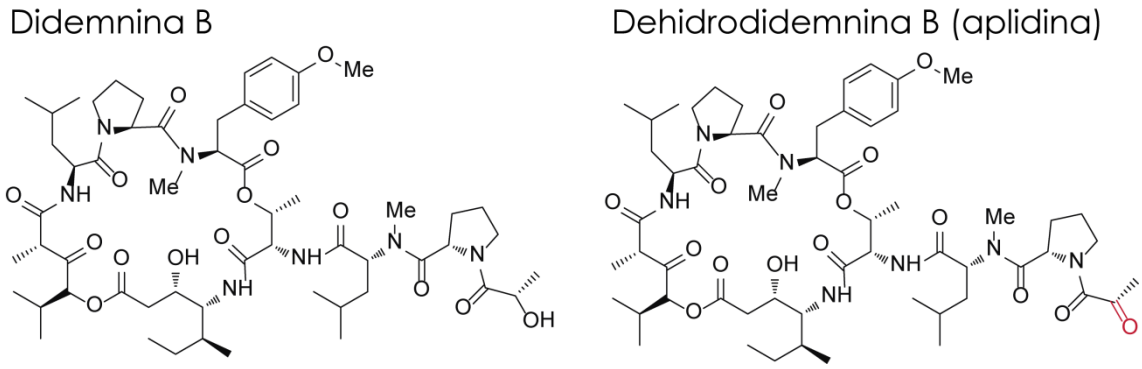

Figura 9 | Diferencias entre las moléculas de la didemnina B y de la aplidina. Representación de la estructura de las moléculas de didemnina B y de dehidrodidemnina B (aplidina). En rojo se marca la única diferencia entre las moléculas de la aplidina y de la didemnina B.

Sin embargo, a pesar del fracaso terapéutico inicial de las didemninas, estas han recuperado el interés farmacéutico debido a la posible utilidad de una molécula muy similar, la aplidina. La aplidina o dehidrodidemnina B es un compuesto derivado de la didemnina que actualmente despierta gran interés por sus implicaciones clínicas (Rinehart y Lithgow-Bertelloni, 1991 (Patente: WO 9104985A1)). Esta molécula se diferencia de la didemnina B únicamente en la sustitución de un grupo hidroxilo por un ceto en el primer residuo incorporado de la síntesis (Fig. 9). La aplidina es uno de los inductores de la apoptosis más potentes descritos hasta la fecha y se postula como un fármaco muy prometedor para el tratamiento de la leucemia. Actualmente, solo se obtiene mediante síntesis química, ya que únicamente ha sido aislada de forma natural del tunicado marino *Aplidium albicans* encontrado en Ibiza (España), pero se desconoce el microorganismo productor y por lo tanto la estructura del su clúster génico de biosíntesis (revisado en Lee *et al*., 2012). Por esta razón, las diversas aproximaciones para abordar la obtención biosintética o semisintética de la molécula podrían disminuir los costes de producción del futuro fármaco. Dado el gran parecido molecular de la aplidina con la didemnina B, un objetivo muy interesante sería el de intentar redirigir la síntesis de las didemninas hacia aplidina utilizando herramientas de ingeniería genética y biología molecular actuando sobre el clúster de biosíntesis de la propia didemnina B.

4. Herramientas para la manipulación de clústeres PKS/NRPS

A la hora de proceder con la extracción de clústeres PKS/NRPS de una fuente ambiental, existen diferencias en las herramientas utilizadas dependiendo de que el microorganismo productor sea cultivable o no cultivable.

Debido a que técnicamente resulta más fácil su manipulación, el hecho que el microorganismo simbionte sea cultivable supone una gran ventaja a la hora de extraer clústeres biosintéticos. Sin embargo, uno de los principales cuellos de botella radica en la pequeña proporción de microorganismos cultivables existentes en las muestras. En el caso de las esponjas marinas, al igual que en otros ambientes, menos del 1% de los taxones observados por otros medios han sido cultivados en condiciones de laboratorio usando de aproximaciones tradicionales. Sin embargo, el caso de que el microorganismo que

contiene el clúster de interés no sea cultivable entraña el uso de herramientas más complejas cuyo funcionamiento implica una mayor dificultad técnica.

4.1. | Herramientas para la extracción de clústeres PKS/NRPS de simbiontes cultivables

Tras la identificación de un compuesto de interés farmacológico el principal objetivo es definir un proceso biotecnológico que garantice su suministro y rebaje los costes derivados de su síntesis química. En el caso de un microorganismo cultivable, la producción se podría abordar cultivando directamente el microorganismo en condiciones de producción o bien intentando transferir los genes biosintéticos a un hospedador que ofrezca ventajas en su cultivo y/o en el proceso de producción. Para ello se debe conseguir rescatar y extraer la información relativa a la biosíntesis de estas moléculas del propio microorganismo productor.

4.1.1. | Cultivo de microorganismos simbiontes productores

Existen multitud de estudios en los que se ha trabajado diversas condiciones de cultivo para ampliar el número de microorganismos cultivables (Muscholl-Silberhorn *et al*., 2008; Kennedy *et al*., 2010). Otros autores se han focalizado en el aislamiento de microorganismos pertenecientes a un grupo taxonómico de interés, ya que de este modo se optimizan las condiciones de cultivo para obtener bacterias cultivables con la máxima variabilidad dentro del taxón en concreto (Xi *et al*., 2012). Otras aproximaciones más innovadoras intentan aislar comunidades bacterianas concretas tras la administración de determinados agentes antibióticos (Richardson *et al*., 2012), o a través de metodologías de cultivo que utilizan filtros flotadores (Sipkema *et al*., 2011). Sin embargo, a pesar de estos esfuerzos, suelen aparecer aislados de forma repetida los mismos phyla bacterianos, que comprenden principalmente miembros de Proteobacteria, Firmicutes, Actinobacteria, Planctomycetes, Verrucomicrobia, Cyanobacteria y Bacteroidetes (revisado en Taylor *et al*., 2007). A pesar de estas limitaciones, el cultivo de simbiontes suele aportar multitud de nuevas especies y es una herramienta útil para localizar nuevos microorganismos productores de productos naturales de interés.

4.1.1.1. | Detección del organismo cultivable productor

Otro de los principales obstáculos a la hora de localizar al microorganismo productor, es el proceso de cribado de los distintos aislados. Habitualmente, este proceso puede llevarse a cabo realizando cultivos para posteriormente detectar la presencia del producto de interés. Pero frecuentemente, a pesar de que el microorganismo sea el productor, el compuesto de interés no llega a sintetizarse en las condiciones de cultivo seleccionadas. En estos casos, es de suma importancia encontrar medios de cultivo y/o condiciones que induzcan la biosíntesis, ya que por diversos motivos, la ruta de producción del compuesto de interés se puede encontrar silenciada.

A veces, cuando se conoce bien la naturaleza del compuesto que se pretende producir y se puede deducir la posible estructura de los genes de biosíntesis, se pueden analizar los clústeres diana presentes en los microorganismos mediante amplificación por

PCR de su genoma, utilizando oligonucleótidos degenerados de secuencias conservadas en los dominios de los genes de biosíntesis. Este el caso de los policetidos y pétidos no ribosomales donde se puede anticipar la presencia de PKS o NRPS (Romero *et al*., 1997; Metsä-Ketelä *et al*., 1999; Ayuso-Sacido y Genilloud, 2005). El cribado mediante amplificación puede resultar útil para hacer una primera aproximación e intentar discernir entre microorganismos que poseen posibles clústeres génicos de interés y aquellos que no.

4.1.2. | Herramientas genómicas

Una vez se ha conseguido identificar y aislar el microorganismo responsable de la biosíntesis del producto de interés, se pueden utilizar diferentes herramientas genómicas con el fin de extraer la información necesaria sobre la naturaleza de los genes de biosíntesis. Esta información será muy útil si se quiere abordar una mejora de la producción del compuesto en cuestión de manera dirigida mediante herramientas de ingeniería metabólica o biología de sistemas.

4.1.2.1. | Secuenciación masiva de DNA para la obtención de las secuencias de rutas de biosíntesis de productos naturales

Una de las herramientas más útiles para obtener la secuencia de los genes de biosíntesis del compuesto de interés consiste en secuenciar el genoma completo del microrganismo productor aislado mediante técnicas de secuenciación masiva. Estos métodos de secuenciación masiva de nueva generación fueron desarrollados a principios del siglo XXI y comenzaron a comercializarse a partir del año 2005 (revisado en Hall, 2007). En la actualidad existen diversas tecnologías diferentes de secuenciación masiva como son por ejemplo la pitosecuenciación 454, Illumina, SOLiD, Ion Torrent semiconductor o la secuenciación SMRT (*Single molecule real time*). Dentro del conjunto de las tecnologías de secuenciación masiva, en el desarrollo de este trabajo de Tesis Doctoral se han utilizado las siguientes:

4.1.2.1.1. | Pirosecuenciación (454 Life Sciences, Roche GS-FLX)

Este método de secuenciación, basado en la síntesis de DNA, se caracteriza por la detección de la liberación de un pirofosfato cuando se incorpora un nucleótido en dicha síntesis. La hebra simple de DNA que sirve de molde se hibrida con el cebador de secuenciación y se incuba con las enzimas DNA polimerasa, ATP sulfurilasa, luciferasa y apirasa además de con los sustratos adenosin 5' fosfosulfato (APS) y luciferina. La adición del dNTP que corresponde al complementario de la siguiente posición en la cadena provoca que al incorporarse gracias a la acción de la polimerasa, se libere pirofosfato (PPi). La enzima ATP sulfurilasa convierte este pirofosfato en ATP (en presencia de APS) el cual actúa como sustrato en la conversión mediada por la luciferasa, de luciferina a oxiluciferina que genera luz visible de forma proporcional a la cantidad de ATP. La luz generada se detecta mediante una cámara y se analiza en un pirograma. Aquellos nucleótidos no incorporados y el ATP se degradan por la enzima apirasa, por lo que la reacción puede volver a comenzar con otro nucleótido. Este proceso se repite para cada uno de los cuatro nucleótidos hasta que se determina la secuencia complementaria de la

cadena molde (Fig. 10) (revisado en Ahmadian *et al*., 2006)]. Una de las principales ventajas que presenta esta tecnología radica en la gran longitud media de lectura que se puede alcanzar, la cual actualmente puede rondar las 700 bp por lectura. Sin embargo, este método puede presentar errores de secuencia al enfrentarse a homopolímeros en la secuencia analizada, especialmente cuando la lectura está muy avanzada.

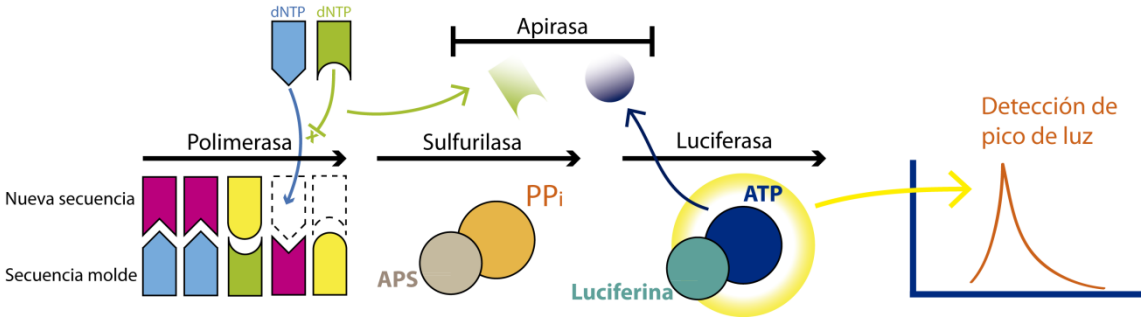

Figura 10 | Esquema del funcionamiento del proceso de pirosecuenciación. Se representa la actuación de las enzimas polimerasa, sulfurilasa, luciferasa y apirasa.

4.1.2.1.2. | Ion Semiconductor (Ion Torrent)

Al igual que la pirosecuenciación, este método también se basa en la secuenciación por síntesis de DNA. La característica principal de este método es que se detectan los protones que se liberan durante la polimerización del DNA. Cada vez que se incorpora un nucleótido a la cadena naciente de DNA junto con un pirofosfato se libera también un protón. Por lo tanto el proceso consiste en detectar los protones liberados en cada paso de secuenciación y saber qué tipo de nucleótido y cuantos se han incorporado. Para ello se añaden los dNTPs de forma secuencial igual que en la pirosecuenciación. Si el nucleótido añadido es complementario al de la hebra de DNA molde, se incorpora a la hebra de DNA complentaria. Esta reacción provoca que se libere un protón, el cual es detectado por un sensor de iones ISFET (Transistor de efecto de campo sensible a iones, *Ion Sensitive Field Effect Transitor*) sensible a pH. La señal electrónica que se obtiene es proporcional al número de protones que se han liberado, por lo que se pueden detectar varias incorporaciones del mismo nucleótido a la cadena en el mismo ciclo (Fig. 11) (Rothberg *et al*., 2011). Entre las principales ventajas que suscita el uso de esta tecnología se encuentra el bajo coste que supone el equipamiento y la mayor cantidad de secuencia generada con respecto a la pirosecuenciación. Sin embargo, con este método también se obtienen errores de secuencia en presencia de homopolímeros y además presenta un tamaño de lecturas de 400 bp, es decir más corto que en la pirosecuenciación.

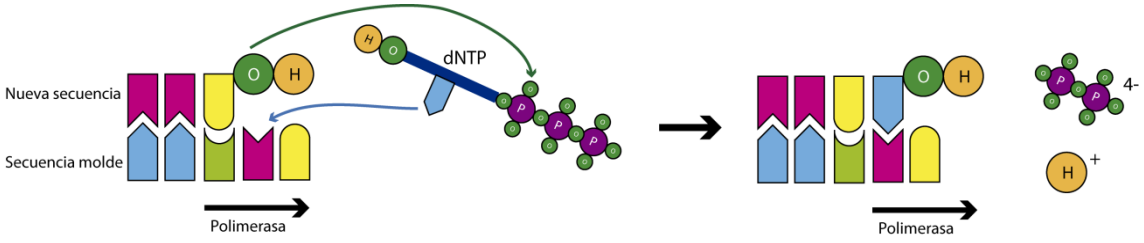

Figura 11 | Esquema del funcionamiento del sistema de secuenciación de DNA del Ion Semiconductor (Ion Torrent).

4.1.2.2. | Herramientas bioinformáticas para la secuenciación de genomas

Debido a la gran cantidad de datos generados por los procesos de secuenciación masiva, para extraer la secuencia del clúster PKS/NRPS de interés resulta esencial el uso de herramientas bioinformáticas pudiéndose obtener de este modo las secuencias génicas asociadas y toda la información que estas contienen.

Una vez se obtienen los datos correspondientes a la secuencia los datos vienen representados de tal forma que para cada uno de los nucleótidos secuenciados, se conoce la fiabilidad con la que ese dato ha sido obtenido. De este modo, los datos de secuencia se presentan con datos la calidad de la misma. Entre los formatos más comunes que se pueden obtener en de las distintas plataformas de secuenciación se encuentran SFF y FASTQ, los cuales han sido utilizados en esta Tesis Doctoral. Además, el hecho de disponer de datos de calidad, permite realizar pasos de previos de procesamiento en la secuencia obtenida para así evitar posibles errores y artefactos, y de este modo generar un posterior ensamblaje de más calidad.

4.1.2.2.1. | Ensamblaje de secuencias genómicas

El ensamblaje de secuencias genómicas se refiere al alineamiento y a la unión de fragmentos pertenecientes a una secuencia de DNA más larga, con el objetivo de reconstruir dicha secuencia original. Para realizar esta tarea existen multitud de algoritmos bioinformáticos diferentes llamados ensambladores. Dependiendo del caso concreto de estudio, existen dos tipos de procesos de ensamblaje diferentes:

- **Ensamblaje *de novo***: se trata del ensamblaje de lecturas pequeñas para obtener secuencias completas que en la mayoría de ocasiones se desconocen previamente.
- **Ensamblaje mediante mapeo**: en este caso las lecturas son ensambladas utilizando una secuencia molde. En este caso la secuencia final obtenida es similar, aunque no necesariamente idéntica al molde utilizado.

Tras el proceso de ensamblaje *in silico*, *e*n la mayoría de los casos de ensamblaje *de novo* no se puede completar la secuencia total, por lo que como resultado se obtienen secuencias ensambladas contiguas llamadas *contigs* las cuales se detienen cuando llegan a zonas donde no se puede continuar con el ensamblaje (*gaps*). En el caso de que la secuenciación se haya realizado utilizando la tecnología de extremos pareados (PE), es posible ordenar la secuencia de los *contigs* aunque existan *gaps*. Estos conjuntos de *contigs* ordenados se denominan *scaffolds*.

La complejidad del ensamblaje de las secuencias generalmente está influida por dos factores, el número de fragmentos y sus longitudes. El caso de que los fragmentos sean más numerosos y más largos permite una mejor identificación de los solapamientos de la secuencias aunque puede conllevar otros problemas de complejidad en los algoritmos. Sin embargo, mientras que las secuencias cortas se alinean de forma más rápida, esto puede

completar algunas fases del ensamblaje debido a que las secuencias cortas son más difíciles de usar con zonas con repeticiones idénticas o casi idénticas.

4.1.2.2.2. | Anotación de genomas

La anotación de genomas es el proceso mediante el cual se asigna información biológica a las secuencias y consiste en tres pasos principales: i) La identificación de porciones del genoma que no codifican proteínas, ii) la identificación de elementos del genoma (predicción génica), iii) la asignación de la información biológica a estos elementos.

El nivel básico de anotación consiste en el uso de herramientas comparativas, como por ejemplo BLAST, para encontrar similitudes entre la secuencia problema y aquellas que previamente han sido depositadas en las bases de datos.

Existen dos fases en los procesos de anotación:

- **Anotación estructural:** consiste en la identificación de elementos genómicos como los son las ORFs y su localización, la estructura de los genes, las regiones codificantes y la localización de elementos regulatorios.
- **Anotación funcional:** consiste en asignar información biológica a los elementos previamente descritos en la anotación estructural. De este modo se puede intentar dilucidar características como las funciones bioquímica, biológica y regulatoria de las secuencias.

Existen multitud de herramientas automáticas de anotación *in silico*, sin embargo, la curación manual de los resultados obtenidos de procesos automáticos resulta un paso esencial para obtener datos de calidad.

4.1.2.2.3. | Anotación funcional especializada en clústeres PKS/NRPS

En los casos en los que se necesiten procesos de anotación especializados que permitan identificar con mayor eficiencia clústeres PKS/NRPS en los genomas secuenciados, es posible recurrir a herramientas de anotación automáticas *in silico* que están orientadas a la búsqueda de este tipo de características. Herramientas como por ejemplo AntiSMASH (Medema *et al*., 2011), permiten indagar en características que otras herramientas automáticas genéricas de anotación no detectarían. Entre estas características resulta de especial importancia por ejemplo, tener en cuenta la arquitectura modular específica que suelen tener los clústeres PKS/NRPS, la detección de los genes circundantes al núcleo PKS/NRPS, la comparación de las estructuras del clúster completas con otros clústeres ya conocidos, la predicción de las subunidades que van a ser incorporadas a la cadena de síntesis e incluso la predicción orientativa de la molécula final resultante. De este modo se pueden dilucidar de forma mucho más efectiva las características del clúster de síntesis del producto natural de interés.

4.1.2.3. | Vectores de clonación de gran capacidad para la extracción de secuencias de clústeres de síntesis de productos naturales

La creación de genotecas puede resultar una herramienta de gran utilidad para extraer secuencias de clústeres PKS/NRPS. El gran tamaño que puede alcanzar este tipo de clústeres, dificulta el uso de otras tecnologías como la amplificación mediante PCR o la síntesis artificial, la cual puede resultar económicamente muy costosa. El gran tamaño de este tipo de clústeres génicos implica ciertas consideraciones a la hora de seleccionar el tipo de vector de clonación que se ha de utilizar. En este sentido, los BACs (Cromosomas Bacterianos Artificiales, *Bacterial Artificial Chromosome*), pueden albergar fragmentos de hasta 350 kb y son una opción que puede resultar útil para el clonaje de clústeres biosínteticos de gran tamaño. La estabilidad de estos vectores radica en el sistema de replicación y estabilidad del plásmido F de *Escherichia coli*, el cual regula la cantidad de copias del vector, manteniéndolo a niveles muy bajos y evitando posibles eventos de recombinación no deseados (revisado en Shizuya y Kouros-Mehr, 2001).

Otros vectores de gran capacidad que se basan en el plásmido F al igual que los BAC son los fósmidos (Kim *et al*., 1992). La principal ventaja que presentan estos vectores al realizar el clonaje, consiste en que únicamente aceptan fragmentos de aproximadamente 40 kb, con lo que se evita obtener genotecas poco eficientes por el hecho de que haya fragmentos de DNA de poco tamaño. Sin embargo, el tamaño de fragmento que pueden contener es menor que el de los BACs y en muchas ocasiones no es suficiente para extraer una clúster PKS/NRPS completo en un único vector.

Como se ha comentado anteriormente, la estructura modular de los clústeres biosínteticos facilita que en ciertos casos se pueda hacer una propuesta de síntesis a partir de la estructura del compuesto. De esta forma, una vez generada la genoteca correspondiente en un hospedador heterólogo como es *E. coli* se puede acordar el cribado mediante amplificación mediante PCR de los clones obtenidos utilizando cebadores específicos o degenerados diseñados para amplificar un fragmento de la secuencia que se está buscando (Schirmer *et al*., 2005, y Tae y Fuerst, 2006).

4.2. | Herramientas para la extracción de clústeres PKS/NRPS de microbiomas

Aunque como se ha mencionado anteriormente existen numerosas herramientas para identificar y caracterizar los clústeres génicos de producción de compuestos de interés en microorganismos cultivables la gran mayoría de la diversidad genética microbiana que existe en el planeta, y en particular en el caso de los endosimbiontes de las esponjas es actualmente inaccesible mediante el uso de estos métodos tradicionales, ya que la mayoría de las especies no son cultivables. Por este motivo la aparición de herramientas y tecnologías emergentes en el campo de la metagenómica ofrece un enorme potencial a la hora de descubrir y explotar nuevas entidades biosintéticas en microbiomas completos (revisado en Dobson *et al*., 2015).

4.2.1. | Herramientas metagenómicas

La metagenómica estudia el conjunto de los genomas de las especies que habitan en un determinado nicho (metagenoma) mediante el análisis del DNA total obtenido a partir de una muestra de ese ambiente, sin necesidad de aislar y cultivar esas especies. De esta manera se genera un perfil de diversidad de las especies que forman parte de la muestra natural al que no se puede acceder mediante métodos basados en el cultivo. Debido a esta propiedad de revelar la diversidad que antes permanencia oculta y gracias a la accesibilidad cada vez mayor de las técnicas de secuenciación masiva, la metagenómica posee un potencial enorme para revolucionar el estudio del mundo microbiano, permitiendo además estudiar su ecología a una escala mucho mayor (Eisen *et al*., 2011).

4.2.1.1. | Genotecas metagenómicas

Tradicionalmente se ha utilizado la generación de genotecas metagenómicas para estudiar la diversidad de los nichos microbianos. Aunque el uso de este tipo de herramientas dependiente de la clonación puede generar cierto sesgo en los resultados, en algunas ocasiones es una herramienta que puede ser de cierta utilidad. Por lo tanto a pesar de sus desventajas, el uso de metagenotecas tiene como fin extraer clústeres génicos de gran tamaño a partir del DNA metagenómico. Además, también puede ser una herramienta útil para extraer e identificar directamente actividades enzimáticas de los metagenomas, si bien en este caso se requieren técnicas complementarias de cribado basadas en la robotización de los análisis. En este sentido, existen varios estudios donde se han realizado metagenotecas de distintos nichos de ecosistemas marinos, entre los que por ejemplo se encuentran algunas esponjas (Abe *et al*., 2012; Juan *et al*., 2006). Aunque esta aproximación ha sido la fuente del descubrimiento de varias enzimas novedosas (Okamura *et al*., 2010; Selvin *et al*., 2012), sin embargo, existen aún muchas limitaciones a la hora de ir un poco más allá de la detección de las actividades enzimáticas más tradicionales en los procesos de cribado cuando se trata de conseguir aislar la maquinaria de síntesis completa de algún producto natural muy complejo.

Como ya se comentó anteriormente, el gran tamaño de los clústeres metabólicos PKS/NRPS requiere la utilización de vectores de alta capacidad de almacenaje. Además, resulta de gran importancia disponer de un DNA de gran calidad y de alto peso molecular, con el fin de poder obtener las rutas biosintéticas completas. Esto hace que la posibilidad de generar genotecas de BACs o de fósmidos utilizando DNA metagénomico sea un reto desde el punto de vista técnico.

Cuando se trata de rastrear metagenotecas para buscar grandes clústeres génicos productores de un compuesto de interés, es muy frecuente que no se pueda ensayar su producción o incluso que sea difícil ensayar una activad enzimática relacionada con la biosíntesis, debido a que no es siempre factible que los genes se expresen en el organismo hospedador. Por eso la opción más eficaz para rastrear genes de interés enlas metagenotecas consiste en realizar hibridaciones o amplificación de PCR con sondas de secuencia conocida idénticas o al menos homólogas a algunos genes del clúster que puedan servir para detectar aquellos clones que contengan los genes de interés (revisado en Kennedy *et al*., 2010). En este tipo de estrategias, la posible fragmentación de la ruta debido a roturas del DNA durante la extracción no resulta tan limitante, ya que se pueden

realizar búsquedas individualizadas de los fragmentos del clúster de interés. Si estos se encuentran en clones distintos, se podrán ensamblar *in vitro* o *in vivo* a posteriori y así obtener la secuencia completa del clúster.

En el caso de que la secuencia del clúster sea completamente desconocida y no se disponga de una propuesta de síntesis, la aproximación que se puede utilizar solo puede estar basada en el uso de protocolos de metagenómica funcional (Brady, 2007). De este modo, es de vital importancia que se genere muy poca fragmentación del DNA extraído para obtener la ruta de síntesis completa en un único clon, y que este sea identificado mediante un cribado basado en la producción del compuesto de interés. Es evidente que la dificultad de esta aproximación es directamente proporcional a la longitud de la secuencia del clúster de interés. Además, el éxito de este tipo de estrategias radica en la posibilidad de una producción heteróloga del compuesto. Por tanto, existen multitud de factores a tener en cuenta para abordar la síntesis del compuesto, como por ejemplo, la diferencia en el uso de codones, la falta de control sobre los elementos reguladores del clúster que incluso pueden estar fuera del mismo, la ausencia de metabolitos precursores, o la dificultad del manejo de grandes estructuras proteicas en microorganismos no especializados, son sólo algunos de los problemas que se pueden encontrar.

4.2.1.2. | Secuenciación masiva de metagenomas

Los avances en el campo de la bioinformática y el aumento de la capacidad computacional de los ordenadores han hecho posible realizar el análisis de las secuencias de DNA que se rescataban de las muestras ambientales, permitiendo de este modo la adaptación de las técnicas de la secuenciación *shotgun* utilizadas para la secuenciación de genomas individuales a estas muestras metagenómicas más complejas. Este tipo de aproximaciones y herramientas bioinformáticas permiten ensamblar las lecturas cortas derivadas de la secuencia del DNA (revisado en Segata *et al*., 2013). Históricamente, cuando solo se disponía del método de secuenciación de Sanger, para facilitar el ensamblaje se construían y secuenciaban las genotecas de fragmentos de DNA previamente clonados, paso que ha sido omitido gracias al desarrollo de la secuenciación masiva, con lo que además se elimina uno de los sesgos y cuellos de botella principales en la toma de muestras medioambientales que es la clonación previa del DNA.

De acuerdo con la estructura que suelen mostrar las comunidades microbianas en el medio ambiente, los genomas de los organismos más abundantes en la muestra, son los más representados en las secuencias metagenómicas obtenidas. En el caso de los microorganismos menos abundantes el acceso a estas fracciones "enmascaradas" puede requerir cantidades de secuencia en ocasiones económicamente inviables. Sin embargo, la naturaleza aleatoria de este método de secuenciación asegura que la mayoría de la diversidad esté representada al menos como fragmentos pequeños de secuencia, generándose así análisis más completos y diversos de la microbiota que aquellos realizados con aproximaciones basadas solamente en el aislamiento mediante cultivos.

Las dos aproximaciones más utilizadas en la actualidad para realizar búsquedas concretas de clústeres biosintéticos PKS/NRPS en metagenomas son la secuenciación masiva junto con la amplificación por PCR de secuencias marcadas (Medema *et al*., 2015). Sin embargo, la secuenciación masiva de genomas suele reportar mejores resultados a la

hora de encontrar clústeres novedosos. Aun así, existen multitud de limitaciones en este campo debido principalmente a la poca longitud de las secuencias obtenidas. No obstante estas limitaciones técnicas serán resueltas pronto debido a las nuevas tecnologías de secuenciación que se anticipa que estarán disponibles en los próximos años. De este modo, en un futuro próximo, la combinación de tecnologías que proporcionen grandes longitudes de secuencia y las nuevas aproximaciones bioinformáticas (Albertsen *et al*., 2013 y Nielsen *et al*., 2014) podrían permitir la obtención de las secuencias genómicas completas de la mayoría de los miembros más representados en las comunidades bacterianas, y con ello, se lograrían obtener multitud de nuevas secuencias de clústeres PKS/NRPS de interés.

4.2.1.3. | Herramientas bioinformáticas para la metagenómica

El tratamiento bioinformático de los datos obtenidos de la secuenciación masiva de metagenomas para la búsqueda de funciones determinadas suele ser más complejo que en el caso de un genoma individual. La gran complejidad de algunas de las poblaciones, sumado al pequeño tamaño de las lecturas obtenidas hace que en ocasiones haya que recurrir a herramientas especializadas de ensamblaje. Aun así, aunque depende de las características poblacionales de la muestra, la proporción de muestra ensamblada suele incluir aquellos microorganismos predominantes y por lo tanto no se representa la mayoría de la misma.

Aunque existen herramientas de anotación metagenómicas especializadas en secuencias cortas, como por ejemplo MG-RAST (Meyer *et al.*, 2008), la cantidad y la fiabilidad de la anotación obtenida hace necesario que en la mayoría de los casos se deba trabajar a la hora de asignar funciones únicamente con aquella secuencia que ha podido ser ensamblada. En concreto, debido a la arquitectura de los clústeres PKS/NPRS, su búsqueda en metagenomas requiere de un mínimo nivel de ensamblaje de la secuencia para poder obtener resultados positivos tras la anotación. Aunque la secuencia del clúster de interés obtenida en primer lugar sea parcial, existen aproximaciones bioinformáticas que permiten la organización de las secuencias para obtener la mayor cantidad de datos posibles (Albertsen *et al*., 2013; Nielsen *et al*., 2014). Además existen otro tipo de herramientas filogenómicas como NaPDoS que trabajan con secuencias de dominios concretos y proporcionan pistas sobre su pertenencia a un posible clúster de interés.

Sin embargo, aunque las herramientas de identificación de clústeres PKS/NRPS en genomas y en metagenomas se siguen desarrollando y mejorando, actualmente la lista de posibles clústeres génicos biosintéticos es enorme y sigue creciendo. Por lo tanto, hoy en día uno de los desafíos en la bioinformática es el desarrollo de estrategias de trabajo para sistematizar grandes cantidades de datos químicos y genéticos que conecten la información genómica y los datos metabolómicos existentes (Medema *et al*., 2015).

4.2.2. | Otras herramientas para la extracción de clústeres PKS/NRPS de microbiomas

Como ya se ha comentado, una de las principales limitaciones de las aproximaciones basadas en la secuenciación masiva de metagenomas es que no se puede acceder a determinados individuos minoritarios de las poblaciones microbianas. En ocasiones, estos individuos pueden contener secuencias de PKS/NRPS de interés y la aplicación de técnicas de secuenciación masiva puede resultar en una representación

parcial o nula de dicha secuencia. Para poder acceder a dichas secuencias se pueden combinar los estudios metagenómicos con otras técnicas relacionadas con la genómica de célula única (Dodsworth *et al*., 2013). De este modo, las células individuales se pueden aislar mediante separación de células activada por fluorescencia (FACS) o técnicas de micromanipulación entre las que se incluyen separación por microfluídica, micropipeteo o el uso de pinzas ópticas. Las células aisladas se lisan para obtener cantidades de DNA del orden de femtogramos, que a su vez pueden amplificarse y secuenciarse (Lasken, 2012; Kalisky *et al*., 2011). Sin embargo, la limitación fundamental de este tipo de aproximaciones radica en la necesidad de establecer algún método de selección para poder acceder a la célula del microorganismo de interés.

IV. Objetivos

El desarrollo y la optimización de herramientas para la obtención del suministro de productos naturales de origen marino sintetizados por PKSs y/o NRPSs es un tema de gran relevancia en los últimos años debido al interés biotecnológico de desarrollar nuevas drogas, como por ejemplo fármacos antitumorales. A pesar de los recientes avances llevados a cabo en este campo, aún siguen existiendo numerosos cuellos de botella en los procesos de obtención de estas moléculas, por lo que se hace necesario evaluar las técnicas y herramientas existentes para así lograr optimizarlas y desarrollar nuevas aproximaciones más eficientes.

Por lo tanto, al comienzo de esta Tesis Doctoral se propusieron los siguientes objetivos:

1. Desarrollar herramientas genómicas para optimizar la producción de moléculas sintetizadas por clústeres génicos que codifican PKSs y/o NRPSs en microorganismos cultivables.

Para llevar a cabo este objetivo se utilizó como prueba de concepto la producción de didemninas por el microorganismo simbionte cultivable *Tistrella mobilis* MES-10-09-028 y se propusieron los siguientes objetivos parciales:

- Obtención de la secuencia y análisis del clúster génico productor de didemninas.

- Modificación del clúster productor de didemninas para producir posibles nuevos intermediarios.

- Traslado de la producción de didemninas a un hospedador heterólogo.

2. Desarrollar herramientas bioinfórmaticas para identificar secuencias pertenecientes a clústeres génicos que codifican PKSs y/o NRPSs en microbiomas.

Para desarrollar este objetivo se utilizaron como pruebas de concepto los microbiomas de tres esponjas marinas diferentes en las cuales se habían identificado previamente compuestos citotóxicos de interés farmacéutico. Por lo tanto, en este caso se propuesieron los siguientes objetivos parciales:

- Secuenciación del DNA metagenómico de las fracciones microbianas de tres esponjas marinas.

- Ensamblaje, clasificación y análisis de las secuencias metagenómicas.

- Utilización de herramientas para la identificación de las secuencias génicas correspondiente a los clústeres de interés.

V. Materiales y Métodos

1. Cepas bacterianas y otros organismos utilizados

Las cepas utilizadas en esta Tesis se detallan en la tabla 1. Las esponjas marinas que utilizadas en esta Tesis fueron *Polymastia littoralis*, PMLT01 y *Lithoplocamia lithistoides*.

Cepa	Genotipo/fenotipo relevante	Referencias
Escherichia coli		
E. coli DH10B	F$^-$, *mcr*A, *Δ(mrr hsd*RMS-*mcr*BC*)*, *Φ80lac*ZΔM15, *Δlac*X74, *deo*R, *rec*A1, *ara*D139, *Δ(ara-leu)*7697, *gal*U, *gal*K, *λ$^-$*, *rps*L (Str^R), *end*A1, *nup*G .	Invitrogen
E. coli Replicator FOS	F$^-$, *mcr*A, Δ(*mrr-hsd*RMS-m*cr*BC), *end*A1, *rec*A1, Φ80d*lac*ZΔM15, Δ*lac*X74 *ara*D139, Δ(*ara,leu*)7697, *gal*U, *gal*K, *rps*L (Str^R), *nup*G (*att*L *ara*C-P$_{BAD}$-*trf*A250 *bla att*R), λ$^-$.	CopyRight® v2.0 Fosmid Cloning Kits, Lucigen
E. coli F1B6	Cepa Replicator FOS con el fósmido 1B6.	Esta Tesis
E. coli F1W5	Cepa Replicator FOS con el fósmido 1W5.	Esta Tesis
E. coli F4P7	Cepa Replicator FOS con el fósmido 4P7.	Este trabajo
E. coli F3H4	Cepa Replicator FOS con el fósmido 3H4.	Esta Tesis
E. coli BAC-Optimized Replicator v2.0	F$^-$, *mcr*A, Δ(*mrr-hsd*RMS-m*cr*BC), *end*A1, *rec*A1, Φ80d*lac*ZΔM15, Δ*lac*X74 *ara*D139, Δ(*ara,leu*)7697, *gal*U, *gal*K, *rps*L (Str^R), *nup*G (*att*L *ara*C-P$_{BAD}$-*trf*A250 *bla att*R), λ$^-$.	CopyRight® v2.0 BAC Cloning Kits, Lucigen.
E. coli B13A10	Cepa BAC-Optimized Replicator v2.0 con el BAC 13A10.	Esta Tesis
E. coli B13A10PPT	Cepa *E. coli* B13A10 con el plásmido pSEVAPPT.	Esta Tesis
Tistrella mobilis		
T. mobilis MES-10-09-028	Cepa silvestre aislada del gusano *Sabellastarte* sp.	PharmaMar
T. mobilis KR3	Cepa MES-10-09-028 con una mutación de inactivación del dominio KR del módulo 3 del clúster de síntesis de didemninas.	Esta Tesis
T. mobilis DidA	Cepa MES-10-09-028 con una deleción de los módulos 1 y 2 del gen *ddnA* y el dominio de condensación del módulo 3.	Esta Tesis
T. mobilis DidAKR3	Cepa MES-10-09-028 doble mutante con las mutaciones KR3 y DidA.	Esta Tesis

Tabla 1 | Cepas bacterianas empleadas en esta Tesis.

2. Medios y condiciones de cultivo

Todas las soluciones y medios de cultivo utilizados en esta Tesis se esterilizaron por calor húmedo en autoclave a 121 °C y 1 atm de presión, o mediante filtración utilizando filtros estériles Millipore de 0,2 µm de diámetro. A menos que se indique lo contrario, las cepas de *E. coli* y *T. mobilis* fueron cultivadas en medio *Lysogeny Broth* (LB) (Sambrook y Russell, 2001) en agitación a 250 rpm a 37 °C y 30 °C respectivamente. A aquellos cultivos que se realizaron en medio sólido se les añadió al medio Bacto Agar (Pronadisa) al 1,5% (p/v). El crecimiento de los microorganismos en cultivo líquido se monitorizó midiendo la densidad óptica a 600 nm (A_{600}) utilizando un espectrofotómetro Shimadzu UVmini-1240. Cuando fue necesario, se añadió a los cultivos 5-bromo-4-cloro-3-indolil-β-D-galactopiranósido (X-Gal) 0,08 mM, así como isopropil-β-D-1-tiogalactopiranósido (IPTG) 0,1 mM. Para seleccionar las cepas bacterianas, se utilizaron por defecto los antibióticos apropiados a las siguientes concentraciones kanamicina (Km) (50 µg/mL), gentamicina (Gm) (5 µg/mL), ampicilina (Ap) (100 µg/mL) y cloranfenicol (Cm) (12,5 µg/mL), los cuales fueron preparados en soluciones 1000 veces concentradas en agua, excepto el Cm que se disolvió en etanol 100%.

El medio mínimo empleado para cultivar las cepas de *T. mobilis* fue el medio MC. Este medio está compuesto por el medio basal MA (pH 7,5) el cual una vez suplementado con vitaminas y elementos traza se denomina MC. La composición del medio MC para 1 L se detalla a continuación:

- Medio MA:
  - KH_2PO_4 0,33 g
  - Na_2HPO_4 1,20 g
  - NH_4Cl 0,11 g
  - $MgSO_4 \cdot 7H_2O$ 0,10 g
  - $CaCl_2$ 0,04 g
  - H_2O destilada hasta 1 L

- Suplemento de elementos traza (1000x) (pH 6,5):
  - Ácido nitrilotriacético (NTA) 1,50 g
  - $ZnSO_4 \cdot 7H_2O$ 0,18 g
  - $MgSO_4 \cdot 7H_2O$ 3,00 g
  - $CuSO_4 \cdot 5H_2O$ 0,01 g
  - $MnSO_4 \cdot 2H_2O$ 0,50 g
  - $KAl(SO_4)_2 \cdot 12H_2O$ 0,02 g
  - NaCl 1,00 g
  - H_3BO_3 0,01 g
  - $FeSO_4 \cdot 7H_2O$ 0,10 g
  - $CoSO_4 \cdot 7H_2O$ 0,18 g
  - $NiCl_2 \cdot 6H_2O$ 0,03 g
  - $NaSeO_3 \cdot 5H_2O$ 0,30 g
  - H_2O destilada hasta 1 L

- Suplemento de vitaminas (1000x):
  - Biotina 20 mg
  - Ácido fólico 20 mg
  - Piridoxina-HCl 10 mg
  - Tiamina-HCl·2H2O 50 mg
  - Riboflavina 50 mg
  - Ácido nicotínico 50 mg
  - D-pantotenato cálcico 50 mg
  - Vitamina B12 50 mg
  - Ácido *p*-aminobenzoico 50 mg
  - H2O destilada hasta 1 L

Durante periodos inferiores a un mes las cepas se conservaron a 4 °C en placas de cultivo. Para conservarlas a largo plazo se congelaron en el medio de cultivo correspondiente con glicerol al 15% (v/v) y se almacenaron a -80 °C.

3. Vectores

En la tabla 2 se describe una relación de los plásmidos utilizados en esta Tesis junto con sus características más relevantes.

Plásmidos	Características relevantes	Referencias
pSEVA224	Vector replicativo de la colección SEVA con origen de replicación RK2 y el sistema de expresión *lacI*^{q}-P_{TRC}. Sm^{R}.	Silva-Rocha *et al.*, 2013
pSEVA424	Vector replicativo de la colección SEVA con origen de replicación RK2 y el sistema de expresión *lacI*^{q}-P_{TRC}. Km^{R}.	Silva-Rocha *et al.*, 2013
pK18*mob*sacB	Vector suicida. Gm^{R}.	Schäfer *et al.*, 1994
pK18KR3	Vector derivado de pK18*mob*sacB con el inserto necesario para provocar la mutación KR3 en el clúster productor de didemninas.	Este trabajo
pk18DidA	Vector derivado de pK18*mob*sacB con los insertos necesarios para provocar la mutación DidA en el clúster productor de didemninas.	Esta Tesis
pSEVAPPT	Vector derivado de pSEVA224 que contiene el gen codificante de la PPTasa putativa de *T. mobilis* MES-10-09-028.	Esta Tesis
pSMART® FOS	Fósmido replicativo en *E. coli*	Lucigen

	con número de copias inducible por arabinosa. Cm^R.	
pSMART® BAC	BAC replicativo en *E. coli* con número de copias inducible por arabinosa. Cm^R.	Lucigen
BAC 13A10	BAC derivado del pSMART® BAC, que contiene el fragmento 13A10.	Esta Tesis
Fosmido 1B6	Fósmido derivado del pSMART® FOS, que contiene el fragmento 1B6.	Esta Tesis
Fosmido 1W5	Fósmido derivado del pSMART® FOS, que contiene el fragmento 1W5.	Esta Tesis
Fosmido 4P7	Fósmido derivado del pSMART® FOS, que contienecon el fragmento 4P7.	Esta Tesis
Fosmido 3H4	Fósmido derivado del pSMART® FOS, que contiene el fragmento 3H4.	Esta Tesis

Tabla 2 | Vectores utilizados en esta Tesis Doctoral.

4. Técnicas de manipulación de DNA

La mayor parte de las técnicas básicas de biología molecular utilizadas en esta Tesis, así como la manipulación del DNA fueron realizadas esencialmente tal y como se describe en Sambrook y Rusell (2001). Las endonucleasas de restricción empleadas en esta Tesis fueron suministradas por New England Biolabs y Takara. La DNA ligasa T4 fue suministrada también por New England Biolabs. Todas las enzimas se emplearon atendiendo a las especificaciones de las distintas casas comerciales.

4.1. | Electroforesis en geles de agarosa

Con el fin de visualizar los fragmentos de DNA se utilizaron geles de agarosa al 0,7% o al 1,5% en tampón TAE (Tris-HCl 40 mM, ácido acético 20 mM, EDTA 2 mM, pH 8,1), utilizando el mismo tampón como electrolito. Como tampón de carga a las muestras se les añadió ¼ de su volumen de una solución compuesta por Ficoll 400 al 30% (p/v), azul de bromofenol al 0,2% (p/v), xilencianol al 0,2% (p/v) y EDTA 40 mM (pH 8,0). El proceso de electroforesis se llevó a cabo a 100 V durante 15-20 min y posteriormente los geles se tiñeron con GEL-RED para poder visualizar los fragmentos de DNA con radiación ultravioleta en un transiluminador. Los marcadores de tamaño utilizados fueron el DNA del fago λ digerido con la endonucleasa de restricción *Bst*EII (Amersham) y la forma replicativa del fago φX174 digerida con *Hae*III (New England Biolabs).

4.2. | Electroforesis de campo pulsado

Para visualizar y separar fragmentos de DNA de alto peso molecular se realizó una electroforesis en campo pulsante. Para ello se utilizaron geles de agarosa al 1% en tampón TBE (Tris-HCl 75 mM, ácido bórico 25 mM, EDTA 0,1 mM, pH 8,0). La electroforesis se

llevó a cabo en el mismo tampón a 200 V con una rampa de pulsos de duración ascendente (de 10 a 200 s) durante 30 h a 14 °C, utilizando un equipo CHEF-DRII (Bio-Rad). Los marcadores de peso molecular utilizados fueron el Yeast Chromosome PFG marker y el Lambda Ladder PFG Marker, ambos proporcionados por NEB.

4.3. | Amplificación mediante PCR

Para llevar a cabo la amplificación de fragmentos de DNA se emplearon las enzimas DNA *Taq* polimerasa o la DNA polimerasa PFU (ambas suministradas por Biotools B. M. Labs) de acuerdo con las instrucciones del fabricante. A menos que se indique lo contrario las mezclas de reacción utilizadas en la PCR contenían dNTPs 0,4 mM, dimetilsulfóxido (DMSO) 10%, 1 unidad de DNA polimerasa, 100 ng de DNA molde y cebadores a una concentración final de 0,5 μM. Los cebadores utilizados para las amplificaciones mediante PCR fueron sintetizados por Sigma y se describen en la tabla 3.

Cebadores	Secuencia	Referencia
DidBKR3AF	CCCAAGCTTGTTCCTGGAACCGGACGATGCC	Amplificación del fragmento 5' para la construcción del mutante KR3.
DidBKR3AR	AACTGCAGGGCAAAGGCCGCGGTACCG	
DidBKR3BF	AACTGCAGCGGTACCGCGGCCTTTGCC	Amplificación del fragmento 3' para la construcción del mutante KR3.
DidBKR3BR	GCTCTAGACTTTCGGTGCCCGGAAGCGC	
DidBKRScr	GCATTGGCGGCGGCAA	Amplificación de un fragmento interno de 815 bp para comprobar la mutación KR3.
DidBAF	CCCAAGCTTGATATCGGCGCGGGATTTGG	Amplificación del fragmento 5' para la construcción del mutante DidA.
DidBAR	GGGATTCCATATGCAGTCTTCATTCCGCTGG	
DidBBF	GGGATTCCATATGGCACTGGCCGATGGCGGAC	Amplificación del fragmento 3' para la construcción del mutante DidA.
DidBBR	GCTCTAGACCAGCGGATCGAAGACAGGTC	
DidACompF	GGGTCAGCAGGTCGAGTTGTG	Cebadores externos de comprobación de la deleción DidA.
DidACompR	GTCTGCGGTGAAGTCGATGC	
SMD13F	GTCAGCAGATGGTCAATCACGG	Amplificación de fragmentos del clúster productor de didemninas.
SMD13R	CAGAAGACGAGCTGGCACCG	
SMD8F	CCAGCCAGTCCTCCCAGTCTTC	Amplificación de fragmentos del clúster productor de didemninas.
SMD8R	CCGCCAAGATCGTTTCGCAG	

SMD3F	GAACAGCTCGGCGCGATAGTC	Amplificación de fragmentos del clúster productor de didemninas.
SMD3R	CACCAACACCCTGCCGCTC	
GAP1F	CCGCGACCTCAGGGATCTC	Amplificación de las zonas del gap 1 en la secuencia de *T. mobilis* MES-10-09-028 .
GAP1R	CAAGCCTGCGGAGATCGTC	
GAP2F	CTGATCCTGGACGGCATGG	Cebadores para la amplificación de las zonas del gap 1 en la secuencia de *T. mobilis* MES-10-09-028.
GAP2R	GAGCAGATAGGCGAGGTCGTC	
GAP3F	CTGGTGATCGACGCGGTG	Cebadores para la amplificación de las zonas del gap 1 en la secuencia de *T. mobilis* MES-10-09-028.
GAP3R	GATCGACGATCAGCGCCAC	
GAP4F	CTGGTGGCGCTGATCAACC	Cebadores para la amplificación de las zonas del gap 1 en la secuencia de *T. mobilis* MES-10-09-028.
GAP4R	CCGTCTCCAGCCCCTGATC	
GAP5F	CACCTCTCCGCTGATCGATCC	Cebadores para la amplificación de las zonas del gap 1 en la secuencia de *T. mobilis* MES-10-09-028.
GAP5R	GTCCAGGCCACCGCTTCG	
GAPOUT1F	CCACGAGGTTGAAGGAAGGG	Cebadores para la amplificación de las zonas del gap out1 en la secuencia de *T. mobilis* MES-10-09-028.
GAPOUT1R	CACCAGAATGATCAGCGCATAG	
SCDidAF	CCTGTTCGCCACCATGCAG	Amplificación de un fragmento interno del gen *ddnA* (201 bp) para análisis mediante PCR semicuantitativa
SCDidAR	GTGCGGCTCGAAAGACAGG	
SCDidBF	CCAATGGGTCGACCGTCTG	Amplificación de un fragmento interno del gen *ddnB* (216 bp) para análisis mediante PCR semicuantitativa
SCDidBR	GTCGCCCTGATCGGTGAAC	
SCDidCF	GCCGAGGTGATGCAGGATC	Amplificación de un fragmento en el extremo 5' del gen *ddnC* (244 bp) para análisis mediante PCR semicuantitativa
SCDidCR	CCGATCAGGATCTGGGCG	
SCDidDF	GGGCGAGGACTATCTGCTGC	Amplificación de un fragmento interno del gen

SCDidDR	CAGTTCCTTTTCGGCCGC	*ddnC* (204 bp) para análisis mediante PCR semicuantitativa
SCDidEF	GTATCGACGATCCGCGCTG	Amplificación de un fragmento interno del gen *ddnD* (206 bp) para análisis mediante PCR semicuantitativa
SCDidER	CATAGACGGCGGTGCGG	
SCDidFINF	CGAGACCCCGCATGAAAGC	Amplificación de un fragmento interno del gen *ddnH* (200 bp) para análisis mediante PCR semicuantitativa
SCDidFINR	CATGTTCGGGCGGGAAGG	
TistPPTF	GCTCTAGACAGTAATACAAGGGGTGTTATGCC GAACTCGCCGCC	Amplificación del gen de la PPTasa putativa de *T. mobilis* MES-10-09-028
TistPPTR	CCCAAGCTTGGATGATGTTCGTGATGATCTCGC	

Tabla 3 | Cebadores utilizados en esta Tesis Doctoral.

En ocasiones, para comprobar si las células de *E. coli* que han sido transformadas contienen el DNA de interés, se realizaron amplificaciones mediante PCR utilizando células enteras procedentes de una colonia aislada como molde. Para ello, se resuspendió una pequeña cantidad de biomasa en la mezcla de reacción, llevándose a cabo posteriormente el protocolo de amplificación habitual.

4.4. | Aislamiento y purificación de fragmentos de DNA

De forma general, los fragmentos de DNA, incluidos los productos de las reacciones de PCR, fueron purificados con el kit *Gene-Clean Turbo* (Q-BIO-gene) o llustra GFX 96 PCR *Purification* Kit (GE Healthcare). Además, la extracción y purificación del DNA plasmídico de *E. coli* y *T. mobilis* se realizó utilizando el kit *High Pure Plasmid Isolation kit* (Roche).

4.4.1. | Extracción del DNA cromosómico de *Tistrella mobilis*

La extracción del DNA cromosómico de *T. mobilis* se realizó a partir de la biomasa obtenida tras un cultivo de 15 mL durante toda la noche en medio LB. Tras obtener un sedimento bacteriano mediante centrifugación del cultivo, las células se resuspendieron en 400 µL de TE (Tris-HCl 10 mM, EDTA 1 mM, pH 7,5) y se incubaron a 80 °C durante 20 min. A continuación se dejó enfriar a temperatura ambiente y se añadió 50 µL de lisozima (10 mg mL^{-1}), se mezcló con vórtex y se incubó 10 min a 37 °C. Seguidamente se añadieron 100 µL de NaCl 5 M y 100 µL de CTAB/NaCl precalentados a 65 °C y se incubó durante 10 min a 65 °C. A continuación se añadieron 750 µL de cloroformo/alcohol isoamílico (24:1), se mezcló suavemente por inversión y se centrifugó 5 min a 14000 x *g*. La solución acuosa se tomó con la pipeta, se transfirió a un tubo eppendorf y se le añadió un volumen equivalente de fenol/cloroformo/isoamílico (25:24:1), se mezcló suavemente por inversión y se centrifugó durante 5 min a 14000 x *g*. La solución acuosa se tomón con una pipeta y se transfirió a un tubo eppendorf donde el DNA se precipitó añadiendo 0,6 volúmenes de isopropanol y manteniéndolo 30 min a 20 °C. Finalmente el DNA se centrifugó 15 min a 14000 x *g* y a 4 °C, se lavó con etanol al 70%, se centrifugó 2 min a

14000 x *g* y se dejó secar. Posteriormente el DNA se resuspendió en 40 – 200 µL de TE o de agua en el caso de que este se destinase a procesos de secuenciación masiva.

4.4.2. | Aislamiento y extracción de DNA metagenómico de esponjas marinas

4.4.2.1. | Aislamiento de los microorganismos contenidos en las muestras de esponja

Se partió de muestras de esponjas congeladas a -80 °C, se descongelaron a temperatura ambiente y se trocearon en fragmentos cúbicos aproximados de 1 cm^3. Estos fragmentos se procesaron en un homogeneizador mecánico (*potter*) añadiendo tampón TE (Tris-HCl 10 mM, EDTA 1 mM, pH 7,5) (aproximadamente 4 mL por 1 g de esponja). Una vez homogenizado el tejido de la esponja se centrifugó 2 min a 200 g para eliminar arenas y restos del tejido de la esponja. Una vez recuperado el sobrenadante se filtró utilizando filtros de membrana de nitrocelulosa de 8 µm tamaño de poro (Millipore) por el cual pasan las bacterias pero se eliminan microrganismos más voluminosos presentes en la muestra como pueden ser algunas microalgas. A continuación se centrifugó a 4000 x *g* durante 30 min a 4 °C para concentrar la muestra que contenía fundamentalmente la fracción bacteriana.

Debido a las distintas características morfológicas y tisulares de las esponjas procesadas en esta Tesis, el protocolo de extracción utilizado se modificó para cada uno de los casos particulares. En concreto, para realizar la extracción microbiana en PMLT01 se tuvo en cuenta, por lo que se incluyeron en el procesamiento muestras de todas las partes diferenciables. Sin embargo, en el proceso de extracción microbiana en *Lithoplocamia*, debido a la viscosidad que presentaba la fracción con las células microbianas tras la homogenización mecánica, fue inviable realizar los pasos de filtrado por lo que directamente se extrajo el DNA del conjunto total de las células contenidas en el sedimento tras centrifugar el extracto.

4.4.2.2. | Extracción del DNA metagenómico de los microorganismos aislados

Para lisar las células microbianas extraídas se utilizó un tampón de lisis con la siguiente composición:

- Urea 8 M
- Sarkosyl 2%
- NaCl 1 M
- EDTA 50 mM
- Tris-HCl (pH = 7,5) 50 mM

En primer lugar, se resuspendieron los sedimentos microbianos en el tampón con una relación de 10 mL/g de esponja utilizados. A continuación se incubó la mezcla 10-20 min a 60 ℃. Seguidamente se realizó una extracción con el mismo volumen de una mezcla con fenol/cloroformo/alcohol isoamílico y se agitó la muestra suavemente. Se centrifugó la mezcla a 4000 x *g* durante 5-10 min y se extrajo la fase superior acuosa. El paso de fenolización seguido de la centrifugación se realizó al menos tres veces. Posteriormente, se añadió un volumen equivalente de cloroformo, se volvió a centrifugar a 4000 x *g* durante 5-10 min y se separó la fase superior acuosa resultante. Se añadió 1/10 de volumen de

acetato sódico 3 M pH 7,0 y 2 volúmenes de isopropanol y para precipitar el DNA. Se mezcló suavemente por inversión y se almacenó un mínimo de 15 min a -20 °C. A continuación se centrifugó a máxima velocidad durante 30 min y se retiró el sobrenadante con precaución. Se realizó un lavado con un volumen de etanol al 70% y se centrifugó a 4000 x *g* durante 10-15 min. Posteriormente se eliminó el sobrenadante y se dejó secar al aire. Finalmente el DNA se resuspendió con 50-300 µL de tampón TE (o agua si se destinó a algún proceso de secuenciación masiva).

En el caso específico de la extracción del DNA de la esponja *Lithoplocamia* y debido a la viscosidad de la muestra ésta se ultracentrifugó a 400000 x *g* a 4°C durante 1 h del DNA en solución.

4.4.2.3. | Extracción del DNA cromosómico de *T. mobilis* de alto peso molecular mediante células embebidas en bloques de agarosa

Con el objetivo de obtener DNA cromosómico de gran tamaño para utilizarlo en la generación de genotecas en BACs, fue necesario utilizar protocolos en los que se evitara la rotura excesiva del DNA cromosómico. Para ello se realizó la lisis de las células previamente embebidas en bloques de agarosa tal y como se detalla a continuación. Se cultivaron 10 mL de células durante toda la noche en medio LB. Una vez centrifugado el cultivo, las células se resuspendieron en 0,5 mL de tampón Cell Suspensión que contenía: Tris-HCl 10 mM, NaCl 20 mM, EDTA 100 mM (pH 7,2). La mezcla se calentó a 42 °C y se añadió 0,5 mL de agarosa al 1,5% preparada en agua. La mezcla se añadió en una jeringa de 1 mL donde se dejó enfriar durante 15 min. Una vez solidificada la muestra se sacó en bloque de la jeringa y con ayuda de un bisturí se dividió en tres partes y se sumergieron en una solución que contenía lisozima (1 mg/mL), Tris-HCl 10 mM (pH 7,2), NaCl 50 mM, EDTA 100 mM, 0,2% de DOC y 0,5% de N-laurylsarcosine y se incubaron a 37 °C durante 2 h. Transcurrido ese tiempo la solución de lisozima se retiró mediante aspiración y los bloques de agarosa se lavaron 3 veces durante 15 min con tamón de lavado que contenía: Tris-HCl 20 mM, EDTA 50 mM (pH 8,0). Posteriormente, los bloques se sumergieron en una solución que contenía proteinasa K (1 mg/mL), EDTA 100 mM, 1% N-laurylsarcosine, 0.2% DOC y se incubaron a 42 °C durante toda la noche. Transcurrido ese tiempo la solución de proteinasa K se retiró mediante aspiración y los bloques de agarosa se lavaron 1 h con tampón de lavado que contenía PMSF 1 mM para inactivar la proteinasa K. Después, se lavó otras dos veces con tampón de lavado y se conservó en tampón de conservación Tris-HCl 2 mM con EDTA 5 mM, pH 8,0 hasta su utilización.

4.5. | Secuenciación del DNA

La secuenciación estándar del DNA fue llevada a cabo por el Servicio de Secuenciación Automática de DNA (SSAD) del Centro de Investigaciones Biológicas (CIB) (Secugen S.L.) mediante el uso de un secuenciador automático ABI Prism 3730 (Applied Biosystems). Para la reacción de la secuenciación se utilizó el *Dye Terminator Cycle Sequencing Ready Reaction Kit* de Applied Biosystems, así como la DNA polimerasa AmpliTaq FS, de acuerdo con las especificaciones del fabricante. Las reacciones de amplificación se llevaron a cabo mediante la técnica de PCR con un termociclador *Mastercycler® gradient* (Eppendorf) o *Veriti®* 384 (Applied Biosystems).

5. Técnicas de manipulación de RNA

5.1. | Extracción de RNA bacteriano

Para realizar la extracción de RNA se partió de 15 mL de los cultivos con una DO_{600} 0,6-0,8 para *E. coli* y 1.5-1.9 para *T. mobilis*. A continuación se utilizó el kit *RNAeasy* (Qiagen) siguiendo las instrucciones recomendadas por el fabricante. Las células se centrifugaron 10 min a 4000 x *g* a 4 °C, se resuspendieron en una solución de tampón TE con lisozima 50 mg/mL y a continuación se incubo la mezcla a temperatura ambiente durante 5 min. Seguidamente se realizaron 2-3 ciclos de congelación a -80 °C y descongelación. Posteriormente, se añadieron 700 µL de tampón RLT (Qiagen) con β-mercaptoetanol (10 µL de β-mercaptoetanol por cada mL de RLT) y se homogenizó en el agitador. Se añadieron 500 µL de etanol al 100% enfriado a -20 °C y se homogenizó con la pipeta. A continuación se realizaron pases por la columna del kit de 700 en 700 µL y se realizó un lavado con 700 µL del tampón RWI y 500 µL del RPE. Las muestras se eluyeron con 50 µL de agua destilada dos veces. Para eliminar el DNA en las muestras se realizaron dos tratamientos

A continuación se realizó a cada muestra dos tratamientos con DNasa I del kit RNase-free (Ambión) siguiendo las especificaciones del fabricante comprobándose la ausencia de DNA por PCR. La concentración de RNA extraído se midió utilizando un NanoPhotometer®Pearl (Implen, GmbH) y mediante electroforesis en gel de agarosa. Las muestras de RNA se conservaron a -80 °C.

5.2. | Retrotranscripción del RNA seguida de PCR (RT-PCR)

Una vez purificado el RNA, se obtuvo el DNA complementario (cDNA) mediante una reacción de retrotranscripción mediada por la enzima transcriptasa reversa SuperScript II (Invitrogen). Cada reacción de retrotranscripción (20 µL) contenía 1 µg de RNA, 200 U de transcriptasa reversa, ditiotreitol 10 mM, dNTPs 0,5 mM y 5 mM de pd(N)6 *random hexamer 5´phosphate* (Amersham Biosciences). Los hexámeros se utilizaron como cebadores, de este modo, se incubaron junto con el RNA a 65 °C durante 5 min para permitir la hibridación. Tras enfriar en hielo, se añadieron los componentes restantes de la reacción, la cual se incubó a 42 °C durante 2 h. La reacción se finalizó con una incubación de 15 min a 70 °C. De forma paralela se realizaron controles en ausencia de retrotranscriptasa para cada una de las muestras procesadas. El cDNA obtenido se utilizó como molde para la PCR posterior en la que los cebadores requeridos se añadieron a una concentración final de 0,5 µM junto con 1 U de DNA polimerasa I (Biotools). La amplificación sobre el cDNA indicó la presencia de expresión de la cepa estudiada en las condiciones concretas.

6. Transferencia de DNA a estirpes bacterianas:

6.1. | Transformación de las células de *Escherichia coli*

Las células de *E. coli* fueron modificadas genéticamente por transformación tras hacerlas competentes mediante el método de RbCl (Sambrook y Rusell, 2001), o bien mediante electroporación (Wirth *et al*., 1989). Las condiciones del equipo de electroporación Gene Pulser/Pulse Controller (Bio-Rad) fueron 2,5 kV, 25 µF, y 200 Ω.

6.2. | Transformación de células de *T. mobilis* MES-10-09-028

La preparación de células competentes de *T. mobilis* MES-10-09-028, se realizó a partir de un preinóculo en medio líquido LB inoculado con una única colonia y cultivado durante toda la noche a 30 °C. Este preinóculo se utilizó para inocular 1 L de medio LB a una densidad óptica inicial de 0,050. El cultivo se recogió a una DO_{600} entre 0,6 y 0,9. A continuación se centrifugó a 3000 x *g* a 4 °C durante 10 min y se realizaron 4 lavados con agua destilada estéril enfriada en hielo en volúmenes decrecientes de 1 L, 0,5 L, 0,25 L y 0,1 L durante 10 min a 3000 x *g*. En todos los pasos las células se resuspendieron suavemente volteando el recipiente, nunca con agitación, ni pipeteando. Finalmente se retira el sobrenadante y las células se resuspendieron en 2 mL de agua destilada que se repartieron en varias alícuotas de 250 µL. Las alícuotas se utilizaron inmediatamente y no se almacenaron congeladas en ninguna ocasión debido a las bajas eficiencias de transformación tras la descongelación. Las condiciones del equipo de electroporación Gene Pulser/Pulse Controller (Bio-Rad) utilizadas para las cepas de *T. mobilis* fueron idénticas que las de *E. coli*, es decir, 2,5 kV, 25 µF, y 200 Ω.

6.3. | Construcción de mutantes mediante doble recombinación homóloga en *T. mobilis*

La construcción de las cepas mutantes KR3, DidA y KR3DidA se llevó a cabo mediante mutación por doble recombinación homóloga utilizando el plásmido suicida pK18*mobsacB* (Schäfer *et al*., 1994) (Fig. 12).

Para la construcción de la cepa *T. mobilis* KR3 se llevó a cabo una amplificación mediante PCR de dos fragmentos solapantes en la región que se va a mutar de unas 800 pb uno en posición 5´ con respecto al gen que se va a mutar (cebadores DidBKR3AF y DidBKR3AR) y otro en posición 3´ (cebadores DidBKR3BF y DidBKR3BR) utilizando el DNA genómico de *T. mobilis* como molde. Los oligonucleótidos DidBKR3AR, y DidBKR3BF introducen dos mutaciones, una que va a generar la sustitución de un codón que codifica tirosina por otro que codifica fenilalanina y otra mutación silenciosa que genera una diana *Kpn*I. Los fragmentos generados se digirieron con las enzimas de restricción correspondientes *Hin*dIII-*Kpn*I, y *Kpn*I-*Xba*I, respectivamente. A continuación ligaron los fragmentos digeridos se clonaron en el vector pK18*mobsacB* previamente digerido haciendo uso de los sitios de restricción únicos *Hin*dIII y *Xba*I generando el plásmido pK18KR3 que se transformó en células de *T. mobilis* MES-10-09-028 mediante electroporación. Los transformantes (que han sufrido una recombinación homóloga y tienen el vector integrado en su cromosoma) se seleccionaron en placas de LB con kanamicina. Para forzar una segunda recombinación se tomaron colonias aisladas y se cultivaron en medio LB con kanamicina hasta una DO_{600} aproximada de 0,6-0,9, momento en el cual se sembraron 10 µL en una placa de MC glucosa 20 mM y otros 10 µL en una placa de MC-glucosa 20 mM con 5% de sacarosa. Las colonias obtenidas se cultivaron en placas de LB con kanamicina para confirmar la pérdida de la resistencia al antibiótico y por tanto del plásmido debido a una recombinación homóloga doble. Los mutantes así obtenidos se analizaron por PCR empleando los cebadores DidBKR3AF y DidBKRScr. El cebador DidBKRScr posee en su extremo 3' la versión de la secuencia con la mutación, de modo que solo existiría amplificación en el caso de que la colonia poseyese la mutación.

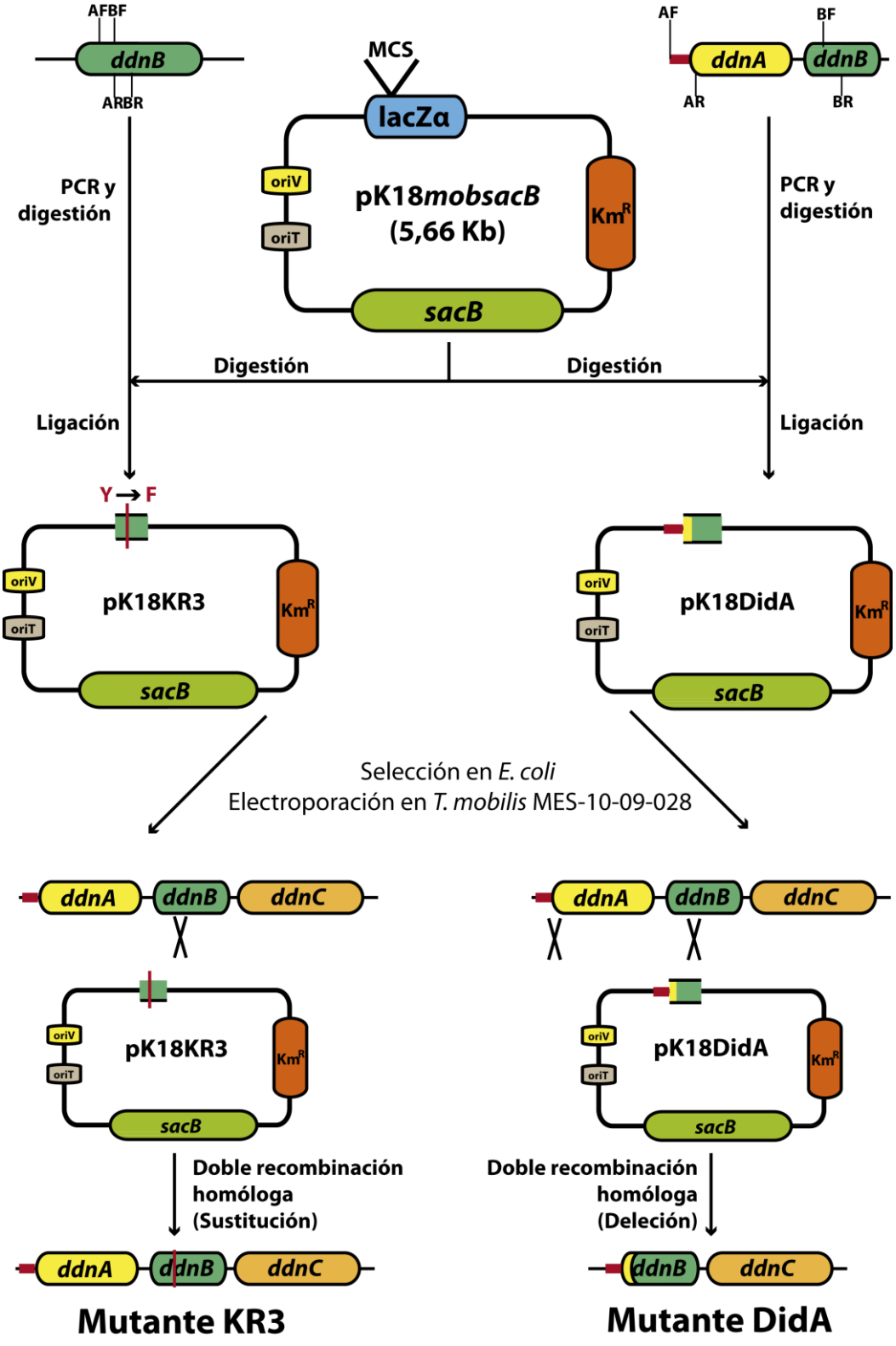

Figura 12 | Esquema de la generación de los mutantes KR3 y DidA de *T. mobilis* MES-10-09-028 mediante doble recombinación homóloga. A la izquierda se muestra el proceso de generación del mutante KR3 representándose mediante una línea vertical roja la mutación puntual generada (sustitución de tirosina por fenilalanina). A la derecha se detalla la generación del mutante DidA, márcandose en rojo la zona de la secuencia del promotor del gen *ddnA*. En la parte superior se detallan las zonas de amplificación de las zonas A y B para cada uno de los casos, siendo los cebadores AF y AR, y BF y BR respectivamente.

Para realizar la deleción del gen DidA sobre la cepa silvestre *T. mobilis* MES-10-09-028 y *T. mobilis* KR3 se siguió la misma estrategia. En primer lugar se amplificaron dos fragmentos de aproximadamente 800 pb uno en posición 5´con respecto al gen a mutar (cebadores DidBAF y DidBAR) y otro en posición 3´ (cebadores DidBBF y DidBBR) utilizando el DNA genómico de *T. mobilis* como molde. Los fragmentos generados se purificaron y se digirieron con las correspondientes enzimas de restricción *Hin*dIII y *Nde*I, y *Nde*I- *Xba*I respectivamente A continuación se ligaron ambos fragmentos digeridos y el fragmento resultante se clonó en el vector pK18*mobsacB* previamente digerido con *Xba*I y *Hin*dIII. La construcción generada se denominó pk18DidA y se transformó en células de *T. mobilis* MES-10-09-028 y de *T. mobilis* KR3 para posteriormente llevar a cabo el proceso de doble recombinación tal y como se ha descrito previamente. Al generarse los mutantes *T. mobilis* KR3 y *T. mobilis* KR3DidA, se realizaron comprobaciones de la mutación mediante amplificaciones con los cebadores DidACompF y DidACompR externos al gen DidA en ambos mutantes.

7. Generación de genotecas del DNA de *T. mobilis* MES-10-09-028 de en vectores de alta capacidad de almacenaje

Partiendo de la extracción del DNA de *T. mobilis* MES-10-09-028 se construyeron genotecas en vectores de alta capacidad de almacenaje.

7.1. | Construcción de genotecas de *T. mobilis* MES-10-09-028 en fósmidos

Para generar la genoteca DNA de *T. mobilis* MES-10-09-028 en fósmidos se utilizó el kit CopyRight v2.0 Fosmid Cloning (Lucigen) el cual contiene el vector pSMART FOS (Fig. 13) según las especificaciones del fabricante. El empaquetamiento en las cabezas de fagos se realizó utilizando el kit Gigapack III Gold Packaging Extract de Agilent Technologies siguiendo las especificaciones del fabricante. Para titular la genoteca se tomó una parte del empaquetado para infectar células EPI300-T1R, cultivadas en LB con $MgSO_4$ 10 mM y 0,2% de maltosa, según se indica en el protocolo del kit *CopyControlTM Fosmid Library Production* (Epicentre), conservando el resto a 4 °C el menor tiempo posible (unos días). Para la infección se mezclaron 100 µL de células EPI300-T1R preparadas a D.O.600 de 0,8-1 y 10 µL de diluciones del empaquetado con factores de dilución 10, 10^1, 10^2, 10^3, 10^4 y 10^5, y se incubaron 1 h a 37 °C. Las células infectadas se sembraron en placas de LB con Cm (12,5 µg/mL) y se incubaron a 37 °C toda la noche, para seleccionar células con fósmido. Se contaron las colonias y se estimó el número de clones que tendría la genoteca si se realizara una infección con el resto del empaquetado. La genoteca obtenida fue de aproximadamente 17000 clones.

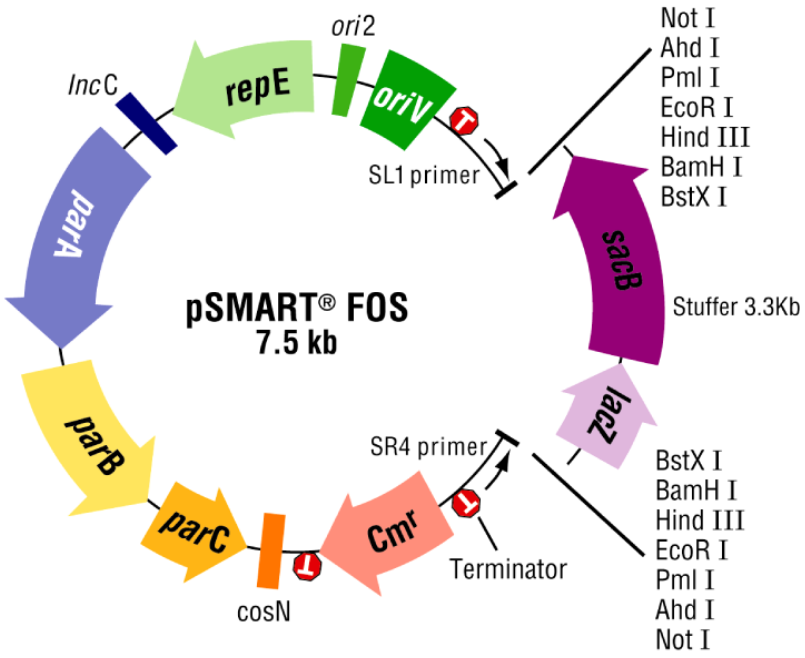

Figura 13 | Esquema del vector pSMART® FOS. Figura extraída del manual del fabricante (Lucigen).

7.2. | Construcción de genotecas de genotecas de *T. mobilis* MES-10-09-028 en BACs

Para obtener fragmentos de mayor tamaño en la genoteca se utilizaron BACs (*bacterial artifitial chromosomes*) como vectores. En este caso es muy importante la obtención de un DNA de alto peso molecular (HMW) para que no existan fragmentos pequeños de DNA genómico que den lugar a una genoteca de BACs en la que predominen aquellos vectores con poco tamaño de inserto.

7.2.1. | Digestión con enzimas de restricción del DNA genómico en bloques de agarosa

Una vez lisadas las células embebidas en la agarosa tal y como se describe en el apartado 4.4.2.3. con el fin de evitar la lisis del DNA, se procedió a realizar una digestión parcial con la enzima de restricción *Bam*HI. Cada una de las condiciones de digestión se realizó por triplicado a partir de cortes finos de los bloques de agarosa (100 µL). Se colocaron en tubos de 2 mL y se dializaron durante 1 h en 200 µL de un tampón que contenía: tampón de restricción 1X proporcionado por el fabricante, espermidina 1 mM y DTT 0,5 mM. A continuación se añadieron 5 unidades de *Bam*HI durante 8 min a 37 °C tiempo tras el cual se paró la reacción con 17 µL de EDTA 0,5 M (pH 8,0).

7.2.2. | Preparación del DNA y generación de la genoteca de BACs

Una vez digerido el DNA se realizó una electroforesis en campo pulsado para separar los fragmentos obtenidos según su tamaño. El gel se tiñó con Gel Red para visualizar el DNA y se cortaron los fragmentos que tenían un tamaño entre 100 y 300 kb. Para extraer el DNA de los bloques de agarosa se introdujeron en membranas de diálisis (MWCO 15000) *Spectra/Por Dyalisis Membrane* (Spectrum Laboratories Inc.) lavadas con agua y 0,5X TBE, se rellenaron con TBE 0.5X y se sellaron con pinzas. Se colocaron en la

cubeta del equipo de campo pulsado y se aplicaron 120 V con una rampa de pulsos (de 35 a 35 s) durante 4 h y media a 12,5 °C. El último minuto se giran las bolsas de diálisis 180° para despegar el DNA de las paredes de la bolsa. La concentración del DNA extraído se midió utilizando el Nanodrop y se visualizó mediante una electroforesis convencional en gel de agarosa al 0,7%.

Posteriormente, los fragmentos obtenidos se clonaron en los vectores pSMART® BAC (Fig. 14) utilizando el kit *CopyRight® v2.0 BAC Cloning* (Lucigen) según las especificaciones del fabricante. La cepa de *E. coli BAC-Optimized Replicator* v2.0 se transformó con el total de las ligaciones mediante electroporación generándose dos genotecas de aproximadamente 3440 y 830 clones, respectivamente.

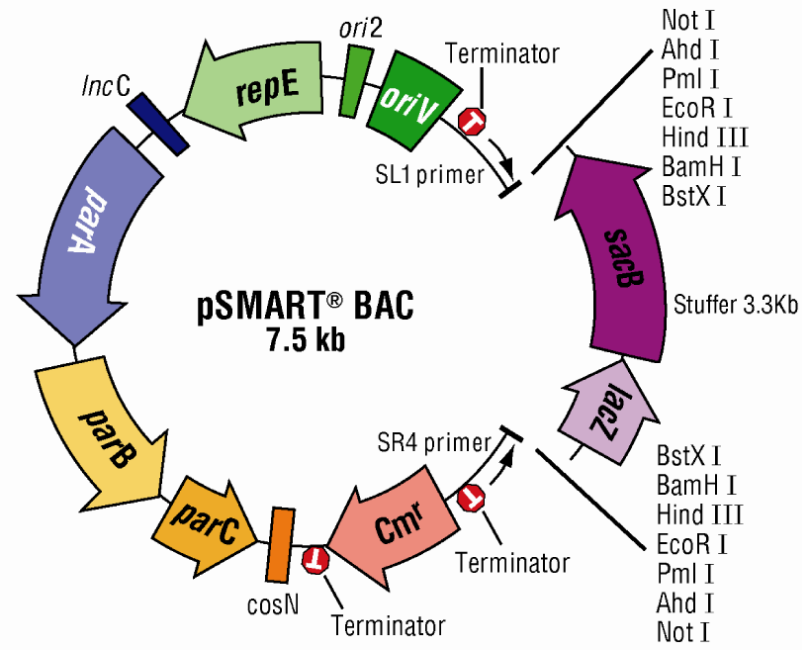

Figura 14 | Esquema del vector pSMART® BAC. Figura extraída del manual del fabricante (Lucigen).

7.3. | Cribado de clones en las genotecas para la búsqueda de fragmentos pertenecientes al clúster productor de didemninas

El cribado de los clones se realizó mediante amplificación por PCR de las colonias. Para ello se diseñaron varias parejas de cebadores que amplificaban regiones pertenecientes al clúster productor de didemninas (ver tabla 3). De este modo para cada ronda de cribado se analizaron 200 clones en grupos de 10 clones. Una vez se detectaba un grupo positivo se volvía a realizar PCR para cada una de las 10 colonias contenidas en dicho grupo con el fin de identificar las colonias positivas.

8. Técnicas cromatográficas

8.1. | Extracción de didemninas de caldos de cultivo

Todos los procesos de extracción de didemninas realizados en esta Tesis fueron llevados a cabo por PharmaMar. Para analizar la producción de didemninas en las cepas de

T. mobilis se realizaron cultivos de 40 mL de medio de producción LB en matraz a 30 °C durante aproximadamente 7 d. En ocasiones se aumentó el volumen a 500 mL con el fin de detectar niveles mayores de didemninas. A continuación, en algunos casos en los que se analizó el caldo de cultivo y las células por separado, para lo cual se separó el sedimento del sobrenadante mediante centrifugación a 4000 x *g* durante 15 min a 4 °C. En los demás casos se procesó la muestra completa.

Para realizar la extracción de las didemninas de las células tras la centrifugación se realizó una extracción 1:1 v/v con una mezcla 2:3 de isopropanol/acetato de etilo y se homogenizó utilizando una *ULTRA TURRAX*®. A continuación se centrifugó a 4000 x *g* durante 10 min y se dejó evaporar en sequedad. Seguidamente se filtró por filtros de placa porosa 3 añadiendo un lecho de celite. En el caso de las muestras del sobrenadante, se añadieron 1:1 v/v de acetato de etilo y se dejó agitar vigorosamente durante 30 min. A continuación se dejó en reposo para separar las fases y se almacenó toda la noche -20 °C. La fase orgánica (no congelada) se filtró por placa porosa 3 con lecho de celite para eliminar el hielo (fase acuosa). Finalmente en ambas variantes se dejó evaporar la fase orgánica en *speed-vac*. La extracción de las muestras completas de cultivo se realizó en primer lugar liofilizando y posteriormente extrayendo con 1:2 v/v de isopropanol/acetato de etilo/agua. Todos los extractos se resuspendieron en 200-300 µL de metanol para su posterior análisis.

8.2. | Cromatografía líquida de alta eficiencia acoplada a espectrometría de masas (HPLC-DAD-MS)

Todos los análisis de presencia de didemninas mediante HPLC-DAD-MS realizados en esta Tesis fueron llevados a cabo por PharmaMar. La separación cromatográfica se llevó a cabo mediante una columna Symmetry C18 RP 150 x 4,6 mm 5 µm (Waters). La detección se llevo a 215 y 254 nm utilizando un equipo DAD y LC/MSD Agilent Serie 1100. La ionización se realizó mediante isoelectronebulizador (ES con polaridad positiva). En cada análisis se inyectó un volumen de 100 µL con un flujo de 0,8 mL/min y se realizó un gradiente entre una fase A que contenía acetonitrilo con 0,4% de trifluoroacético y una B con agua Milli Q con 0,4% de trifluoroacético. La relación de porcentajes de gradiente en el tiempo se puede observar en la tabla 4. El tiempo de análisis utilizado fue de 45 min.

Tiempo (min)	A%	B%
0	5	95
5	5	95
35	100	0
40	100	0
42	5	95
45	5	95

Tabla 4 | Relación en porcentaje de las soluciones A y B durante el tiempo de análisis mediante HPLC.

9. Secuenciación masiva de DNA

Todos los procesos de secuenciación masiva realizados en esta Tesis fueron llevados a cabo por la empresa Life Sequencing S.L. (Paterna, España).

El DNA genómico extraído de *T. mobilis* MES-10-09-028 fue secuenciado utilizando el kit Titanium y el equipamiento GS-FLX de pirosecuenciación generándose librerías con adaptadores *Paired End* (Roche), según las instrucciones del fabricante y en total se generaron 2 carreras independientes. El DNA metagenómico extraído de la esponja *P. littoralis* fue secuenciado en una única carrera utilizando también el kit Titanium y el equipamiento GS-FLX de pirosecuenciación (Roche).

El DNA metagenómico de las esponjas PMLT01 y *Polymastia* fue secuenciado utilizando el chip de 316 y la química de 400 bp de la plataforma Ion Torrent PGM según las instrucciones del fabricante.

10. Herramientas y procedimientos bioinformáticos

A la hora de realizar búsquedas de secuencia por comparación se utilizaron los algoritmos BLAST (incluyendo todas sus variantes) en las distintas bases de datos que presenta el servidor del *National Center for Biotechnology Information* (NCBI), (http://www.ncbi.nlm.nih.gov/cgibin/Entrez/genom_table_cgi) y HMMER3 (http://hmmer.janelia.org/) (Finn *et al*., 2011). Para detectar dominios funcionales conservados se realizaron búsquedas en la base de datos Pfam (http://pfam.xfam.org/) (Finn *et al*., 2010) y para conocer las implicaciones metabólicas de ciertas secuencias se acudió a la base de datos KEGG (http://www.genome.jp/kegg/). A la hora de realizar alineamientos múltiples se utilizó MUSCLE (Edgard, 2004) con parámetros por defecto dentro del programa MEGA6 (Tamura *et al*., 2013). Todos los análisis taxonómicos de genes 16S RNA así como la posterior generación de árboles filogenéticos fueron llevados a cabo en la base de datos RDP (https://rdp.cme.msu.edu/) (Cole *et al*., 2006) con sus herramientas correspondientes. Para obtener los valores de ANIb y ANIm se utilizó el software JSpecies (Richter y Rosselló-Móra, 2006).

10.1. | Ensamblaje y anotación de la secuencia de *T. mobilis*

El ensamblaje de las lecturas generadas en la secuenciación masiva del DNA genómico de *T. mobilis* MES-10-09-028 fue llevado a cabo con el software Newbler v2.7 (454 Life Sciences) utilizando los datos de las librerías *Paired End* y con los parámetros establecidos por defecto. El conjunto de *scaffolds* resultantes se anotó de forma general y automática utilizando el servidor RAST (Aziz *et al*., 2008). La anotación específica de los putativos clústeres productores de metabolitos secundarios en la secuencia se realizó utilizando la herramienta AntiSmash en su versión 1.0 (Medema *et al*., 2011) y posteriormente se revisó con la versión 2.0 (Blin *et al*., 2013). La anotación específica del clúster productor de didemninas y de su entorno se revisó de forma manual con el fin de corregir errores en la secuencia y en la anotación automática, para ello se revisaron todas ORFs en busca de posibles cambios en la pauta abierta de lectura que provocasen defectos en la anotación estructural y funcional. En las zonas donde se comprobó la existencia de un posible error, este se trató de corregir utilizando otras versiones de la secuencia sin errores contenidas las en el conjunto de las lecturas.

10.2. | Ensamblaje y anotación de la secuencia obtenida de los metagenomas de esponja

La anotación y el análisis parcial de las lecturas pertenecientes al metagenoma de cada una de las esponjas se llevó a cabo utilizando la herramienta MG-RAST (Glass y Meyer, 2011).

El ensamblaje de las lecturas obtenidas fue llevado a cabo utilizando distintos programas ensambladores (Mira (Chevreux *et al*., 1999), CLC *Genomics Workbench* (CLC Bio) y Newbler (Roche)) importando sus correspondientes archivos en formato FASTQ. Las versiones dadas por el software Newbler con sus parámetros establecidos por defecto fueron las utilizadas finalmente (ver Resultados).

El análisis y la anotación funcional general de los genomas individuales extraídos de los metagenomas fue realizado, utilizando la herramienta RAST (Aziz *et al*., 2008).

10.3. | Análisis bioinformático de metagenomas

El proceso de extracción de datos, de procesamiento y de clasificación de los fragmentos ensamblados (Fig. 15) de cada uno de los metagenomas fue llevado a cabo basándose en los protocolos generados en Albertsen *et al*. (2013) y posteriormente revisados en las guías y tutoriales dispuestos por el mismo autor (http://madsalbertsen.github.io/multi-metagenome/ y http://madsalbertsen.github.io/mmgenome/). Los programas de Perl y R utilizados entre los que se encuentra el paquete de R *mmgenome*, fueron en todos los casos obtenidos tanto de estos recursos web como de sus correspondientes repositorios en GitHub (https://github.com/MadsAlbertsen).

Una vez extraídos y parcialmente procesados, los datos se terminaron de procesar y se visualizaron utilizando el entorno de R, RStudio (https://www.rstudio.com/).

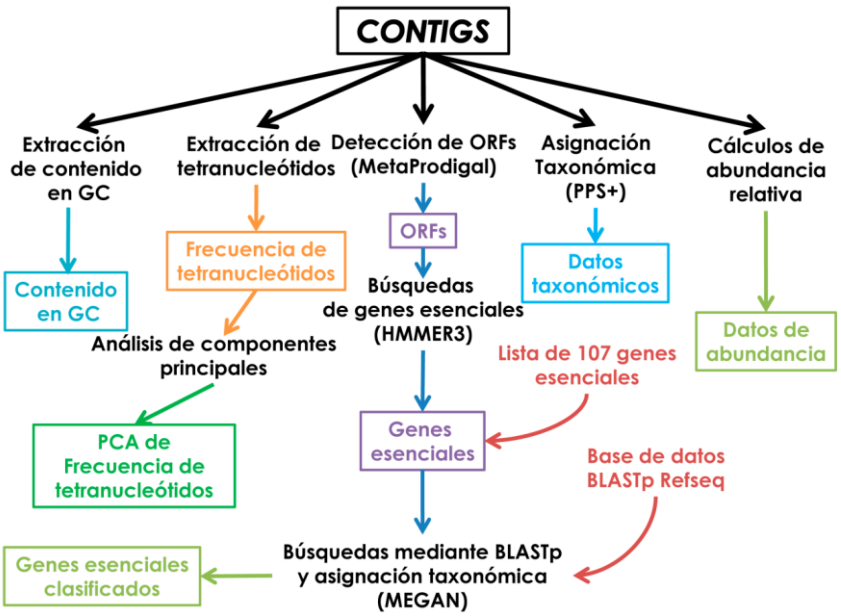

Figura 15 | Diagrama de flujo del tratamiento *in silico* de las secuencias ensambladas. Los pasos correspondientes al procesamiento de los datos se representan en negro mientras que los datos generados en sí se representan en color y con un recuadro. Entre paréntesis se muestra el software utilizado.

10.3.1. | Abundancia de los fragmentos ensamblados

Para realizar el cálculo de los niveles de abundancia de cada uno de los contigs generados con respecto al conjunto total de las lecturas, en primer lugar se mapearon las lecturas con un mínimo de similitud del 95% sobre el 100% de la longitud frente a los contigs utilizando el software bowtie2 (Langmead y Salzberg, 2012) y se generó un archivo en formato SAM. A continuación utilizando las herramientas contenidas en el paquete samtools, se transformó archivo en formato BAM, se ordenó y se obtuvieron los datos de profundidad, los cuales se procesaron mediante la script de perl *calc.coverage.in.bam.depth.pl* para obtener un archivo CSV utilizable posteriormente en R.

10.3.2. | Cálculo del contenido en GC y de la frecuencia de tetranucleótidos

El cálculo del contenido en GC y de la frecuencia de tetranucleótidos de cada uno de los *contigs* ensamblados se realizó mediante la función *mmload* en R, contenida en el paquete *mmgenome*. El análisis estadístico de componentes principales en la frecuencia de tetranucleótidos fue llevado a cabo también utilizando dicha función.

10.3.3. | Identificación de los genes marcadores conservados

Las ORFs en la secuencia ensamblada fueron predichas utilizando la versión de Prodigal para metagenomas (Hyatt *et al.,* 2012). Se realizaron búsquedas locales mediante la herramienta HMMER 3 (http://hmmer.janelia.org/) (Eddy, 2011) contra una recopilación de 107 modelos de tipo HMM de genes esenciales monocopia (Dupont *et al.,* 2012) con los parámetros establecidos por defecto. Las proteínas identificadas fueron clasificadas taxonómicamente realizando BLASTP contra la base de datos de proteínas RefSeq (versión 70) con un valor máximo del valor esperado (*e-value*) de $1xe^{-5}$. Posteriormente se utilizó la herramienta MEGAN (Huson *et al*., 2011) para extraer los niveles taxonómicos de las asignaciones mediante BLAST y se exportaron a un archivo XML.

10.3.4. | Asignación taxonómica total de la secuencia ensamblada

El total de la secuencia ensamblada fue clasificado taxonómicamente utilizando la herramienta PhyloPythiaS+ (Gregor et al., 2014) con todos los parámetros establecidos por defecto. El archivo de salida fue utilizado posteriormente en el proceso de agrupamiento de los datos en R.

10.3.5. | Representaciones de la distribución espacial de la secuencia ensamblada y extracción de genomas individuales

Una vez importados los datos extraídos, cada una de las secuencias ensambladas fue representada gráficamente utilizando la función *mmplot*, según las características extraídas previamente como son contenido en GC, en tetranucleótidos y su cobertura relativa. Para realizar la extracción espacial de un conjunto de secuencias a un archivo formato FASTA se utilizaron las funciones *mmplot_locator* y *mmextract* y el análisis de la presencia de genes esenciales según la taxonomía se realizó con la función *mmref*, todas dentro del paquete *mmgenome*.

10.4. | Búsqueda de genes del metabolismo de PKSs y NRPSs en las secuencias metagenómicas de esponja

10.4.1. | Búsqueda de dominios relacionados con clústeres PKS y NRPS

La búsqueda de dominios relacionados con clústeres PKS y NRPS se realizó utilizando la herramienta HMMER3 (http://hmmer.janelia.org/) (Eddy, 2011). Para ello se realizaron las búsquedas tanto en la secuencia ensamblada como en el conjunto de las lecturas seleccionando distintos modelos de dominios comúnmente encontrados en este tipo de clústeres (ver tabla 5). Para reducir parcialmente la redundancia de los positivos en las lecturas se *clusterizaron* los resultados obtenidos con la herramienta CD-HIT (Li y Godzik, 2006) permitiendo un 98% de identidad total.

Dominio	Nombre Pfam	Número de acceso
Acil- Transferasa	Acyl_transf_1	PF00698.16
Unión a AMP	AMP-binding	PF00501.23
Unión a AMP C-terminal	AMP-binding_C	PF13193.1
Condensación	Condensation	PF00668.15
Cétido-sintasa N -terminal	ketoacyl-synt	PF00109.21
Cétido-sintasa C -terminal	Ketoacyl-synt_C	PF02801.17
Ceto-reductasa	KR	PF08659.5
Sitio de unión de Fosfopanteteina	PP-binding	PF00550.20
Dehidratasa	PS-DH	PF14765.1
Tioesterasa	TE	PF00975.15

Tabla 5 | Perfiles de dominios utilizados en las búsquedas de genes que codifican PKSs y NRPSs

10.4.2. | Curación manual de las secuencias ensambladas con dominios relacionados con PKSs y NRPSs

Se realizó un cribado manual de todas aquellas secuencias ensambladas realizando búsquedas en la base de datos Pfam y haciendo BLAST frente a la base de datos del NCBI. Solo se seleccionaron aquellos fragmentos que posiblemente podían pertenecer a clústeres PKS/NRPS complejos. A continuación se llevó a cabo un ensamblaje manual de dichos fragmentos seleccionados realizando BLAST con las secuencias de los extremos frente a la base de datos de lecturas correspondiente.

10.4.3. | Clasificación taxonómica y representación gráfica de secuencias ensambladas con dominios relacionados con PKSs y NRPSs.

Las secuencias fueron clasificadas taxonómicamente por comparación realizando BLAST frente a la base de datos del NCBI. Se asignó manualmente aquel taxón que más se repetía entre los 5 mejores resultados, priorizando aquellos con mejor puntuación. Para obtener una distribución espacial gráfica de los *contigs* seleccionados según sus características, se realizó un protocolo de asignación similar al utilizado para marcar los genes esenciales en las secuencias metagenómicas, generándose de este modo un archivo XML con la misma estructura, y de este modo se introdujeron los datos correspondientes en el entorno de R. La estructura de dominios de aquellos *contigs* que podrían pertenecer al mismo clúster se comparó con las propuestas de síntesis generadas.

10.5. | Ensamblaje y análisis de la secuencia mitocondrial de *P. littoralis*

La secuencia mitocondrial de *P. littoralis* se ensambló de forma manual partiendo de la secuencia de los *contigs* generados en el ensamblaje del metagenoma de la esponja. Se realizó un proceso de anotación estructural y funcional automática utilizando la herramienta especializada MITOS (http://mitos.bioinf.uni-leipzig.de/index.py) (Bernt *et al.,* 2013). A continuación se revisó manualmente cada una de las ORFs generadas y se realizó una anotación posterior también manual corrigiendo los posibles errores en la secuencia mediante comparación con las lecturas.

VI. Resultados

1. Utilización de herramientas genómicas para la identificación, caracterización y expresión heteróloga del clúster de síntesis de didemninas contenido en *Tistrella mobilis* MES-10-09-028

1.1. | Secuenciación del genoma de *T. mobilis* MES-10-09-028 y análisis de clúster de síntesis de didemninas

1.1.1. | Detección de la producción de didemninas

La cepa *T. mobilis* MES-10-09-028 fue aislada por PharmaMar en el proceso de descubrimiento de nuevas moléculas antitumorales de ambientes marinos (Schleissner et al., 2013). En concreto *T. mobilis* MES-10-09-028 se aisló del gusano *Sabellastarte* sp., un poliqueto marino recolectado en el Océano Índico en el que se detectó la producción de didemnina B, nordidemnina B, didemnina X y didemnina Y.

Para comprobar que la cepa recibida en el laboratorio era productora de didemninas se realizaron distintos ensayos de crecimiento y producción en distintos medios de cultivo. Como resultado de estos ensayos se observó crecimiento de la cepa MES-10-09-028 en placa Petri en medio rico LB y en medio mínimo MC a 30 °C con glucosa o sacarosa como fuentes de carbono.

A continuación se analizó la producción de didemninas en medio rico LB. Para ello se llevaron a cabo cultivos en matraz a 30 °C en agitación (250 rpm) y se tomaron muestras a distintos tiempos (24, 48, 72 y 96 h). Se analizó la presencia de didemninas en la fracción celular y en el sobrenadante y se detectaron los picos pertenecientes a las distintas didemninas mediante HPLC-MS (Fig. 16) (ver Materiales y Métodos apartado 8). Se identificaron los picos pertenecientes a la didemnina B, didemnina X, didemnina Y y en menor cantidad la nordidemnina B en la fracción celular mientras que en el medio extracelular, solo fue posible detectar didemnina B. Si se tienen en cuenta el área de los picos para determinar las concentraciones relativas de los metabolitos, se deduce que las concentraciones de didemnina X e Y van disminuyendo en el interior celular a la vez que la de didemnina B aumenta hasta las 72 h de cultivo (Fig 17).

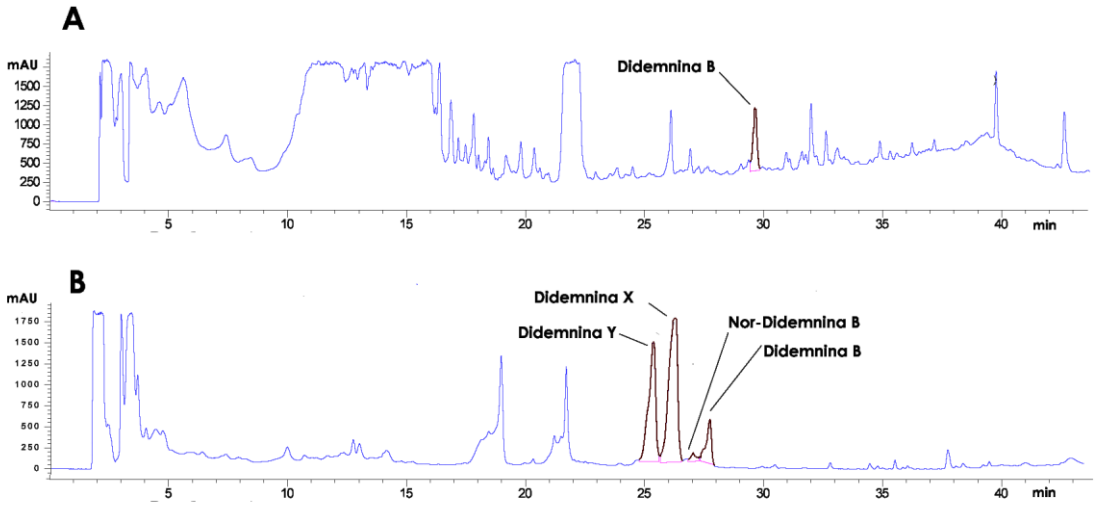

Figura 16 | Análisis mediante HPLC-DAD de la producción de didemninas en *T. mobilis* MES-10-09-028. Los extractos celulares (A) y sobrenadantes de cultivo (B) se analizaron tras 24 h de cultivo. Los picos correspondientes a didemninas aparecen señalados en color rojo y fueron detectados a 215 nm.

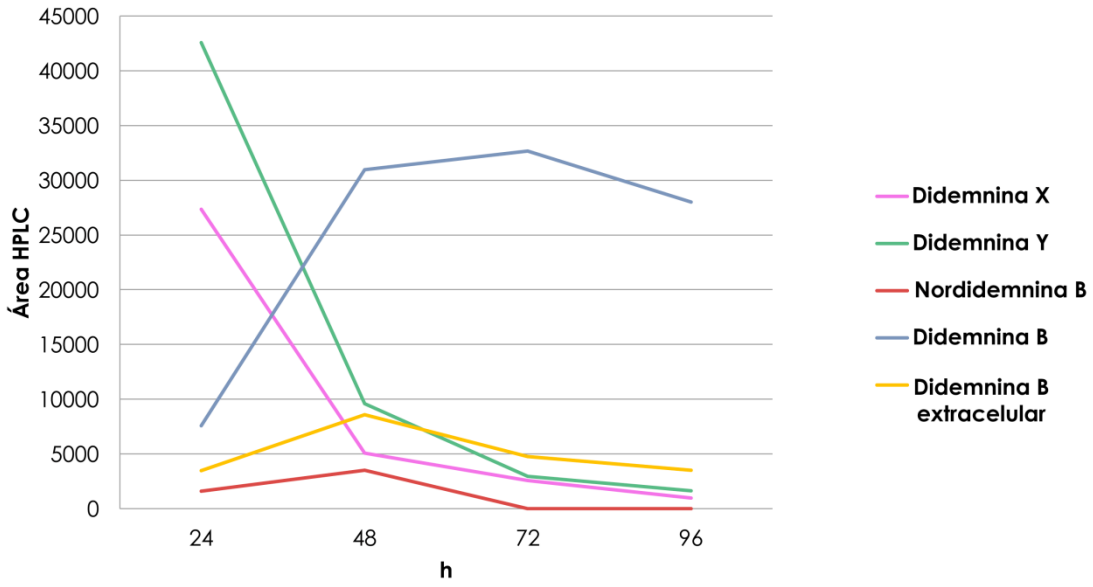

Figura 17 | Curvas de producción de didemninas en cultivos de *T. mobilis* MES-10-09-028 en medio LB.

1.1.2. | Secuenciación y ensamblaje del genoma de *T. mobilis* MES-10-09-028

Para obtener la secuencia del genoma de *T. mobilis* MES-10-09-028, se llevó a cabo en primer lugar un proceso de pirosecuenciación mediante la estrategia de *Paired End* (PE). La secuencia obtenida de la librería PE generada, estaba distribuida en un total de 185 394 lecturas con una longitud media de 275 bp, abarcando la secuencia un total de 51 Mb. A continuación se procedió a ensamblar *de novo* la secuencia obtenida resultando 4651 *contigs* de los cuales 3390 tenían una longitud mayor de 500 bp. Tras utilizar los datos obtenidos de los extremos pareados, estos *contigs* se ordenaron en 186 *scaffolds* con una longitud total de 8,8 Mb. El gran número de fragmentos obtenidos en este borrador

del genoma hizo replantear la estrategia de secuenciación considerándose necesaria la adición de una segunda tanda de secuencia al proceso. Esta secuencia se obtuvo al aplicar el proceso de pirosecuenciación a una librería generada mediante una aproximación *shotgun* tradicional (ver Materiales y Métodos). De este modo se obtuvieron un total de 210 030 lecturas con una longitud media de 377 bp dando lugar a un total de 79,35 Mb.

Partiendo de ambos conjuntos de secuencias (130 Mb totales) se logró ensamblar *de novo* el conjunto de las lecturas en un total de 291 *contigs* de los cuales 216 tienen una longitud de más de 500 bp. La media de tamaño de los *contigs* se situó en 20920 bp teniendo el de mayor tamaño 318346 bp. La secuencia fue ordenada utilizando los datos provenientes de la estrategia *Paired End* dando lugar a un boceto preliminar del genoma de *T. mobilis* MES-10-09-028 que consta de 6 *scaffolds* los cuales abarcan un total de 6 170 371 bp de secuencia.

1.1.3. | Clasificación taxonómica de la cepa *T. mobilis* MES-10-09-028

En estudios anteriores realizados en la empresa PharmaMar sobre este aislado, se realizó su clasificación taxonómica mediante la secuenciación de los amplicones generados por PCR que cubrían parcialmente la secuencia del gen 16S rRNA. Tras estos estudios la cepa fue clasificada de forma preliminar dentro del género *Tistrella* y de la especie *mobilis* (Schleissner *et al*., 2013).

Una vez obtenida la secuencia genómica, se obtuvo la secuencia completa del gen 16S rRNA de la cepa constando ésta de 1460 bp. Utilizando esta secuencia se realizaron comparaciones mediante las herramientas incluidas en la base de datos RDP (Ribosomal Database Project) pudiendo confirmar que la cepa es del género *Tistrella,* un género bacteriano englobado dentro de la familia *Rhodospirillaceae,* en el orden *Rhodospirillales*, en la clase *Alphaproteobacteria* y en el phylum *Proteobacteria*. A continuación se realizó un análisis filogenético involucrando las especies tipo más cercanas a la cepa dentro del orden Rhodospirillales (Fig. 18). Mediante este análisis *T. mobilis* MES-10-09-028 se situaría en un nodo relativamente alejado del resto de géneros contemplados. Se observó entonces la relación filogenética de la cepa con todas las cepas del género *Tistrella* cuyas secuencias de 16S rRNA han sido depositadas en la base de datos RDP. De este modo se observó (árbol filogenético no mostrado) que (al igual que se muestra en la Fig. 18), la secuencia no solo resulta idéntica a la secuencia parcial del aislado de la cepa tipo de *T. mobilis* (AB071665) (Shi *et al*., 2002) sino también a otros aislados ambientales. Este resultado sugiere la pertenencia del aislado de PharmaMar a la especie *mobilis*.

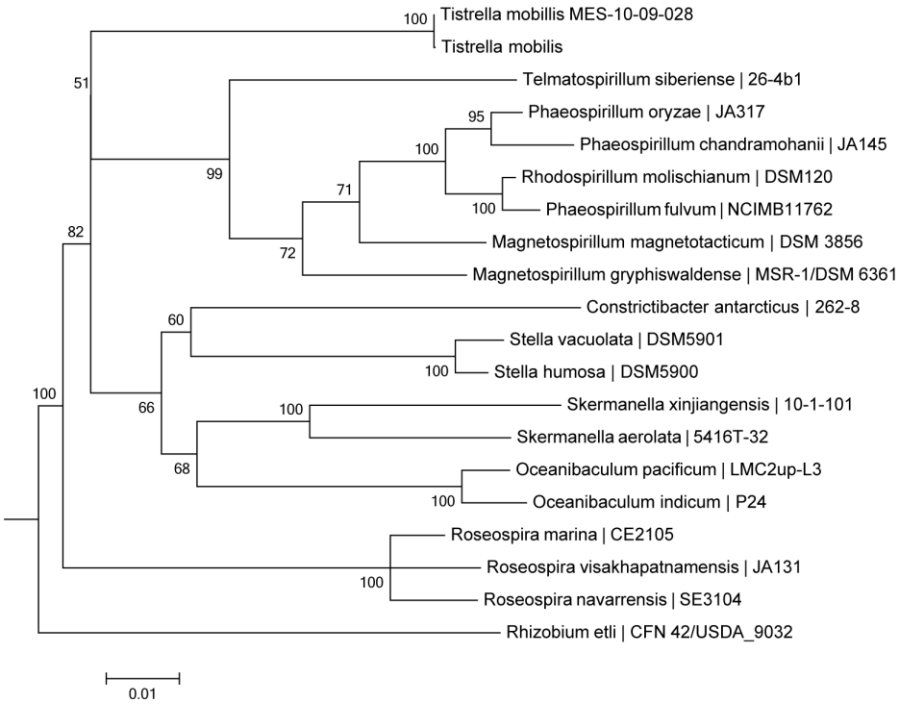

Figura 18 | Arbol filogenético de especies tipo del orden *Rhodospirillales* incluyendo *T. mobilis* MES-10-09-028. El árbol fue construido utilizando la herramienta correspondiente en la base de datos RPD con los parámetros establecidos por defecto. El árbol incluye las 20 secuencias tipo del orden *Rhodospirillales* más cercanas al gen del 16S de la cepa MES-10-09-028 incluida dicha secuencia. En los nodos se señala el valor de *boostrap* correspondiente.

La única cepa cuyo genoma se encuentra secuenciado y depositado en las bases de datos que pertenece al género *Tistrella* es *T. mobilis* KA081020-065 (Xu *et al*., 2012). Esta cepa posee un genoma superior en 300 kb (6,5 Mb) en comparación con la secuencia obtenida para *T. mobillis* MES-10-09-028. Para comprobar si estas cepas, que comparten un 16S rRNA similar, son igualmente similares a un nivel genómico, se obtuvieron los valores de ANIb y ANIm (Richter y Rosselló-Móra, 2009), siendo estos de un ANIb de 95,65 (con un 88,69% alineado) y un ANIm de 95,96 (con un 91,70% alineado). Estos valores indican que ambas cepas están muy cercanas evolutivamente y que tal y como sus secuencias de 16S rRNA señalaban, se encuentran en el umbral de poder ser consideradas dos cepas distintas de la misma especie.

1.1.4. | Anotación del genoma de *T. mobillis* MES-10-09-028

El conjunto de 6 *scaffolds* resultante del proceso de ensamblaje se introdujo en el servicio de anotación automática RAST (Aziz *et al*., 2008). El resultado se visualizó con SEED (*theseed.org*) (Overbeek *et al*., 2014) obteniéndose un total de 49 RNAs y 5430 secuencias codificantes en las cuales se representaban 491 de los subsistemas definidos por SEED. Se clasificaron un total de 2674 secuencias (aproximadamente un 50%), conteniendo 171 proteínas hipotéticas, dentro de estos subsistemas; mientras que 2756 de entre las cuales 1272 fueron anotadas como proteínas hipotéticas quedaron fuera de esta clasificación. Las secuencias codificantes de *T. mobillis* MES-10-09-028 representaron 2293 del conjunto de las funciones y características contenidas dentro de estos subsistemas (Fig. 19), incluyéndose las posibles redundancias de funciones encontradas.

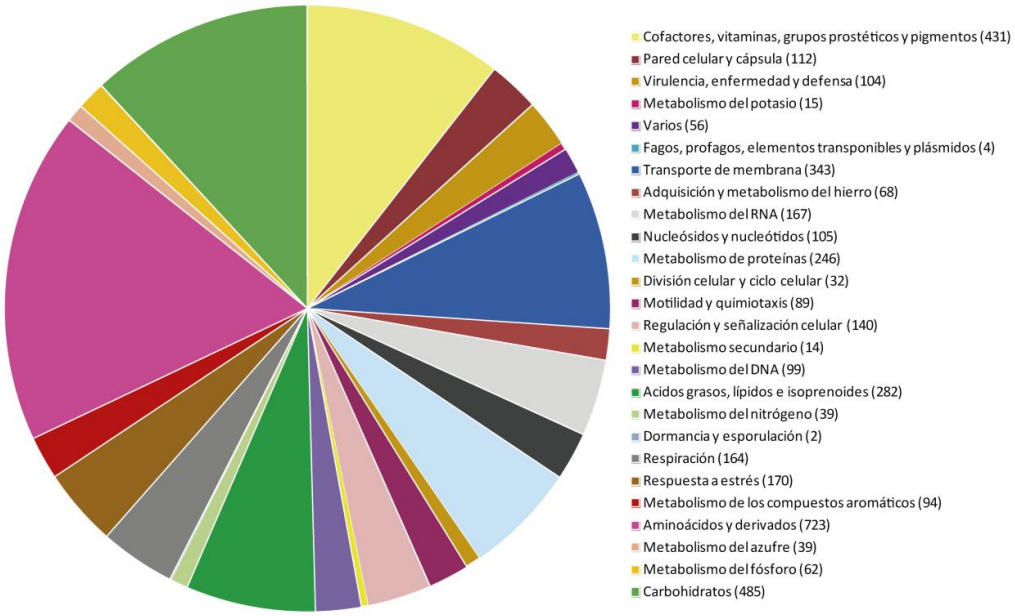

Figura 19 | Gráfica de la distribución por subsistemas de la anotación de *T. mobillis* MES-10-09-028. En el gráfico proporcionado por la herramienta RAST se representan el número de positivos para cada uno de los subsistemas agrupados en categorías generales tras el proceso de anotación automática.

1.1.5. | Análisis del clúster génico responsable de la síntesis de didemninas

Para identificar el clúster génico responsable de la producción de la didemnina B, se realizó un análisis pormenorizado de cada uno de los monómeros que forman la molécula con el fin de predecir la estructura de dominios en cada uno de los módulos responsables de la síntesis. Basándose en la estructura química de la didemnina B (ver Introducción, figura 9) se asumió la hipótesis de que esta molécula fuera producida por un clúster génico de estructura modular híbrido entre PKS y NRPS. En concreto, si se observa la molécula en la figura 9 ésta muestra una circularización probablemente generada al liberarse la molécula del complejo multiproteico de ensamblaje. Este proceso podría ser llevado a cabo por un dominio tioesterasa en la posición final. Por lo tanto, asumiendo que la arquitectura del clúster génico sea secuencial, el monómero de lactato que se encuentra en uno de los extremos de la secuencia sería el primero en añadirse a la molécula. Este monómero podría incorporarse mediante una NRPS cuyo dominio de adenilación añadiese directamente la molécula de lactato. Otra opción posible es que el módulo podría estar activando una molécula de piruvato que sería posteriormente modificada por un dominio ceto reductasa (KR), incorporándose así la molécula lactato a la didemnina B (Fig. 20). Dado que no existen ejemplos claros de dominios de adenilación que tengan este tipo de afinidades, no resulta fácil predecir la incorporación del residuo inicial.

A continuación es posible detectar en la estructura de la molécula de didemnina varios monómeros que resultan de la incorporación de aminoácidos o de sus análogos metilados. Estas adiciones, posiblemente son llevadas a cabo por módulos que contienen NRPSs. En concreto, el segundo módulo añadiría un residuo de prolina, el tercero uno de N-metil-leucina seguidos de residuos de treonina e isoleucina incorporados por los módulos cuarto y quinto, respectivamente. Seguidamente se encuentran monómeros de isostatina y ácido α-(α-hidroxi-isovaleril)-propiónico, lo que sugiere extensiones cétidas

de aminoácidos que podrían ser llevadas a cabo por algún tipo de combinación entre módulos pertenecientes a PKSs y NRPSs.

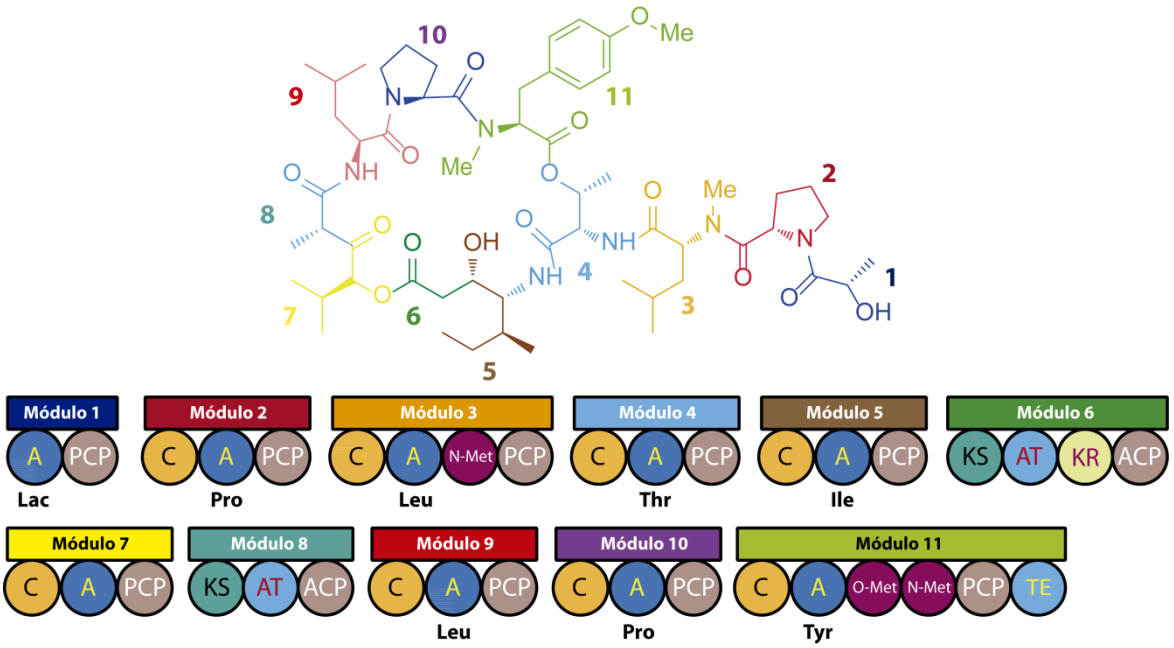

Figura 20 | Propuesta de síntesis de la molécula de didemnina B. Representación de los módulos teóricos que podrían dar lugar a cada una de las subunidades de la molécula de didemnina B. Cada una de las posibles subunidades de la molécula aparece coloreada y numerada de forma correspondiente su respectivo módulo de síntesis. El posible precursor incorporado aparece señalado únicamente en los casos en los cuales se puede pronosticar de forma evidente.

Para terminar se observan residuos de leucina, prolina y tirosina doblemente metilados, de cuya síntesis serían responsables 3 módulos más de tipo NRPS. En total y con la premisa de que actúen de forma lineal se propone la acción de aproximadamente 10-11 módulos de síntesis mixtos entre PKS y NRPS para sintetizar la molécula completa de didemnina B.

1.1.6. | Localización del clúster génico productor de didemninas

Basándose en la propuesta de síntesis de la didemnina B, se realizó un análisis del conjunto de los distintos clústeres génicos de *T. mobilis* MES-10-09-028 representados en la secuencia genómica obtenida. Para ello se utilizó el programa AntiSMASH (Medema *et al*., 2011). Esta herramienta está especializada en la anotación de agrupaciones génicas de producción de metabolitos secundarios, siendo más efectiva y manejable en estas tareas si se compara con otras herramientas de anotación automática como por ejemplo RAST. En primera instancia se utilizó esta herramienta en su versión 1.0 y se consiguió identificar un total de 12 clústeres génicos con presencia de genes que codifican proteínas implicadas en la producción de metabolitos secundarios. Únicamente 2 de las agrupaciones de genes resultaron ser híbridos entre PKS y NRPS y por lo tanto cumplirían con las condiciones de la propuesta de síntesis. Realizando un análisis más exhaustivo se observa que únicamente el numerado como clúster 2, tendría una cantidad y una distribución de módulos de síntesis semejante a los propuestos anteriormente. El núcleo de NRPS/PKS, sin contar con

los genes adyacentes que probablemente estuviesen implicados en la síntesis, se encontraría entre la posición 473187 y 527748 del *Scaffold 1*. Sin embargo se puede observar que la predicción por parte de AntiSMASH genera zonas donde la anotación aparentemente no es consistente, provocando lagunas en la secuencia en los que se detectan pequeñas ORFs sin sentido biológico, o directamente, la existencia de zonas sin anotar. También, al observar la predicción del contenido de todas las ORFs del clúster génico se hace patente la presencia de algunos módulos *a priori* incompletos. Esto podría deberse a la existencia de *gaps* y otros errores de secuenciación, por lo tanto sería necesario un refinado manual en detalle de la anotación. Aun así existen evidencias claras que indicarían que el clúster 2 podría ser el responsable de la producción de didemnina B, como por ejemplo: i) existen algunas ORFs con sucesiones de dominios de adenilación coincidentes con la propuesta de síntesis ii) existe un módulo con la presencia de un dominio tioesterasa.

1.1.7. | Curación manual de la secuencia del clúster génico productor de didemninas

A la vista de estos resultados se decidió realizar un refinamiento manual de la secuencia perteneciente a la zona del clúster 2. Teniendo en cuenta únicamente la secuencia del núcleo de síntesis formado por los genes que codifican NRPS o PKS, se observaron 5 *gaps* en la secuencia cuya longitud ha sido predicha con los datos de la secuenciación de *Paired End* (ver tabla 6). Para obtener la secuencia completa se decidió realizar amplificaciones mediante PCR de las zonas en cuestión, diseñando los cebadores en los extremos conocidos de la secuencia que flanquea cada *gap*. Estos amplicones se secuenciaron utilizando el método Sanger obteniéndose la secuencia de los 5 *gaps*.

Gap	Inicio	Fin	Longitud pronosticada	Longitud final	Cebadores de secuenciación
1	524119	524081	20	-37	GAP1F/GAP1R
2	514271	514318	807	48	GAP2F/GAP2R
3	507562	507614	20	53	GAP3F/GAP3R
4	487402	487408	689	7	GAP4F/GAP4R
5	486786	486826	371	41	GAP5F/GAP5R
Out1	540239	540538	640	300	GAPOUT1F/GAPOUT1R

Tabla 6 | *Gaps* presentes en la secuencia correspondiente al clúster de síntesis de didemninas en *T. mobillis* MES-10-09-028. Para cada uno de los *gaps* encontrados se detallan las posiciones de inicio y de fin, el pronóstico de longitud extraído de la secuenciación PE, la longitud final y la pareja de cebadores utilizada para su secuenciación (ver Materiales y Métodos). La longitud negativa del *gap* 1 muestra que en este caso existían repeticiones que generan artefactos casi idénticas en ambos extremos del gap.

Al analizar la secuencia adyacente al núcleo NRPS/PKS se observó que el *gap* más cercano al extremo 5' del clúster se encontraba a 25175 bp de distancia, por lo que se descartó para su posterior análisis, ya que por la anotación de las ORFs cercanas, no se consideró que pudiese afectar a genes adyacentes pertenecientes al clúster de la didemnina B. Por otro lado, en el extremo 3' se encontró un *gap* a una distancia de 12782 bp. Este fragmento si fue secuenciado, ya que los datos obtenidos de la anotación de RAST y AntiSMASH, se consideró que esa zona podría contener genes adyacentes que perteneciesen a la agrupación de genes de la síntesis de la didemnina B. El *gap* que se podía encontrar en el extremo 3' está a 54373 bp de distancia y por las ORFs circundantes se consideró que lo más probable es que no se encontrara en el clúster de producción de didemninas, por lo que se descartó su posterior análisis por secuenciación

Una vez conocida la secuencia de los *gaps* del núcleo PKS/NRPS del clúster se procedió a anotar de nuevo manualmente cada una de las ORFs. Para ello, para cada uno de los genes truncados se extendió la ORF correspondiente hasta el siguiente codón de parada, con lo cual se obtuvo un boceto mejorado de lo que podría ser el clúster génico producto de didemninas.

Sin embargo, al realizar un análisis de las nuevas ORFs aún se podían apreciar *gaps* en la secuencia y tras analizar la estructura de dominios de las nuevas ORFs se pudo observar que algunas podrían continuar truncadas. Esto hizo pensar que era posible que hubiese errores en la secuencia arrastrados desde el ensamblaje. Para subsanar estos errores se realizó un examen exhaustivo de todas las zonas que parecían conflictivas, para cada caso se siguió el siguiente procedimiento general:

- Se detectaron todas las ORFs posibles en la zona problema y se compararon en la base de datos de Pfam.
- En el caso de encontrarse un dominio no contenido en ninguna ORF, se realizó la comparación con las ORFs adyacentes ya anotadas, con el fin de comprobar si se trataba de una continuación.
- Si se trataba de una posible ORF truncada, se localizó la zona que contenía el error de secuencia que causó dicho truncamiento.
- Una vez detectado el error, se comprobó si se trataba de un error común de secuenciación como por ejemplo la inserción o deleción de un nucleótido en una secuencia repetida de 3 o más veces el mismo nucleótido.
- Se realizó el mapeo del conjunto de las lecturas en la zona problema y se contrastó que la versión correcta estuviese representada en un porcentaje razonable de las lecturas.

Tras tener en cuenta estos criterios se realizaron correcciones a lo largo de todo el núcleo PKS/NRPS del clúster y se procedió a anotar la secuencia manualmente.

1.1.8. | Anotación manual del núcleo del clúster productor de didemninas

Una vez corregidos los errores de la secuencia se consiguió generar manualmente una anotación más precisa. En concreto, el núcleo que contiene las enzimas PKS y NRPS posee 54562 bp y está anotado en lo que parece un único operón que contiene 8 genes (Fig. 21). El gen inicial *ddnA* contiene un primer dominio de condensación seguido de dos módulos completos de NRPS que incorporarían dos residuos de glutamina. A continuación se encuentra un módulo completo de NRPS con un dominio ceto reductasa (KR). Este módulo tiene la particularidad de que no es posible predecir la afinidad de su posible dominio de adenilación con certeza con ninguna de las herramientas de predicción que incorpora AntiSMASH. El gen *ddnC* contiene el cuarto, quinto, sexto y séptimo módulos que agregarían residuos de prolina, leucina, treonina e isoleucina, respectivamente, conteniendo el módulo de la leucina un dominio epimerasa y otro de metilación. El cuarto gen es una PKS con un dominio ceto reductasa. El gen *ddnE* contiene un único módulo NRPS sin la presencia de un dominio de condensación y con un dominio ceto reductasa en el cual no se puede predecir su afinidad mediante los distintos algoritmos de análisis que utiliza AntiSMASH. Este módulo es seguido por otro de tipo PKS que contiene un dominio metilasa. El gen *ddnG* contiene dos módulos de NRPS que añadirían residuos de leucina y

de prolina y finalmente el gen *ddnH* contiene un único módulo de tipo NRPS con dos dominios de metilación y un dominio tioesterasa en el extremo C-terminal que sería el responsable de circularizar la molécula y de su posterior liberación del complejo multiproteico.

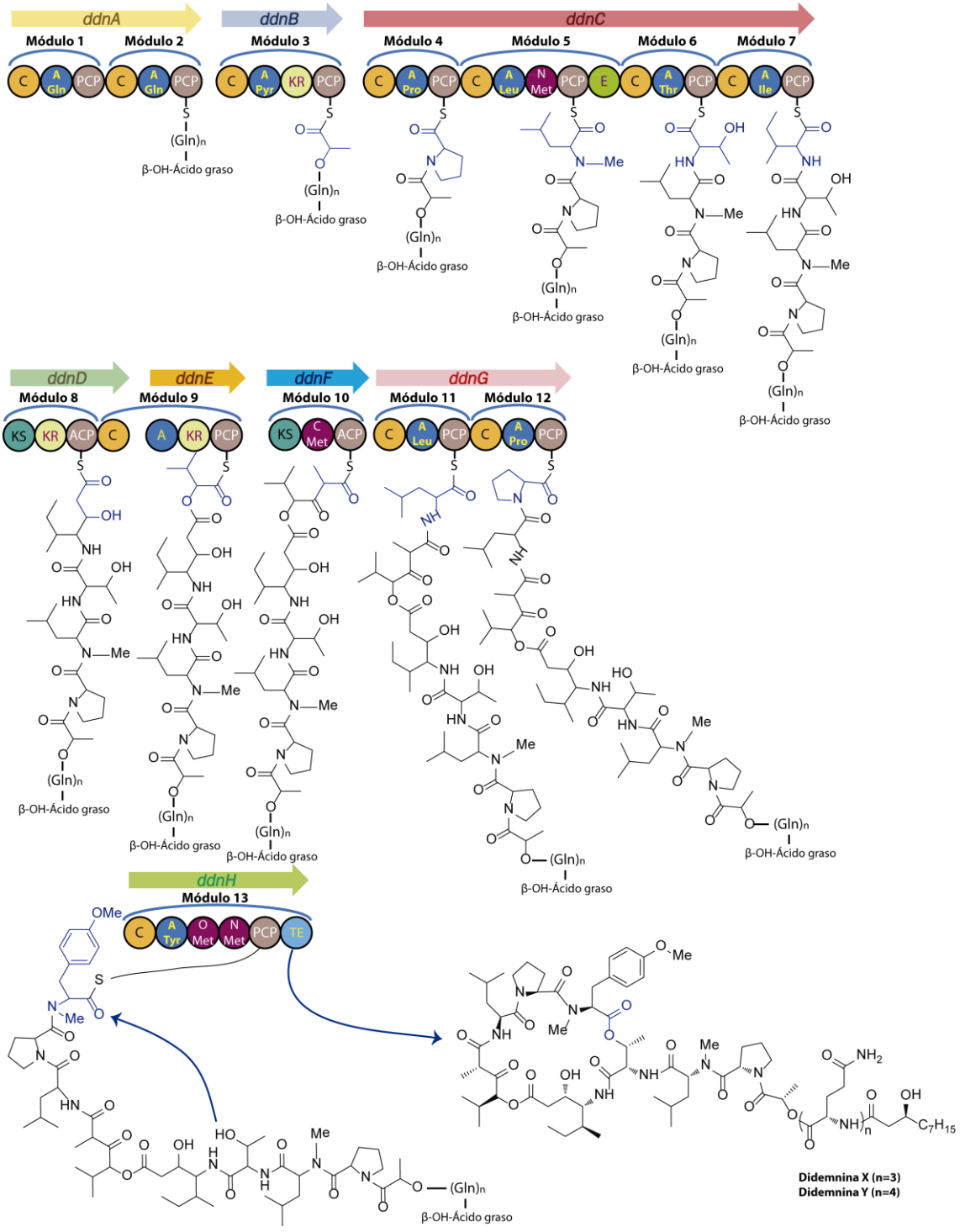

Figura 21 | Esquema de la síntesis de didemninas en *T. mobilis* MES-10-09-028. El núcleo PKS/NRPS del clúster de síntesis de didemninas en *T. mobilis* MES-10-09-028 contiene un total de 8 genes representados por flechas en los cuales se pueden definir 13 módulos de síntesis representados mediante círculos. La nueva molécula incorporada a la cadena en cada paso se marca en azul. En un último paso, el dominio TE del módulo 13 libera la molécula provocando una circularización de la misma y produciendo didemnina X o Y.

1.1.9. | Diferencias entre la anotación y la propuesta de síntesis

Al comparar la anotación final obtenida para el posible clúster génico productor de la didemnina B y la propuesta de síntesis generada inicialmente es evidente que aunque en general ambos coinciden en muchos puntos, existen algunas diferencias. La que puede resultar más llamativa está situada en el módulo inicial, ya que según la propuesta de síntesis incorporaría una molécula de lactato, sin embargo, el conjunto génico obtenido es consistente con este modelo unicámente si se ignora la presencia de los dos primeros módulos contenidos en el gen *ddnA*. Estos dos módulos son de tipo NRPS y su posible función sería la de incorporar cada uno dos moléculas de glutamina, que no se encuentran en la estructura de la molécula de didemnina B. Sin embargo, analizando otras didemninas distintas que también están presentes en los extractos de *T. mobilis* MES-10-09-028, se observa que las didemninas X e Y poseen una estructura en la cual hay una repetición de 3 o 4 residuos de glutamina respectivamente cerca de donde se encontrarían los residuos iniciales. Además de estas repeticiones peptídicas, la síntesis de estas dos moléculas comenzaría por un componente lipídico que consiste en un β-hidroxi-ácido. Por lo tanto, los genes encontrados podrían corresponder con la síntesis de las moléculas de didemnina X e Y en el caso de que la adición de glutamina sea iterativa, obteniéndose variantes de 3 y 4 residuos, y además teniendo en cuenta que el primer dominio de condensación del módulo 1 pudiese iniciar la síntesis añadiendo el precursor lipídico. Esto sugiere que el módulo 3 sería el equivalente al módulo inicial de la propuesta de síntesis y su dominio de adenilación incorporaría una molécula de piruvato que sería ceto-reducida hasta lactato.

Otra de las diferencias que se hace patente es que los módulos PKS no poseen dominios acil-transferasa (AT), lo que hace pensar en un sistema trans-AT (ver Introducción apartado 3.2.1.1). Sin embargo, al observar las inmediaciones del clúster, no se localiza ninguna enzima con dominio AT, por lo que puede que este clúster necesite enzimas con esta función del metabolismo de la síntesis de ácidos grasos, como es por ejemplo el caso de FabD (Wesener *et al*., 2011).

También es necesario comentar la presencia de lo que en primera instancia parecería un módulo incompleto de tipo NRPS en el gen *ddnE*, cuyo dominio de condensación inicial parece encontrarse en el gen anterior, por lo tanto, el módulo 8 estaría compartido entre dos ORFs distintas. Este módulo debería ser el encargado de incorporar ácido 2-hidroxisovalérico y el siguiente continuaría extendiendo la molécula con malonato.

1.1.10. | Análisis de los genes adyacentes al núcleo del clúster de producción de didemninas

Una vez generada la propuesta de anotación del núcleo PKS/NRPS, se procedió a anotar las secuencias adyacentes con el objetivo de comprobar la presencia de otros genes que pudieran estar implicados en la síntesis de didemninas. En primer lugar se partió de la anotación generada anteriormente con la herramienta RAST. A continuación se revisó manualmente cada una de las ORFs detectadas intentando reparar los posibles genes truncados y se anotaron manualmente aquellas ORFs que por distintas razones no habían sido anotadas con las herramientas de anotación automáticas.

Para cada una de las ORFs detectadas se realizó un proceso manual de anotación funcional que consistió en realizar comparaciones utilizando las herramientas HMMER y las variantes de BLAST en la bases de datos del NCBI. Se prestó especial atención a la detección de distintos dominios funcionales utilizando la base de datos Pfam. De este modo finalmente se generó la anotación manual estructural y funcional de los genes adyacentes (ver tabla 7).

Gen	Anotación	Inicio	Fin	Longitud (bp)
orf1	PPTasa	543322	542447	876
orf2	Intercambiador sodio/hidrógeno	540796	542529	1734
orf3	Proteina de transporte con dominio MMPL	539984	537647	2338
orf4	Tioesterasa	537564	536779	786
orf5	Posible proteasa de membrana (Similar a estomatina y prohibitina)	535324	536775	1452
orf6	Transportador ABC de unión a ATP	533645	535327	1683
orf7	GTPasa	532592	533176	585
orf8	Transportador de eflujo multidroga	529222	532514	3293
orf9	Transportador de eflujo multidroga de la familia RND	528089	529222	1134
ddnA	NRPS	527748	521350	6399
ddnB	NPRS	521302	515963	5340
ddnC	NRPS	515966	500421	15546
ddnD	PKS	500396	495279	5118
ddnE	NRPS	495263	490413	4851
ddnF	PKS	490416	486166	4251
ddnG	NRPS	486169	479726	6444
ddnH	NRPS	479729	473187	6543
orf10	Proteina con dominio MbtH	473120	472881	240
orf11	Proteina hipotética	472795	472568	228
orf12	Posible transposasa	472419	472188	232
orf13	Posible transposasa	472160	471810	351
orf14	Posible transposasa	471820	471254	567
orf15	Posible lipasa	470317	471297	981
orf16	Posible Integrasa	469959	470210	252
orf17	Posible Transposasa/Integrasa	469305	469922	618
orf18	Posible transposasa	469078	469305	228
orf19	Proteasa amino terminal de la familia CAAX	467983	468951	969

Tabla 7 | Relación de los genes del clúster de producción de didemninas de la cepa *T. mobilis* MES-10-09-028. En la tabla se muestra el nombre del gen, la anotación las coordenadas y su longitud en nucleótidos. En rojo se señalan aquellos genes pertenecientes al núcleo PKS/NRPS del clúster y en blanco y gris aquellos genes del entorno.

Entre los genes adyacentes del extremo 5' del clúster génico responsable de la síntesis de didemninas podemos encontrar varios genes que codifican posibles proteínas de unión a membrana que podrían estar relacionadas con el transporte y con la resistencia a la didemnina B. La delimitación del conjunto génico en este extremo no es clara, ya que existen genes que codifican proteínas de función desconocida. Sin embargo, destaca la

presencia de una ORF que codifica una posible fosfopanteteinil transferasa (Walsh *et al*., 1997; Lambalot *et al*., 1996), enzima que es necesaria para activar las proteínas portadoras presentes en la biosíntesis de policétidos, péptidos no ribosomales y ácidos grasos. Esta actividad suele ser esencial para la biosíntesis y curiosamente realizando una búsqueda más detallada en el genoma de *T. mobilis*, se trata de la única ORF que podría ser responsable de esta función. Por lo tanto no es posible concretar si esta ORF pertenece de manera específica al clúster génico productor de didemninas, ya que resultaría también esencial para otros muchos procesos celulares.

En el extremo 3' se puede apreciar una concentración de ORFs pequeñas e de función hipotética y una ORF que codifica una posible transposasa, la cual podría delimitar el grupo génico si se piensa en que inicialmente éste pudo ser adquirido por la bacteria mediante transferencia horizontal. Sin embargo, existen ORFs a continuación de las mencionadas que parecen tener funciones reguladoras, proteolíticas y de transporte, por lo que no se puede descartar que estos genes estén implicados en un procesamiento posterior de las moléculas resultantes de la biosíntesis del núcleo del clúster.

1.1.11. | Comparación de los clústeres de síntesis de didemninas de las cepas *T. mobilis* MES-10-09-028 y *T. mobilis* KA081020-065

Una vez anotados manualmente los genes del núcleo del clúster génico y los genes adyacentes, se procedió a comparar la secuencia con la única existente en las bases de datos perteneciente a un clúster productor de didemninas (Xu *et al*., 2012). Está secuencia pertenece a *T. mobilis* KA081020-065, una bacteria muy parecida a *T. mobilis* MES-10-09-028 que fue publicada en el transcurso de este trabajo.

Al comparar la secuencia del núcleo PKS/NRPS del clúster de la productor de didemninas resulta evidente que ambos clústeres son prácticamente idénticos en lo que a presencia de dominios se refiere. Ambos codifican el mismo número de módulos y estos están conformados por los mismos dominios proteicos. Sin embargo, al observar la anotación realizada, se puede ver que la cepa KA081020-065 posee dos módulos que parecen estar anotados de forma fraccionada (Xu *et al*., 2012). En concreto, se trata de los dominios de condensación del módulo 5 y 12 que se encuentran divididos entre los genes *didC* y *didD*, y *didH* y *didI* respectivamente. Este fenómeno no se observa en la anotación realizada sobre la cepa MES-10-09-028, y es complicado pensar que estos dominios pudieran encontrarse fragmentados para posteriormente operar unidos en un plegamiento que implique dos proteínas distintas, cuando lo común es encontrarlos en la misma ORF (Fig. 22). Esto hace pensar que la cepa KA081020-065 contiene errores que han sido arrastrados desde la secuenciación en los cuales hay codones de parada prematuros que truncan la proteína generando 2 ORFs distintas, cuando en realidad deberían estar formando una única ORF.

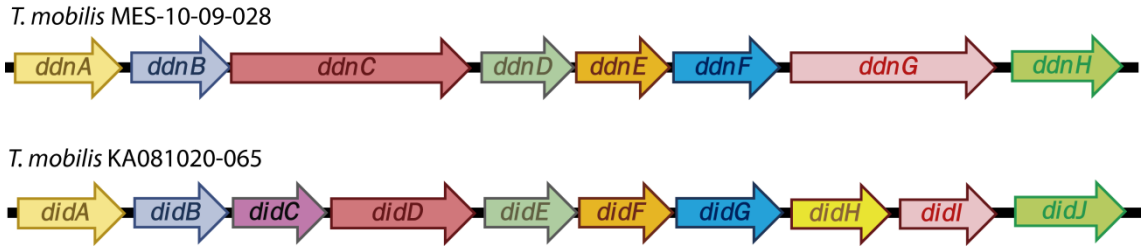

Figura 22 | Comparación de los clústeres de producción de didemninas de las cepas MES-10-09-028 y KA081020-065. Disposición esquemática de las ORFs del núcleo PKS/NRPS de los clústeres de producción de didemninas de las cepas MES-10-09-028 y KA081020-065.

Al comparar los genes adyacentes al núcleo PKS/NRPS, se puede observar que las delimitaciones de genes posiblemente pertenecientes al cluster DidB no coinciden totalmente. En las ORFs coincidentes del extremo 3' del clúster se observan algunas diferencias. En concreto, las diferencias se encuentran en la zona entre la *orf13* de la cepa KA081020-065 (no anotada funcionalmente como transposasa) y su transposasa homóloga *orf18* de la cepa MES-10-09-028, y la *orf8* y su homóloga *orf10* en la cepa MES-10-09-028, que resulta ser una proteína que contiene un dominio MbtH. Entre estos dos genes existen sucesiones de varias ORFs que codifican posibles proteínas pequeñas, muchas de ellas aparentemente relacionadas con elementos móviles. Sin embargo, en la cepa MES-10-09-028 se han anotado 7 posibles ORFs mientras que en la cepa KA081020-065 únicamente se han detectado 4, que además difieren estructuralmente de las 7 anotadas en nuestra cepa en estudio. Además, en este extremo del clúster se propone que el límite 3' podría estar en la *orf18* y no dos ORFs más allá como se plantea en la cepa KA081020-065 (ver tabla 7).

En el extremo 5' del núcleo PKS/NRPS, en el caso de la cepa KA081020-065, se han anotado un total de 7 ORFs que coinciden estructuralmente sin grandes diferencias con las ORFs homólogas de la cepa MES-10-09-028, excepto por la *orf4* (codifica 70 aa) que no se anotó en la cepa MES-10-09-028. Cabe destacar que la propuesta hecha en este trabajo extiende a 3 ORFs más los límites del clúster con respecto a la propuesta de la cepa KA081020-065, incluyéndose la *orf1* que codifica la posible enzima fosfopanteteinil transferasa antes mencionada (ver tabla 7). Esta enzima no aparece en la propuesta de anotación de genoma completo de Xu *et al*. (2012) (CP003239), aunque un análisis detallado de la secuencia de la cepa KA081020-065 indica que dicha secuencia existe pero no se ha detectado y por lo tanto no ha sido estructuralmente anotada.

1.1.12. | Análisis de la expresión del clúster productor de didemninas en *T. mobilis* MES-10-09-028

Una vez anotados los genes del núcleo PKS/NRPS del clúster productor, se decidió analizar la expresión mediante RT-PCR semicuantitativa de dicho clúster en condiciones de producción en la cepa MES-10-09-028. Para llevar a cabo estos experimentos se extrajo RNA de cultivos realizados en medio de producción LB a 30 °C, que se encontraban en la fase exponencial de la curva de crecimiento (DO_{600nm} = 1,8). Una vez extraído el RNA, se realizó la correspondiente retrotranscripción a cDNA. Para la posterior PCR se diseñaron 5 parejas de cebadores que amplificaban fragmentos de aproximadamente 200 bp situados en los 4 primeros genes del núcleo PKS/NRPS del clúster (ver tabla 3 en Materiales y Métodos). Se obtuvo la amplificación del cDNA con todas las parejas de cebadores

utilizadas (Fig. 23). En el caso del fragmento en el gen *ddnA* se obtuvo una menor cantidad de amplicón que podría deberse a una pequeña degradación del extremo 5´ del RNA policistrónico.

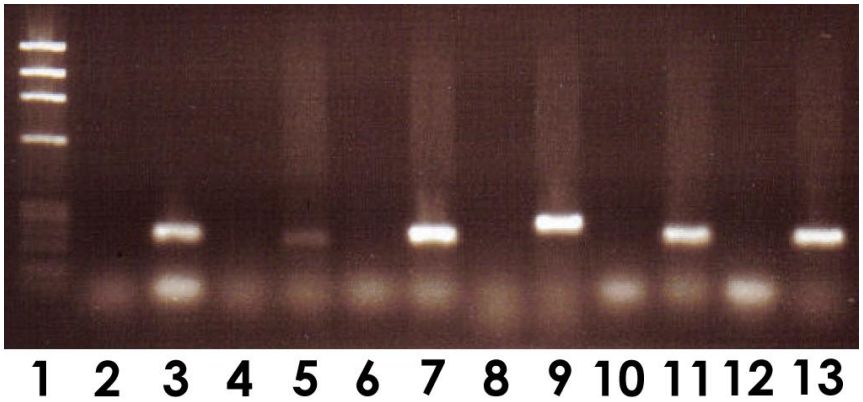

Figura 23 | Gel de la RT-PCR semicuantitativa del clúster productor de didemninas en la cepa *T. mobillis* MES-10-09-028. 1: φX174 *Hae*III. **2:** Control negativo sin DNA. **3:** Control positivo con DNA genómico de la cepa MES-10-09-028. **4 y 5:** Respectivamente, control sin retrotranscriptasa y cDNA amplificados con la pareja de cebadores SCDidA. **6 y 7:** Respectivamente, control sin retrotranscriptasa y cDNA amplificados con la pareja de cebadores SCDidB. **8 y 9:** Respectivamente, control sin retrotranscriptasa y cDNA amplificados con la pareja de cebadores SCDidC. **10 y 11:** Respectivamente, control sin retrotranscriptasa y cDNA amplificados con la pareja de cebadores SCDidD. **12 y 13:** Respectivamente, control sin retrotranscriptasa y cDNA amplificados con la pareja de cebadores SCDidE.

Estos resultados sugieren que la expresión transcripcional del clúster productor de didemnina en *T. mobilis* se encuentra activa en fase exponencial, ya que se ha detectado la expresión de al menos las 4 primeras ORFs del núcleo PKS/NRPS del clúster.

1.2. | Generación de mutantes de la cepa *T. mobillis* MES-10-09-028 en el clúster de síntesis de didemninas y análisis de las diferencias en la producción

1.2.1. | Transformación genética de la cepa de *T. mobilis* MES-10-09-028

Hasta la fecha no existen ejemplos en la literatura de modificación genética de una cepa bacteriana perteneciente al género *Tistrella*. Por este motivo, al inicio de este trabajo se desconocían las características para que un vector pudiera ser introducido y replicara con éxito en *T. mobilis* MES-10-09-028. Para estudiar estas propiedades en primer lugar se comprobó la resistencia de la cepa a diferentes antibioticos utilizando concentraciones de uso común en el laboratorio. En concreto, se utilizó 100 µg/mL de ampicilina, 10 µg/mL de gentamicina, 50 µg/mL de kanamicina, 20 µg/mL de cloranfenicol, 50 µg/mL de estreptomicina y 7 µg/mL de tetraciclina. Se cultivó la cepa MES-10-09-028 durante 48 h a 30 °C en placas de medio LB-agar con cada uno de los antibióticos descritos. La cepa resultó sensible a todos los antibióticos probados con la única excepción de la ampicilina. Es interesante destacar que las colonias de *T. mobilis* cultivadas en presencia de ampicilina mostraban un fenotipo de colonia distinto, al que se puede observar en el control sin antibiótico. En concreto las colonias poseían una estructura más mucosa.

El siguiente paso fue la selección de vectores replicativos atendiendo a distintos criterios. Para ello se acudió a la serie SEVA (*Standard European Vectors Architecture*) en la que se dispone de una batería de plásmidos con diferentes características (Silva-Rocha *et*

al., 2013) En primer lugar se procedió a seleccionar vectores que contuviesen genes de resistencia a antibióticos a los que *T. mobilis* MES-10-09-028 no es resistente. En concreto los vectores pSEVA224 (Km^R) y pSEVA424 (Sm^R) poseen el origen de replicación del plásmido *RK2* de amplio espectro de hospedador, y además cuentan con el sistema de inducción de la expresión *lacI*^{q}-P_{TRC}. Para poder transformar *T. mobilis* con estos plásmidos se prepararon células competentes tal y como se describe en Materiales y Métodos. Ambos plásmidos se electroporaron en la cepa MES-10-09-028 y tras 3 d de incubación a 30 °C se obtuvieron colonias transformantes en placas de LB con kanamicina en el caso del pSEVA224 o estreptomicina en el caso de pSEVA424. Para cada caso se realizó una extracción de plásmido de los cultivos generados a partir de colonias aisladas y se visualizaron mediante electroforesis en gel de agarosa. Estos resultados nos permitieron concluir que la cepa puede ser transformada por electroporación, que los genes de resistencia se expresan correctamente, y que el origen de replicación de tipo *RK2* permite que estos vectores se repliquen de manera estable en *T. mobilis*.

1.2.2. | Generación del mutante KR3 de *T. mobilis* MES-10-09-028

Partiendo de la base de que *T. mobilis* MES-10-09-028 es genéticamente modificable y con el objetivo de comprobar que el posible clúster génico de la síntesis de didemninas es el verdadero responsable de la producción de estos compuestos, se decidió realizar un mutante que afectase a la generación de la molécula. Para ello se decidió generar un mutante que tuviese inactivado el dominio cetoreductasa (KR) del módulo 3 de la síntesis (Fig. 24).

Como ya se ha comentado anteriormente el dominio KR es el responsable de reducir grupos ceto a grupos hidroxilo. Inactivando este dominio teóricamente se podría determinar si es posible incorporar directamente una molécula de piruvato en el módulo 3 en lugar de la molécula de lactato generada tras el paso de ceto-reducción. Así, se comprobaría si es posible generar una nueva molécula sin los residuos incorporados por los dos primeros módulos, ya que no existiría la posibilidad de enlazar la molécula al piruvato que se añade en el módulo 3 porque este posee en lugar del grupo hidroxilo un grupo ceto (Fig. 24). La didemnina derivada de este mutante sería la dehidrodidemnina B, también conocida como aplidina (ver Fig. 9 en Introducción).

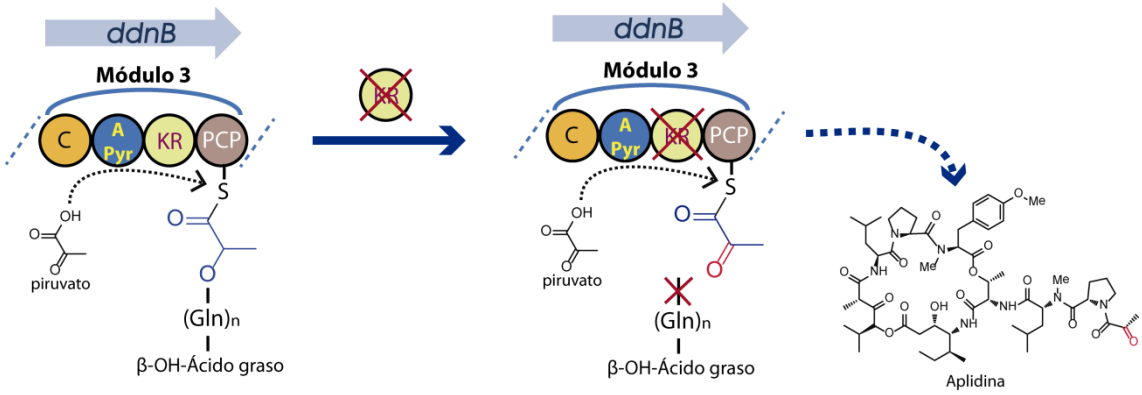

Figura 24 | Mutación KR3 en el clúster de producción de didemninas. La inactivación del dominio KR no actuaría sobre la molécula de piruvato que se incorpora, manteniendo así su grupo ceto (rojo) e impidiendo la unión de la cadena generada por los módulos 1 y 2. En el caso de que la síntesis fuese posible, la molécula que teóricamente debería producirse sería la aplidina.

1.2.2.1. | Diseño del mutante KR3 de *T. mobilis* MES-10-09-028

Para genera el mutante, en primer lugar se identificó y se delimitó el dominio KR dentro del módulo 3 haciendo una búsqueda con la ORF perteneciente al gen *ddnB* en la base de datos Pfam. Esta búsqueda delimitó el dominio entre los aminoácidos 1404 y 1563 de la ORF.

Con la intención de encontrar el centro catalítico de los dominios ceto-reductasa (KR) en clústeres génicos PKS/NRPS para posteriormente inactivarlo, se identificó un residuo de tirosina muy conservado en el centro activo de este tipo de dominios y que puede actuar en la catálisis (Reid *et al*., 2003). Para ello se analizó en el modelo HMM del dominio KR (PF08659) depositado en Pfam, situándose en la posición 149. Seguidamente, se realizó un alineamiento múltiple con las 85 secuencias de dominio KR que habían generado el dominio KR de Pfam y la propia secuencia del módulo 3. De esta manera se consiguió encontrar la posición de la tirosina catalítica en el gen *ddnB*, que en concreto, sería el aminoácido 1531 de la secuencia proteica. Con el fin de reducir drásticamente la actividad de este dominio se decidió llevar a cabo la estrategia utilizada previamente por Power et al. (2008) para el caso de la anfotericina en la que se sustituye la tirosina catalítica por fenilalanina reduciendo así la síntesis de esta molécula. La sustitución de la tirosina por la fenilalanina sólo debería afectar a la actividad del dominio reductasa ya que no implica cambios estructurales muy drásticos que pudieran afectar a las otras actividades enzimáticas de los otros dominios.

1.2.2.2. | Estrategia de sustitución de la tirosina catalítica por un residuo de fenilalanina en el dominio KR del módulo 3

Para realizar la sustitución aminoacídica propuesta se eligió una estrategia de doble recombinación mediante el plásmido suicida pK18*mob*sacB (Schäfer *et al*., 1994). Para conseguir una sustitución dirigida en la secuencia se amplificaron dos fragmentos del gen *ddnB* a los que se llamarán A y B, los cuales solapan en la zona donde se encuentra el codón que codifica para la tirosina catalítica del dominio KR (ver Materiales y Métodos). El amplicón A de 805 bp está delimitado en el extremo 5' por la secuencia del cebador DidBKR3AF que contiene una diana de restricción para *Hin*dIII. En el extremo 3' sin embargo se encuentra limitada por la secuencia del cebador DidBKR3AR que es el responsable de introducir 2 mutaciones en la secuencia. Una de estas mutaciones modifica la secuencia del codón que codifica la tirosina catalítica por la de un codón que codifica una fenilalanina, provocándose una sustitución de TAT por TTT. La otra mutación es una mutación silenciosa, ya que el codón pasa de GGC a GGT, y ambos codifican glicina. Este intercambio nucleotídico genera una diana de restricción *Kpn*I en la secuencia, que permite realizar las construcciones correspondientes a las zonas A y B en el plásmido pK18*mob*sacB.

El amplicón B, de 850 bp, limita en su extremo 5' con la secuencia del cebador DidBKR3BF que al igual que el cebador DidBKR3AR antes mencionado va a generar las mismas mutaciones pero en la hebra contraria. En el extremo 3' la zona B limita en la secuencia del cebador DidBKR3BR que contiene una diana de restricción para la enzima *Xba*I. De este modo se puede observar que los productos amplificados A y B son solapantes en una zona de 19 bp en la cual comparten la diana de *Kpn*I.

Una vez realizadas ambas amplificaciones el fragmento A fue digerido con las enzimas *Hin*dIII y *Kpn*I, mientras que el fragmento B fue digerido con *Kpn*I y *Xba*I. Esto permitió realizar un clonaje dirigido de la construcción AB, unidas por la diana *Kpn*I en el plásmido pK18*mob*sacB, generándose así el vector pK18KR3, el cual contiene una reconstrucción de un fragmento total de 1636 bp perteneciente a la secuencia del gen *ddnB*, pero que contiene las dos mutaciones antes mencionadas. La introducción de la mutación en esta construcción se comprobó por secuenciación.

1.2.2.3. | Generación de la cepa mutante *T. mobilis* KR3

Una vez generado el vector pK18KR3 éste se electroporó en células competentes de *T. mobilis* MES-10-09-028. Los transformantes tras la primera recombinación homóloga se seleccionaron en placas de medio LB + Km. Una vez seleccionados se cultivan en medio líquido en presencia de sacarosa para forzar la segunda recombinación homóloga. De este modo se seleccionaron aquellas colonias que habían perdido la resistencia al antibiótico sembrando la misma colonia en placas de LB con y sin Km. Las bacterias transformantes que portaban el DNA recombinante se identificaron mediante PCR y posterior secuenciación. Esta nueva cepa generada se nombró como *T. mobilis* KR3.

1.2.2.4. | Análisis de la producción de didemninas de la cepa mutante de *T. mobilis* KR3

Para comprobar si la capacidad de la cepa KR3 de producir didemninas se había visto afectada tras realizar las mutaciones en el dominio KR del módulo 3, se analizó la producción de didemninas en cultivos en matraces con 40 mL de medio LB a 30 °C. Como control positivo se utilizó la cepa silvestre MES-10-09-028 con la que anteriormente se había detectado en estas condiciones producción de didemninas.

La producción de didemninas se analizó a las 96 h de cultivo. Se realizaron extracciones con isopropanol y acetato de etilo tanto de las células como del sobrenadante del cultivo, y tras su evaporación se analizaron mediante HPLC-MS. En los extractos celulares de la cepa KR3 sólo se detectó un pequeño pico perteneciente a la didemnina B, el cual se identifica correctamente mediante masas (Fig. 25). Sin embargo en la cepa silvestre utilizada como control se detectaron los picos pertenecientes a la didemnina X, Y, nor-didemnina B y didemnina B. Al analizar los sobrenadantes también se detectó aunque en muy baja concentración didemnina B en el mutante KR3, mientras que en la cepa silvestre, además de didemnina B también se detectó una pequeña proporción de nor-didemnina B.

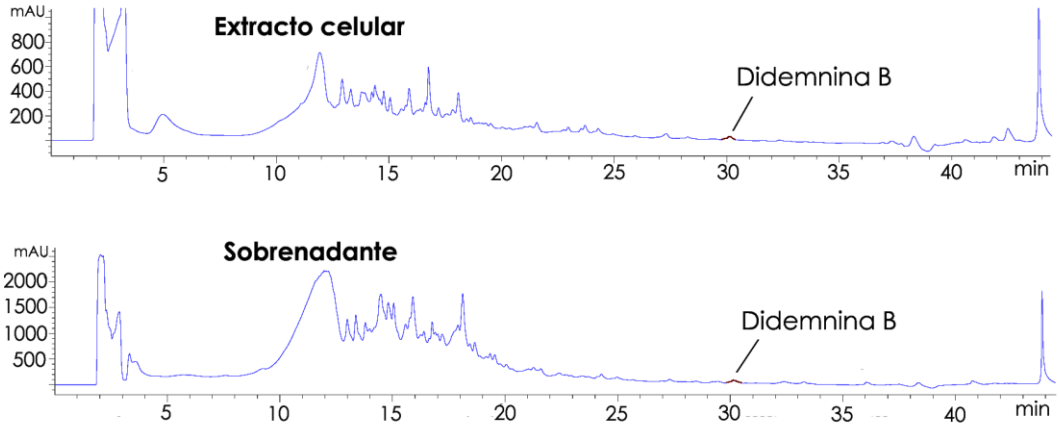

Figura 25 | Análisis mediante HPLC-DAD de la producción de didemninas del mutante *T. mobilis KR3*. Cromatogramas a 215 nm correspondientes al extracto celular y al sobrenadante a las 96 h de una fermentación de la cepa KR3. En rojo se señalan los picos correspondientes a didemninas.

Con el fin de aumentar la sensibilidad de la detección de los derivados de didemnina en el caso del mutante KR3, se realizó otro cultivo de 500 mL para concentrar el extracto y comprobar si se identificaban otras didemninas. De esta forma la muestra se concentró 12,5 veces con respecto al ensayo anterior. En este caso, los tiempos finales para proceder al análisis del cultivo se fijaron en 120, 144 y 168 h de cultivo y se analizaron los extractos de la mezcla del sobrenadante y del contenido celular. Como se puede observar en los cromatogramas de la figura 26, se detecta un único pico que corresponde a didemnina B de forma concluyente en los tres tiempos de cultivo.

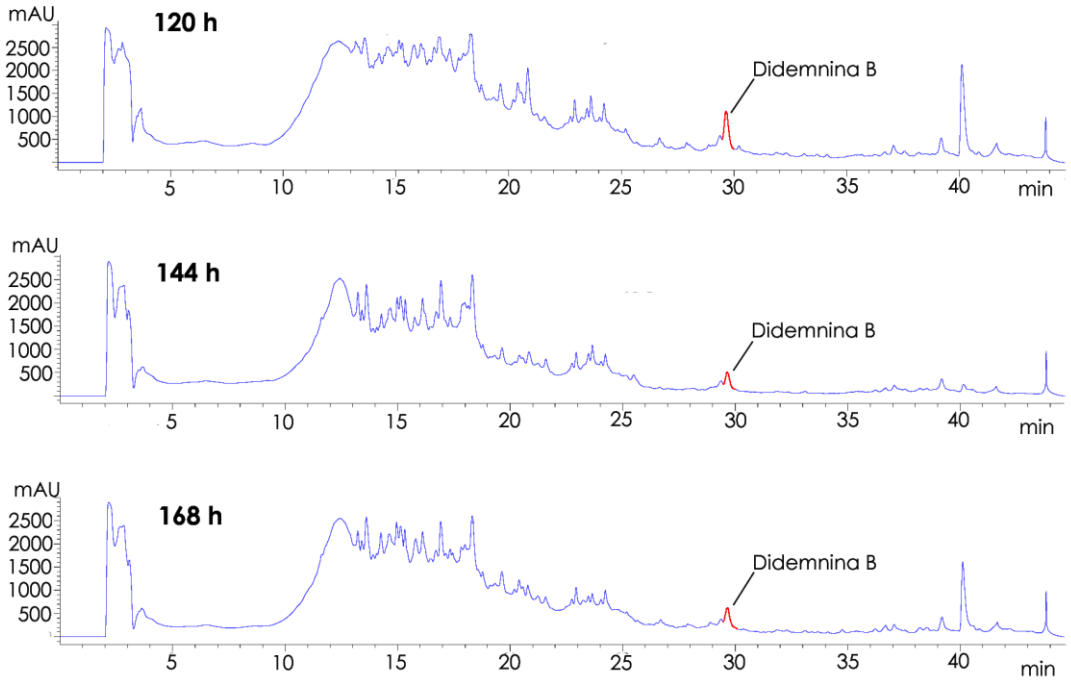

Figura 26 Análisis mediante HPLC-DAD de la producción de didemninas en fermentaciones concentradas de la cepa KR3. Los cromatogramas a 215 nm se han obtenido de cultivos concentrados a las 120, 144 y 168 h de fermentación de la cepa KR3. En rojo se señalan los picos correspondientes a didemninas.

Con estos resultados se puede concluir que una mutación en el dominio KR del módulo 3 afecta de manera drástica a la producción total de didemninas, demostrándose que este clúster es el verdadero responsable de la síntesis de didemnina B. Por otra parte, con esta mutación se ha eliminado la producción de otras didemninas que habitualmente se sintetizan junto a la didemnina B. Esto podría deberse a la aparición de un nuevo cuello de botella en la síntesis. Además, al detectarse trazas de didemnina B en los extractos de la cepa KR3, se puede concluir que la sustitución de la tirosina por la fenilalanina no inactiva completamente la actividad catalítica del dominio KR del módulo 3. Por último, se comprueba que no se ha conseguido producir la aplidina, lo que implica que el módulo 3 no puede funcionar como iniciador la hora de producir este derivado.

1.2.3. | Generación de los mutantes DidA y del doble mutante KR3DidA en *T. mobilis* MES-10-09-028

La razón por la cual no se genera aplidina en el mutante KR3 podría deberse a la incapacidad del módulo 3 de actuar como módulo inicial al ir precedido por los módulos 1 y 2 y tener que realizar una reacción de condensación entre la molécula incorporada y la resultante de la acción de los dos primeros módulos. Por ello se planteó la posibilidad de eliminar los dos primeros módulos y el dominio inicial de condensación del módulo 3 tanto en la cepa silvestre de la cepa MES-10-09-028 como en la mutante KR3 (Fig. 27).

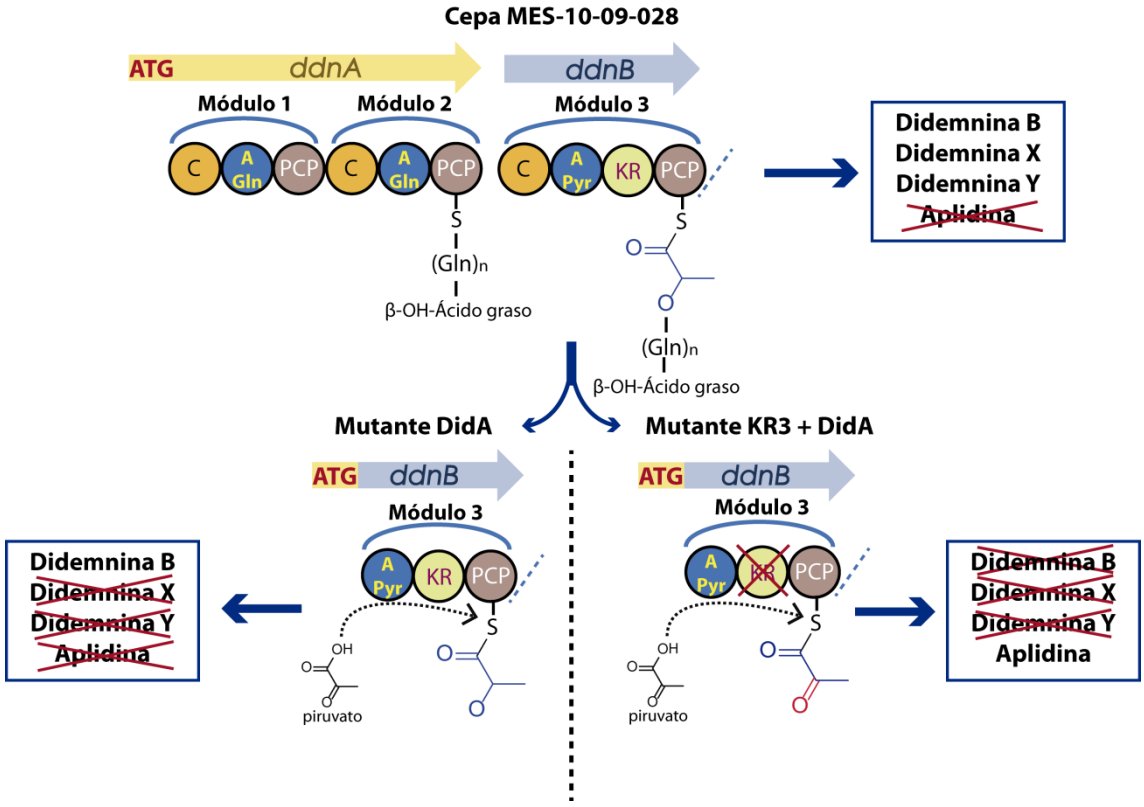

Figura 27 | Mutación DidA en la cepa silvestre MES-10-09-028 y en el mutante KR3. La mutación DidA elimina los dos primeros módulos de síntesis y el dominio de condensación del módulo 3 manteniendo el codón de inicio del gen ddnA. Estas modificaciones teóricamente implicarían cambios en el perfil de producción de didemninas, tal y como se especifica para uno de los casos.

1.2.3.1. | Diseño del mutante DidA en la cepa silvestre de *T. mobilis* MES-10-09-028 y en el mutante KR3

Para diseñar el mutante de deleción de los dos primeros módulos de síntesis contenidos en el gen *ddnA* se acudió a la anotación manual realizada. Se tuvo en cuenta la estructura de dominios de los 3 primeros módulos, y como se puede observar en la figura 27, el módulo 3 comienza con un dominio de condensación que sería el responsable de la incorporación de la molécula generada en los dos módulos anteriores. En un diseño en el cual la síntesis comenzase en el módulo 3, no sería necesario contar con un dominio de condensación inicial, ya que no se necesitaría incorporar un precursor a la molécula como en el caso del β-hidroxi-ácido graso probablemente incorporado por el módulo 1. Por este motivo se decidió eliminar por completo ese dominio de condensación junto a los dos primeros módulos. Este dominio se delimitó utilizando la herramienta Pfam y se localiza desde el aminoácido 40 al aminoácido 335 de la proteína DdnB. Para mantener el promotor original del clúster y para que el nivel de traducción sea similar al del gen original se decidió mantener el codón de inicio del gen *ddnA* y fusionarlo con la ORF del gen *ddnB* en la zona que está a continuación del dominio de condensación. En total la zona a delecionar es un fragmento de 8521 bp que contiene los módulos 1 y 2 junto al dominio de condensación del módulo 3.

Para llevar a cabo al deleción planteada se utilizó el vector pK18*mob*SacB y se amplificaron dos fragmentos, uno en la zona 5' fuera del gen ddnA y otro en la del gen *ddnB*, denominados A y B, que contienen la secuencia flanqueante a la región que se pretende delecionar (Fig. 12). En concreto el fragmento A comprende una secuencia de 814 bp que en su extremo 5' limitaría con la secuencia del cebador DidBAF que contiene la diana de restricción *Hin*dIII. En el extremo 3' se encuentra la secuencia del cebador DidBAR que coincide con la zona del codón de inicio del primer gen. Este cebador va a incorporar una mutación en los dos nucleótidos más cercanos al codón de inicio sustituyendo CG por AT y generando así una diana de restricción *Nde*I que contiene el codón de inicio de la traducción. El fragmento B en este caso consta de 873 bp de longitud y está delimitado en el extremo 5' por la secuencia del cebador DidBBF, que se sitúa 141 nucleótidos antes del comienzo del dominio de adenilación del módulo 3. Este cebador contiene también una secuencia de restricción *Nde*I que va a permitir la ligación de los fragmentos A y B manteniendo el codón de inicio ATG. Este fragmento se encuentra delimitado en el extremo 3' por la secuencia del cebador DidBBR, que coincide con la zona central del dominio de adenilación del módulo 3. Este cebador contiene la secuencia de restricción *Xba*I, que permite un posterior clonaje dirigido. Los fragmentos A y B únicamente solapan en los 6 nucleótidos de la diana *Nde*I, la cual contiene el codón de inicio ATG.

El fragmento A se digirió con las enzimas de restricción *Hin*dIII y *Nde*I, y a su vez, el fragmento B con *Nde*I y *Xba*I. Estos dos fragmentos se ligaron y fueron clonados de forma dirigida en el vector pK18*mob*SacB previamente digerido con las dianas *Hin*dIII y *Xba*I. A esta nueva construcción se le dio el nombre de pk18DidA, que contiene una secuencia de 1687 bp que corresponde a las zonas flanqueantes de la región que se quiere delecionar, pero que incorpora las modificaciones antes mencionadas.

1.2.3.2. | Obtención de los mutantes DidA de *T. mobilis* MES-10-09-028 y *T. mobilis* KR3

Del mismo modo que se generó el mutante KR3, el vector suicida pk18DidA se transformó mediante electroporación en células competentes de la cepa silvestre MES-10-09-028 así como en la cepa mutante KR3. Los transformantes obtenidos tras la primera recombinación homológa se seleccionaron en medio LB con Km. La segunda recombinación se forzó cultivando los transformantes en presencia de sacarosa. Los dobles transformantes se identificaron seleccionando aquellas cepas que había perdido la resistencia al antibiótico sembrando en placas de LB con y sin Km, y posteriormente se comprobó mediante PCR. Tras este paso se obtuvieron dos nuevas cepas mutantes, la primera de ellas fue nombrada como *T. mobilis* DidA y se trata de la cepa silvestre sin los dos primeros módulos de la síntesis de didemninas y sin el dominio de condensación del módulo 3. La segunda cepa, llamada *T. mobilis* DidAKR3 es un doble mutante con la deleción de *T. mobilis* DidA y además la mutación en la tirosina catalítica del dominio KR del módulo 3 del mutante *T. mobilis* KR3.

1.2.3.3. | Análisis de la producción de didemninas de las cepas mutantes *T. mobilis* DidA y *T. mobilis* KR3DidA

La producción de didemninas de las cepas *T. mobilis* DidA y *T. mobilis* KR3DidA se realizó en medio LB a 30 °C y se extrajeron muestras a las 96 y 144 h de cultivo que se analizaron mediante HPLC-masas (cromatogramas no mostrados). Para ambos mutantes el resultado fue idéntico. A las 96 h de cultivo se identificaron pequeñas trazas de didemnina B, sin embargo a las 144 h no se detectó ninguna didemnina.

Queda patente para ambos mutantes, que la eliminación de los dos primeros módulos de síntesis de la molécula, se elimina casi por completo la capacidad para sintetizar didemninas. Se observa por lo tanto que la eliminación de estos módulos iniciales no permite convertir el módulo 3 en módulo inicial eficaz, y por lo tanto no se generan otros análogos teóricamente esperados como la aplidina en el caso del doble mutante KRBDidA.

1.3. | Clonaje y monitorización de la expresión y la producción del clúster de síntesis de didemninas en un hospedador heterólogo

1.3.1. | Generación de genotecas a partir del DNA genómico de *T. mobilis* MES-10-09-028

Con el objetivo de trasladar la capacidad de síntesis didemninas a otras bacterias se decidió clonar el cluster productor de la cepa MES-10-09-028 en un vector de clonación/expresión que admitiese fragmentos de gran tamaño y que además fuese de amplio espectro de hospedador. Debido a la gran longitud del clúster, se decidió evitar un abordaje que se basara en el ensamblaje de pequeños fragmentos de DNA procedentes de amplificaciones de PCR y por ello se procedió a realizar genotecas mediante una clonación directa a partir del genoma bacteriano. Estas genotecas se realizaron en primer lugar utilizando fósmidos y posteriormente con BACs (ver Materiales y Métodos apartado 7).

1.3.1.1. | Generación de una genoteca de fósmidos a partir del DNA genómico de *T. mobilis* MES-10-09-028

Tal y como se indica en Materiales y Métodos apartado 7.1, una vez empaquetados en las cabezas de fago lambda, los productos de la ligación entre el DNA genómico de la cepa MES-10-09-028 y el vector, se infectó la cepa de *E. coli* Replicator FOS. A continuación se tituló la genoteca conteniendo aproximadamente 17000 clones, por lo que teniendo en cuenta que el tamaño del genoma de la cepa MES-10-09-028 es de aproximadamente 6,2 Mb, y que el tamaño de inserto medio de los fósmidos es de 40 kb, se calculó que el genoma de MES-10-09-028 estaría representado unas 110 veces.

1.3.1.2. | Rastreo de fragmentos pertenecientes al clúster productor de didemninas en la genoteca de fósmidos

Para comprobar la presencia de fragmentos pertenecientes al clúster de síntesis de didemninas en la genoteca se diseñaron 3 parejas de oligonucleótidos que hibridaban en distintas zonas del clúster para poder detectar producto de PCR en el caso de que el fragmento deseado estuviese presente. La primera pareja de cebadores (SMD13F y SMD13R) (ver tabla 3) amplifican un fragmento de 373 bp situado en el módulo 13 del clúster (último módulo del cluster). La segunda pareja (SMD8F y SMD8R) (ver tabla 3) amplifican un fragmentode 360 bp del módulo 8 (zona central del clúster). La tercera pareja (SMD2F y SMD2R) (ver tabla 3) hibrida en el otro extremo del clúster, en concreto en el tercer módulo de síntesis. De este modo, se obtendrían amplificaciones pertenecientes a tres zonas distantes que cubrirían todo el clúster (Fig. 28).

A continuación se llevaron a cabo reacciones de PCR utilizando grupos de 10 colonias que previamente habían sido cultivadas en placa. En los casos en que un grupo dio un resultado positivo, las colonias del grupo se analizaron individualmente. Con el objetivo de organizar los clones analizados, cada uno se nombró dependiendo de la ronda de análisis, el grupo al que pertenecía y su posición dentro de dicho grupo. Para detectar los módulos 8 y 13 se analizaron un total de 600 clones, lo que supone recorrer el genoma 3,9 veces. De este modo se obtuvieron 5 clones positivos para el módulo 13 y 2 para el módulo 8. Para detectar el módulo 3 se analizaron un total de 300 clones, lo que supondría recorrer el genoma casi 2 veces y se obtuvieron un total de 6 clones positivos, lo que estadísticamente resulta poco probable. Por lo tanto, para eliminar posibles falsos positivos, estos clones también fueron verificados utilizando la pareja de cebadores DidBKR3A5' y DidBKR3A3', diseñada para amplificar una zona de 805 bp también en el módulo 3. En este caso únicamente 4 de los clones fueron positivos, lo cual concordaba más con lo esperado.

Posteriormente, y para comprobar la amplitud del fragmento clonado en cada uno de los clones positivos, se realizaron amplificaciones mediante PCR con el resto de los oligonucléotidos diseñados. De este modo se determinó que algunos clones resultaron positivos para más de una de las sondas detectadas. Estos clones, que probablemente contendrían fragmentos más completos del clúster contenían los fósmidos 1B6, 1W5, 4P7 y 3H4. Todos los clones mencionados resultaron positivos para dos de las sondas utilizadas en este proceso de rastreo excepto el fósmido 1W5 el cual únicamente resultó positivo para la sonda SMD13 (Fig. 28).

Para concretar la zona exacta del clúster que abarcaban los fósmidos 1B6, 1W5, 4P7 y 3H4, se secuenciaron los extremos de los insertos. Los resultados de la secuenciación mostraron que el clon F1B6 contiene gran parte del clúster ya que contiene parte del módulo 4 y el final del módulo 13. Además, este clon incluye ORFs adyacentes a la transposasa del extremo 3' del clúster (Fig. 28). Por otro lado, el clon F1W5 contiene muy poca secuencia perteneciente al núcleo PKS/NRPS y se extiende desde el módulo 13 hasta 36 kb hacia el extremo 3´de la última ORF del núcleo del clúster. Los clones F3H4 y F4P7 solapan entre el módulo 8 y 9 y entre los dos contienen la totalidad de los genes pertenecientes al núcleo PKS/NRPS y gran parte de los genes adyacentes que pueden ser considerados parte del clúster (Fig. 28).

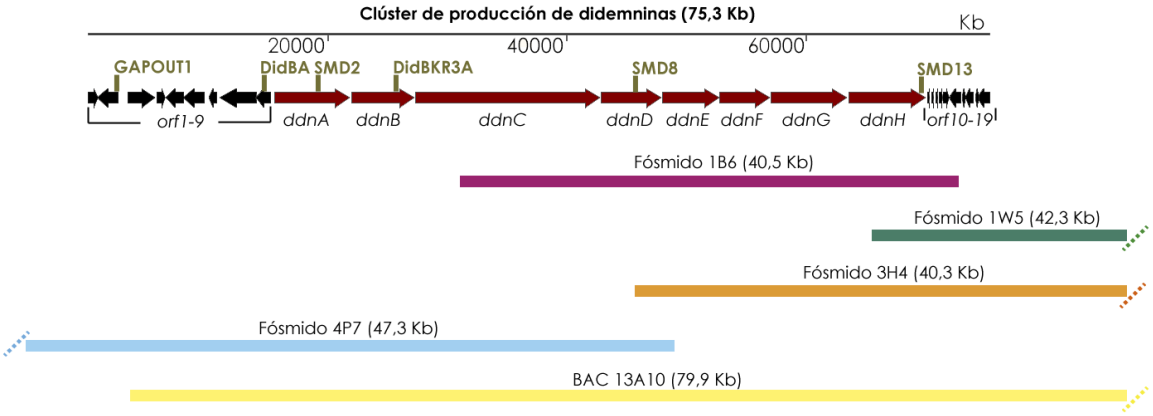

Figura 28 | Esquema de los fragmentos del clúster de producción de didemninas obtenidos en los distintos clones. Los genes del clúster correspondientes al núcleo PKS/NRPS se representan en rojo mientras que los genes de las zonas adyacentes en negro. En dorado se representa la posición de los oligonucleótidos utilizados para la detección de fragmentos pertenecientes al clúster en el rastreo de los clones. Cada uno de los fragmentos obtenidos clonados en las genotecas se ha representado en función a la zona del clúster de producción de didemninas que cubre su secuencia. Las líneas discontinuas indican que el fragmento continúa más allá de la zona delimitada del clúster.

Por tanto, mediante los clones aislados se ha conseguido obtener la secuencia de la totalidad de los módulos que conforman el núcleo PKS/NRPS del clúster de producción de didemninas. En particular, los clones solapantes F3H4 y F4P7 contienen toda la secuencia del cluster, incluyendo los genes incluidos en el entorno.

1.3.1.3. | Generación de una genoteca de BACs a partir del DNA genómico de *T. mobilis* MES-10-09-028

Tal y como se detalla en Materiales y Métodos, apartado 7.2, se generó una genoteca de 830 clones con insertos mayores de 200 bp, lo que correspondería aproximadamente a una cobertura de 26x del genoma de la cepa MES-10-09-028. A continuación se construyó otra genoteca de 3440 clones de insertos de entre 100 y 200 kb, que proporcionó una cobertura de 55x del genoma. El tamaño de los insertos se verificó para un número de clones representativos para ambas genotecas, resultando éstos en todos los casos mayores de 100 kb.

1.3.1.4. | Identificación de BACs con fragmentos del clúster productor de didemninas

Para realizar el rastreo de los clones se realizaron amplificaciones por PCR de grupos de 10 colonias. Para comprobar la presencia del extremo 3' del núcleo PKS/NRPS

del clúster (módulo 13) se utilizaron los cebadores SMD13F y SMD13R (373 bp) (ver tabla 3). Además para amplificar la zona del promotor del gen *ddnA* se utilizó la pareja de cebadores DidBAF y DidBAR (814 bp) (utilizados también para generar el mutante DidA) (ver tabla 3). Por último, para amplificar la región 5' de las ORFs que flanquean el núcleo del clúster se utilizaron los cebadores GAPOUT1F y GAPOUT1R (772 bp) (ver tabla 3), que hibridan a 12 184 bp del codón de inicio del gen *ddnA*. De este modo se analizaron un total de 1800 clones, detectándose un único clon positivo (para todos las parejas de oligonucleótidos excepto GAPOUT1F-R) que se le denomino B13A10.

Para comprobar la zona que abarca el clon B13A10 se secuenciaron los extremos del inserto y se pudo comprobar que contiene un fragmento de 80 kb que va desde la posición 459 299 a la 539 288 del genoma, estando incluidas las ORFs adyacentes al núcleo PKS/NRPS en el extremo 5'. En concreto contiene parte de la *orf3*, y termina a 14 kb del fin del núcleo PKS/NRPS en el extremo 3' (Fig. 28).

Por lo tanto, el clon B13A10 contiene el núcleo PKS/NRPS del clúster de síntesis de didemninas, además de gran parte de las ORFs adyacentes que podrían estar asociadas a dicha síntesis. Se debe tener en cuenta, que en este fragmento los genes *orf1 y orf2* no están presentes y además el gen *orf3* estaría incompleto. Como ya se ha comentado (ver Resultados apartado 1.1.11.), el gen *orf1* codifica una posible fosfopanteteinil transferasa, que podría tener un papel importante para la activación de la producción del clúster.

1.3.2. | Análisis de la expresión y la producción heterólogas del clúster productor de didemninas

Una vez se consiguió clonar una gran parte del clúster productor de didemninas de forma heteróloga en *E. coli*, se decidió analizar la expresión del clúster para valorar las posibilidades de obtener producción de didemninas en durante su cultivo.

1.3.2.1. | Monitorización de la expresión heteróloga del clúster de síntesis de didemninas

El primer paso para abordar la producción heteróloga de didemnina es detectar si los genes responsables de la síntesis se están expresando en la bacteria receptora. De este modo se comprobó la expresión heteróloga del clúster productor de didemnina B de la cepa *E. coli* B13A10.

1.3.2.1.1. | Expresión heteróloga del clúster productor didemninas en la cepa B13A10 de *E. coli*

La expresión de los genes del clúster se comprobó mediante RT-PCR semicuantitativa. Para ello se realizó un cultivo de la cepa en matraz a 37 °C y 250 rpm utilizando el medio rico LB en presencia de arabinosa para inducir el número de copias del vector (ver Materiales y Métodos). A continuación, se extrajo el RNA del cultivo de *E. coli* B13A10 en fase exponencial de crecimiento (DO_{600nm}=0,8). Seguidamente se realizó la correspondiente retrotranscripción a cDNA y posteriormente la amplificación por PCR. A diferencia de con la cepa MES-10-09-028 (ver Resultados apartado 1.1.12.), en este caso se utilizó la pareja de cebadores que amplificaban la zona del gen *ddnB* (SCDidBF y SCDidBR)

obteniéndose amplificación (Fig. 29). Este resultado sugiere que el clúster productor de didemnina se está expresando en la cepa B13A10 de forma heteróloga en *E. coli*.

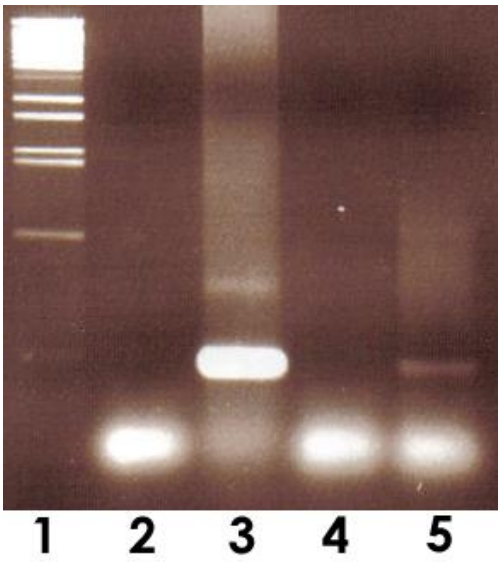

Figura 249 | Gel de la RT-PCR semicuantitativa del gen *ddnB*. La amplificación se realizó con la pareja de cebadores SCDidBF y SCDidBR. **1:** λ *Bst*EI. **2:** Control negativo sin DNA. **3:** Control positivo de amplificación con DNA del BAC 13A10. **4:** Control negativo de retrotranscripción (muestra procesada sin retrotranscriptasa). **5:** cDNA de *E. coli* B13A10.

1.3.2.2. | Producción heteróloga de didemninas en *E. coli*

Una vez comprobado que existe expresión heteróloga de los genes del núcleo PKS/NRPS se decidió comprobar la producción didemninas en *E. coli* B13A10.

1.3.2.2.1. | Ensayos de producción de didemninas en *E. coli*

Se realizaron cultivos en matraz de 40 mL de cultivo en medio rico LB de la cepa B13A10 a 37 °C y 200 rpm. Se realizaron extracciones tanto de las células como de los sobrenadantes del cultivo a diferentes tiempos (24, 48, 72 y 96 h) tal y como se describe en materiales y métodos, y se analizaron por HPLC-MS. Sin embargo y a diferencia de lo que ocurría en la cepa MES-10-09-028, no se detectó la presencia de ninguna didemnina en ninguno de los tiempos de cultivo para ninguna de las n las fracciones ensayadas (cromatogramas no mostrados). De estos resultados se puede concluir que aunque aparentemente los genes se estén expresando la producción no resulta posible en estas condiciones.

1.3.2.3. | Análisis de la producción de didemninas en *E. coli* B13A10 expresando la actividad PPTasa de la *orf1* de la cepa MES-10-09-028

Tal y como se ha comentado en la introducción la actividad PPTasa es esencial para la producción de compuestos naturales ya que actúa sobre las proteínas portadoras de acilo (ACP) transformando la forma inactiva apo a la forma catalíticamente activa *holo*. Aunque en *E. coli* existe un gen que codifica una PPTasa, ésta podría no ser muy eficaz en este tipo de clústeres debido a que presenta muy poca promiscuidad de sustrato (Li y Neubauer, 2014; Lambalot *et al*., 1996) Por tanto el hecho de que no se detectara

presencia de didemninas en la cepa B13A10 podría deberse a una falta de función PPTasa específica para este tipo de clústeres (Ongley *et al*., 2013a). Por esta razón, se decidió construir una cepa de *E. coli* B13A10 en la cual se expresase de forma inducible el único gen que codifica una enzima con una función fosfopanteteinil transferasa en el genoma de la cepa MES-10-09-028.

1.3.2.3.1. | Clonaje de la *orf1* en el plásmido pSEVA224 en la cepa *E. coli* 13A10

El gen *orf1*, que codifica una posible PPTasa en *T. mobilis* MES-10-09-028 se amplificó utilizando los cebadores TistPPTF y TistPPTR. El uso del cebador TistPPTF añade en el extremo 5' la secuencia shine-dalgarno del gen de resistencia a kanamicina del vector pSEVA224, la cual se ha comprobado que es funcional tanto en *E. coli* como en la propia *T. mobilis* MES-10-09-028. El cebador TistPPTR incorpora un fragmento de DNA que contiene la secuencia posterior al codón de STOP del gen, para así poder mantener posibles terminadores El producto de PCR de 1090 bp, fue purificado y digerido con las enzimas *Hin*dIII y *Xba*I y se ligó con el vector pSEVA224 previamente digerido con las mismas enzimas. El plásmido resultante pSEVAPPT se transformó en *E. coli* B13A10 por electroporación generándose la cepa *E. coli* B13A10PPT.

1.3.2.3.2. | Análisis de la producción de didemninas en la cepa *E. coli* B13A10PPT

Al igual que con la cepa B13A10, se analizó la producción de didemninas de la cepa *E. coli* B13A10PPT. Para ello se realizaron cultivos de 40 mL en matraz a 30 °C y 200 rpm, y se obtuvieron los extractos correspondientes a las 24, 48, 72 y 96 h, tanto para la fracción celular como para la extracelular. Dichos extractos fueron analizados por HPLC-MS y de nuevo no se observó la presencia de ningún pico correspondiente a didemninas en las muestras analizadas (cromatogramas no mostrados).

A la vista de estos resultados, se puede decir que muy probablemente la razón por la cual no se detecte producción heteróloga en la cepa 13A10 no sea una falta de actividad PPTasa necesaria para activar el núcleo del clúster, y que quizás esta causa esté más ligada a fenómenos relacionados por ejemplo, con las moléculas precursoras, la correcta conformación de los complejos macroproteicos u otras deficiencias funcionales críticas que este sistema pueda presentar.

2. Utilización de herramientas metagenómicas para la identificación de secuencias de clústeres de síntesis de metabolitos secundarios en cepas no cultivables

2.1. | Secuenciación y análisis del metagenoma de *Polymastia littoralis* para la obtención secuencias de interés que codifican PKSs y NRPSs

2.1.1. | Análisis metagenómico del microbioma

La esponja marina *P. littoralis* fue descrita por primera vez por Jane Stephens en 1915 en especímenes encontrados en Sudáfrica (Stephens, 1915) El género *Polymastia*, que actualmente cuenta con 30 especies identificadas, es un género de esponjas marinas que pertenece a la familia *Polymastiidae* (clase *Demospongiae*, subclase *Tetractinomorpha* y orden *Hadromerida*). En esta esponja PharmaMar detectó la presencia de un compuesto con actividad antitumoral mediante HPLC/MS. En concreto, la muestra con la que se trabajó fue en el Océano Índico africano y fue congelada inmediatamente a -20 °C. La compleja estructura molecular de este compuesto de 58 carbonos sugirió que este podría ser sintetizado por un clúster PKS/NRPS bacteriano de aproximadamente 18 módulos de síntesis de tipo PKS y un único módulo de tipo NRPS. Dado que no se consiguió aislar ninguna bacteria cultivable capaz de producir este compuesto nos planteamos la búsqueda del clúster de PKS/NRPS responsable de su síntesis realizando un análisis de la secuencia metagenómica de la fracción microbiana extraída de la esponja.

2.1.1.1. | Aislamiento de la fracción microbiana

Tal y como se ha comentado en la Introducción, teóricamente los clústeres PKS/NRPS de síntesis de metabolitos secundarios pueden encontrarse en la fracción microbiana de la muestra, en concreto en la fracción bacteriana. Por lo tanto, para poder analizar el metagenoma presente en la fracción microbiana de la esponja *P. littoralis* se llevó a cabo un proceso de aislamiento de dicha fracción. Para ello se partió de 33 g de muestra de esponja y se siguió el protocolo de aislamiento del microbioma que se indica en la sección de Materiales y Métodos. Con la fracción microbiana resultante se realizó una extracción de DNA genómico (Materiales y Métodos) obteniéndose una fracción de 5 µg de DNA purificado que fue utilizado para ser secuenciado mediante técnicas de secuenciación masiva.

2.1.1.2. | Secuenciación de la fracción microbiana

La secuenciación del DNA metagenómico de la fracción microbiana aislada de *P. littoralis*, se llevó a cabo mediante pirosecuenciación. La secuencia obtenida estaba distribuida en un total de 371 533 lecturas con una longitud media de 550 bps (Fig. 30) abarcando un total de 204,5 Mb. El contenido medio en GC del DNA metagenómico resultó ser del 47% (Fig. 30).

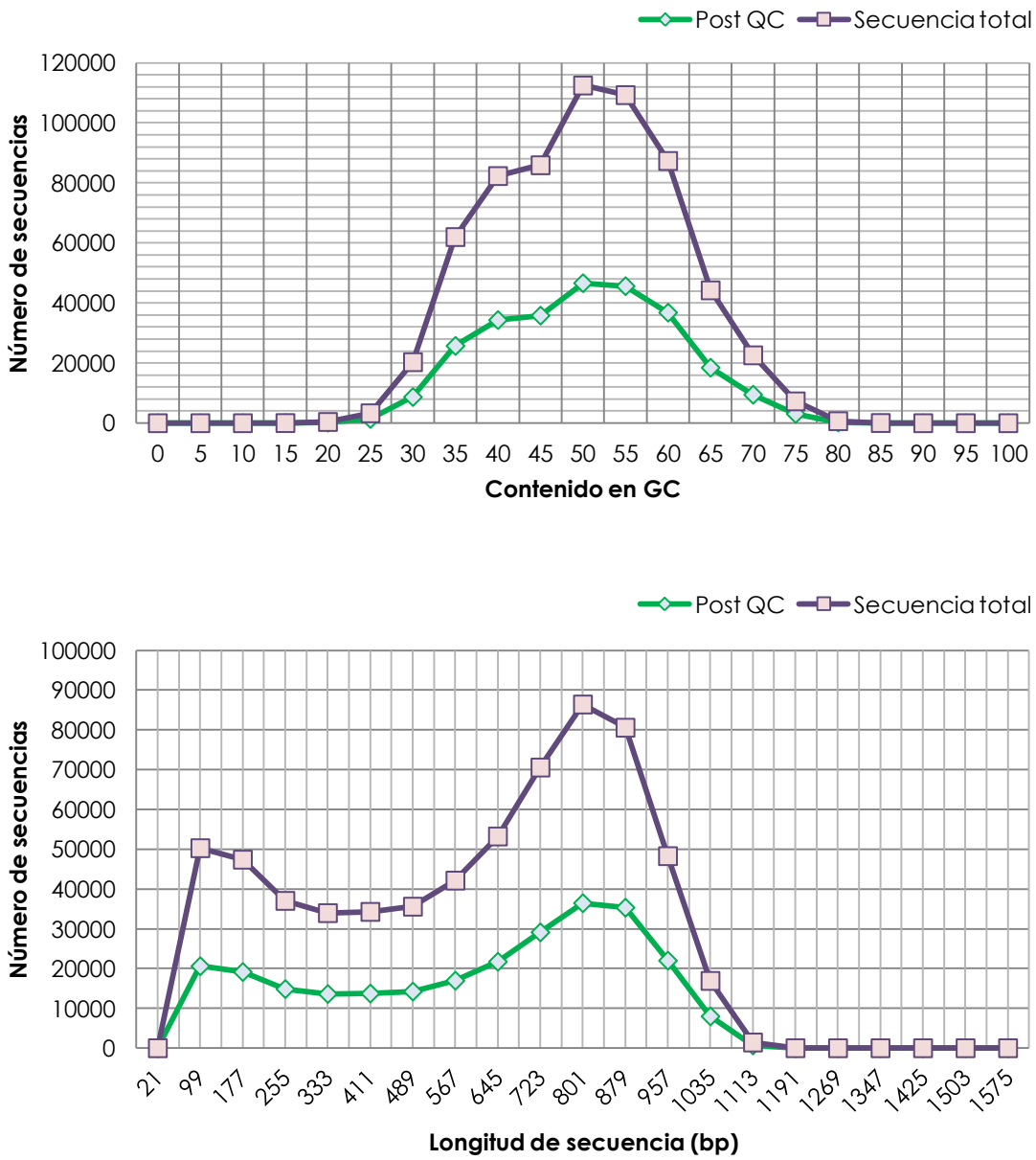

Figura 30 | Distribución de las secuencias del metagenoma de *P. littoralis* según su longitud y contenido en GC. En morado se muestra la secuencia obtenida directamente de la secuenciación masiva mientras que en verde las secuencias que han pasado el control de calidad (QC) del servidor MG-RAST. En el gráfico superior se observa la distribución del número de secuencias según su porcentaje del contenido en G+C mientras que en el inferior se puede ver la distribución del número de secuencias obtenidas según la longitud.

2.1.1.3. | Análisis de la secuencia del metagenoma microbiano en MG-RAST

Para realizar un análisis general de las secuencias obtenidas del metagenoma se utilizó el servidor especializado MG-RAST (Meyer *et al.*, 2008). La ventaja de usar esta herramienta radica en que este servidor puede trabajar con secuencias desde 75 bp, y se adapta bien a las lecturas que se suelen obtener tras realizar una secuenciación masiva. Es necesario tener en cuenta, que el límite de la precisión de la información obtenida normalmente es dependiente de la longitud de secuencia, por lo tanto, los análisis realizados mediante este tipo de herramientas, son de índole general y resultan útiles para

evaluar características del metagenoma en su conjunto y son menos fiables a la hora de apreciar detalles puntuales.

2.1.1.3.1. | Resultados de la asignación funcional de secuencias del metagenoma

En primer lugar, siguiendo el protocolo de trabajo de MG-RAST, se realizó un procesamiento previo de las secuencias atendiendo su calidad. Del total de las lecturas un 28,2% (104859 secuencias) no pasaron las restricciones de calidad (Fig. 31). Del conjunto de las secuencias que sí superaron dicho procesamiento, el 94% fueron detectadas como portadoras de secuencias codificantes, en concreto se detectaron un total de 226999 características proteicas. El 4,2% correspondían a secuencias de genes de RNA ribosómico y el resto podrían ser secuencias no codificantes. Del conjunto de todas las características proteicas detectadas, el 40,1% fueron anotadas, mientras que el 59,9% restante no poseían similitud significativa con otras proteínas de las bases de datos. De todas las secuencias anotadas el 77,5% (70579 características proteicas) fueron asignadas a categorías funcionales (Fig. 31).

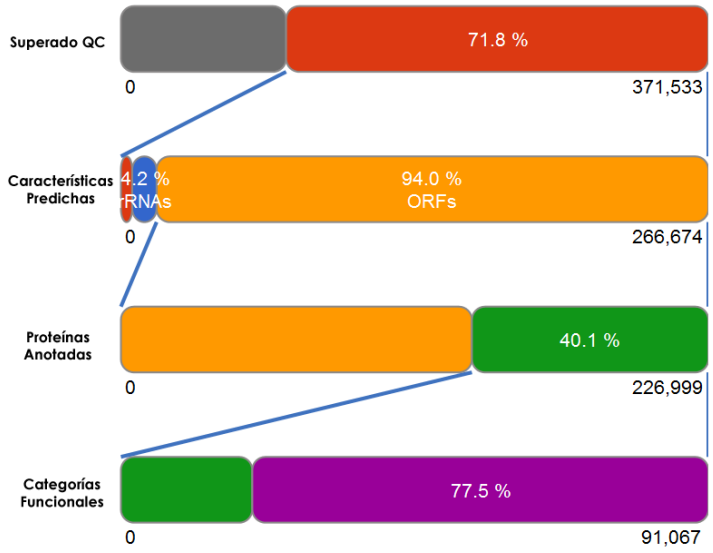

Figura 31 | Esquema del proceso de anotación en MG-RAST. En el esquema se puede observar la cantidad de secuencias procesadas en los pasos de control de calidad (QC), predicción de características, anotación de proteínas y asignación de categorías funcionales.

Como se puede observar en la distribución de funciones por subsistemas, los más representados, sin contar los grupos que recopilan categorías sin determinar, son aquellos relacionados con el metabolismo de los aminoácidos, de los carbohidratos y de las proteínas (Fig. 32). Además llama la atención el porcentaje relativamente alto de secuencias en el subsistema que agrupan funciones fágicas, plasmídicas y aquellas relacionadas con elementos transponibles.

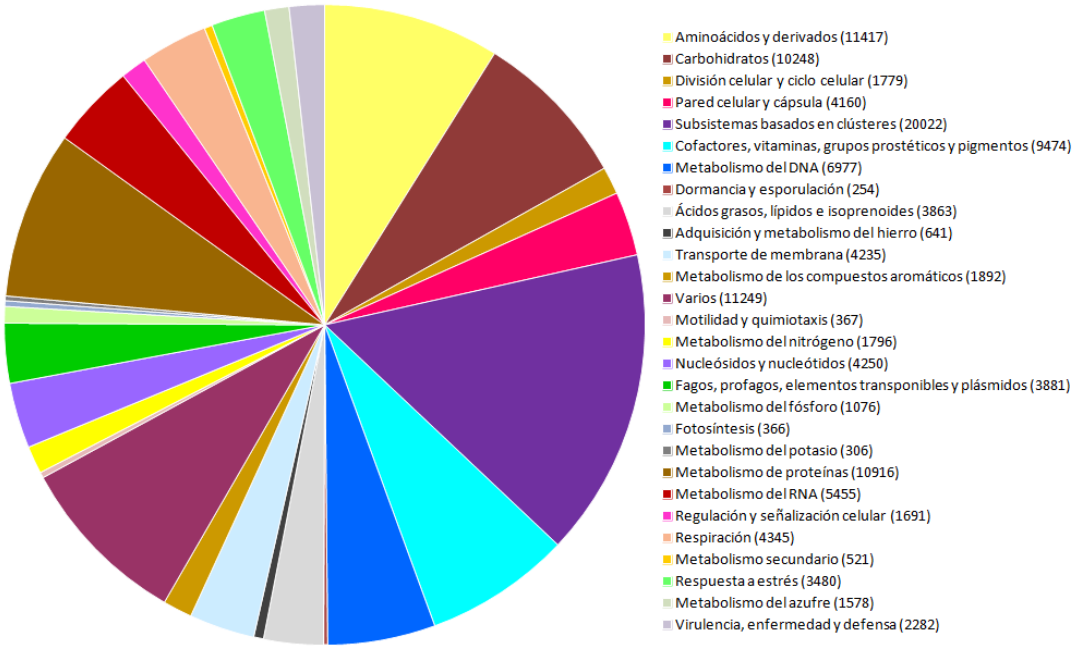

Figura 32 | Gráfica de la distribución por subsistemas de la anotación del metagenoma de *P. littoralis*. En el gráfico proporcionado por la herramienta MG-RAST se representan el número de positivos para cada uno de los subsistemas agrupados en categorías generales tras el proceso de anotación automática.

2.1.1.3.2. | Resultados de la asignación taxonómica de las secuencias metagenoma total

La herramienta MG-RAST es capaz de obtener datos taxonómicos de las secuencias a la vez que realiza las anotaciones de las mismas. De este modo se obtuvo la distribución de las secuencias ordenadas por dominios pudiéndose observar que la mayoría de las secuencias fueron clasificadas como bacterianas (83,9%) (Fig. 33). Sin embargo también existe una representación de DNA de arqueas cercana al 8%. Al profundizar más en el siguiente nivel de clasificación se obtuvo la distribución de la secuencia en los distintos phyla (Fig. 34), siendo el phylum proteobacteria el más representado en las secuencias, seguido aunque muy de lejos por Actinobacteria, Thaumarchaeota, Cyanobacteria y Firmicutes. Es interesante destacar que las secuencias víricas y eucarióticas presentes quedan parcialmente desclasificadas, ya que no se le asigna ningún phylum.

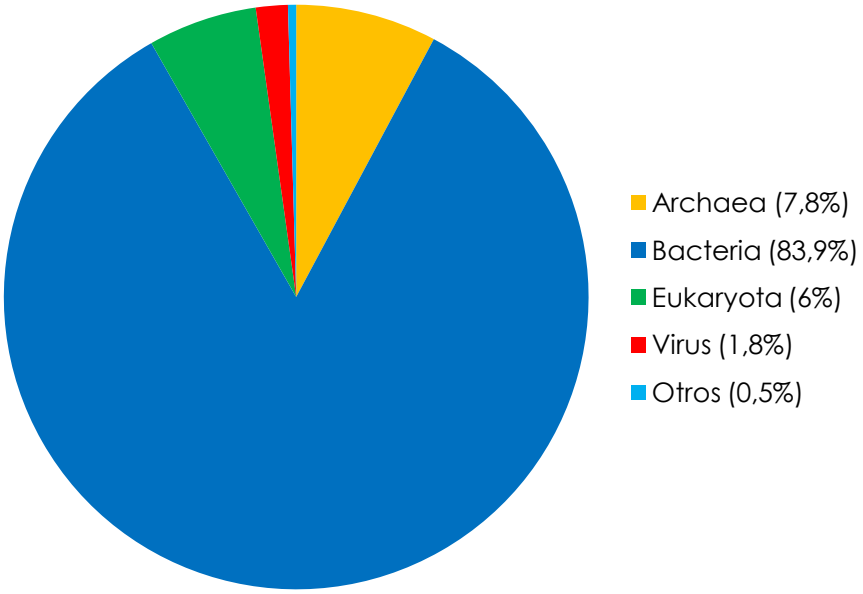

Figura 33 | Distribución por dominios de las secuencias del metagenoma de *P. littoralis*.

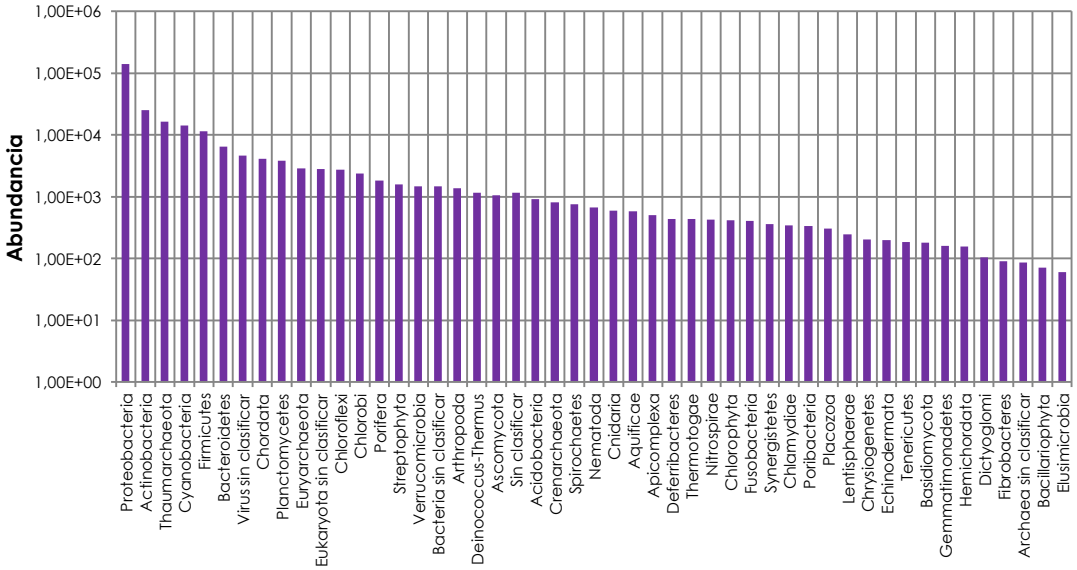

Figura 34 | Abundancia de cada phylum en las secuencias clasificadas de *P. littoralis*.

El número total de especies distintas que se identifican *a priori* en la muestra proviene de analizar la correspondiente curva de rarefacción calculada por MG-RAST (Fig. 35) y para el número de secuencias dado sería aproximadamente de 3500 especies diferentes.

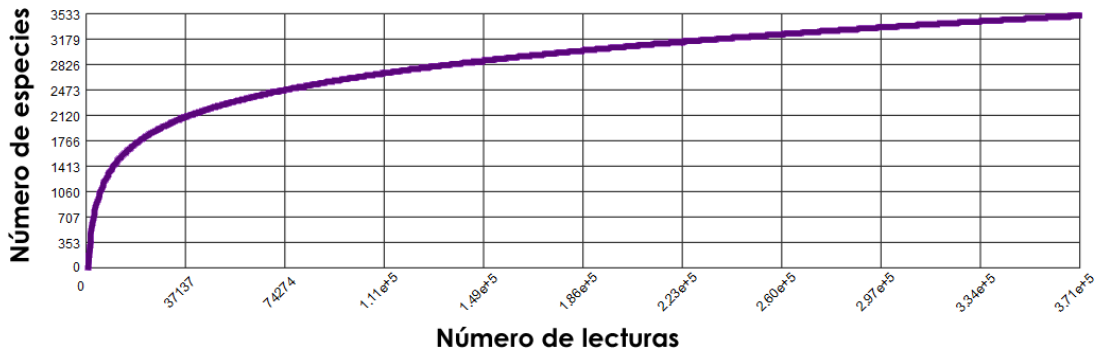

Figura 35 | Curva de rarefacción de las secuencias de *P. littoralis*. En la gráfica obtenida de MG-RAST se muestra el grado de saturación en especies de la muestra identificando el número de especies distintas conforme se van analizando las lecturas.

2.1.1.4. | Ensamblaje del metagenoma microbiano

A continuación se procedió a ensamblar la secuencia *de novo* en busca de aquellos fragmentos más representados y por lo tanto con probabilidad de ensamblarse más fácilmente. Para realizar dicho ensamblaje se realizaron pruebas con varios algoritmos ensambladores especializados para metagenomas, y en este caso el mejor resultado se obtuvo al emplear el software Newbler (ver Materiales y Métodos). Tras el proceso de ensamblaje se consiguió un total de 9042 *contigs*, que representan 12,5 Mb de secuencia. A continuación se realizó un mapeo de las lecturas totales frente al ensamblaje realizado utilizando bowtie2 y se observó que un 51,69% de las lecturas estaban representadas en la secuencia ensamblada. Esta fracción estaría formada por un número relativamente pequeño de individuos probablemente más abundantes.

2.1.1.5. | Asignación taxonómica del metagenoma ensamblado

Una vez ensamblada la secuencia, se obtuvieron fragmentos de DNA lo suficientemente largos como para realizar asignaciones taxonómicas con distintos métodos que ofrecen más precisión que los métodos descritos anteriormente. En concreto, utilizando variaciones del protocolo *in silico* desarrollado por Albertsen *et al*. (2013) (ver Materiales y Métodos) resulta posible extraer distintas características de cada una de las secuencias con el objetivo de clasificarlas atendiendo a estos criterios, por lo que de este modo se puede separar y agrupar las secuencias que posiblemente pertenezcan a un mismo individuo.

En primer lugar, de los *contigs* generados se extrajeron los datos correspondientes a la abundancia, mapeando las lecturas con cada una de las secuencias ensambladas y posteriormente calculando la densidad de lecturas en cada caso (ver Materiales y Métodos). A continuación se calcularon otros parámetros estructurales de la secuencia, como son el contenido en GC, la longitud y la frecuencia de tetranucleótidos. A continuación, se realizaron búsquedas de genes esenciales de copia única y estos se clasificaron taxonómicamente. Por último se llevó a cabo una clasificación taxonómica basada en estructura con el algoritmo PhyloPythiaS+ (Gregor *et al*., 2014) para cada secuencia (ver Materiales y Métodos).

Una vez extraídos todos los datos, se distribuyeron de forma gráfica aquellos *contigs* de mayor tamaño, atendiendo a las características antes mencionadas. Por ejemplo, en la figura 36 se puede observar cómo se distribuiría cada uno de los *contigs* mayores de 2500 bp en relación a su contenido en GC y su abundancia en el metagenoma. Además en esta figura se puede apreciar el tamaño de cada secuencia atendiendo al diametro del círculo que la representa. Esta representación hace posible diferenciar cúmulos de secuencias con las mismas características, los cuales podrían pertenecer probablemente a una misma especie. En la figura 36 se pueden apreciar 3 cúmulos principales y varios círculos de gran tamaño en la zona de las secuencias altamente representadas.

En primer lugar se analizó la secuencia de estos contigs de mayor tamaño con alta representación. Además se realizaron búsquedas mediante Blast, HMMER y la herramienta de detección de dominios de Pfam, para intentar clasificar el posible origen de estas secuencias.

Se realizó un análisis de aquellos contigs con más de 8000 bp que estuviesen representados con una abundancia relativa mayor de 10x y que a su vez no estuvieran incluidos dentro de las zonas pertenecientes a los 3 cúmulos de secuencias antes descritos. Como resultado, se puede observar en la figura (Fig.36), que la mayoría de secuencias son fragmentos de posibles fagos. Por la longitud de toda las secuencias de fagos que aparecen y por las distintas características que tienen, se podría decir que existen varios fagos distintos en el metagenoma. Se pueden apreciar también otros fragmentos pertenecientes a posibles elementos móviles como transposasas o incluso posibles plásmidos. Es interesante destacar la presencia de un único *contig* que aparentemente contiene una secuencia mitocondrial. Cabe destacar también que del total de los 12 contigs más largos (mayores de 19800 bp), la mayoría de ellos se encuentran áltamente representados y poseen secuencias de fagos, elementos móviles o secuencias mitocondriales. Tras realizar los análisis correspondientes se comprobó que únicamente 3 de estos *contigs*, contienen fragmentos genomicos bacterianos y de estos, únicamente el *contig* 7 se encuentra un poco desplazado con respecto a los cúmulos principales de secuencias, el resto se halla inmerso en estos clúster.

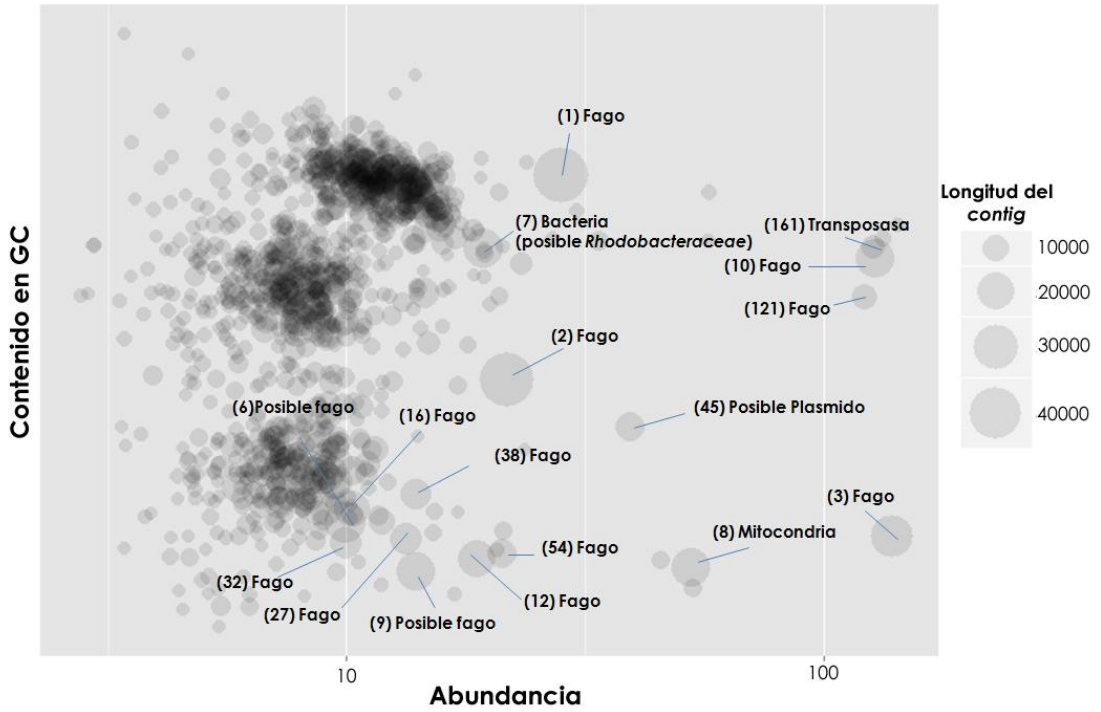

Figura 36 | Distribución de la secuencia ensamblada de *P. littoralis* y análisis de los *contigs* individuales y representativos. En la distribución se representan los *contigs* mayores de 2500 bp según su contenido en GC y abundancia relativa. El tamaño del punto es relativo a la longitud del *contig* (bp). Se muestran los resultados de asignaciones taxonómicas manuales con HMMER, y BLAST para aquellas secuencias no agrupadas en cúmulos principales, representadas más de un 10x y mayores de 8000 bp.

Una vez analizados aquellos *contigs* de gran tamaño y altamente representados, se procedió a superponer los datos obtenidos de la búsqueda de genes esenciales y su clasificación taxonómica (Fig. 37). Con este tipo de representaciones se analiza si los cúmulos de secuencia que se han descrito contienen secuencias con genes esenciales, un requisito imprescindible para poder comprobar si la agrupación de secuencias representa fragmentos de un mismo genoma, ya que un genoma completo poseerá genes esenciales clasificados taxonómicamente de una forma similar. En este caso, al tratarse de una muestra reducida de secuencias la clasificación de los genes esenciales se realizó usando algoritmos de comparación, los cuales requieren más capacidad de computación.

Al observar la representación de la figura 37 se puede apreciar la existencia de genes esenciales en el conjunto de las secuencias que forman cada una de las agrupaciones. Para este análisis se utilizaron secuencias ensambladas de más de 2500 bp de longitud. La agrupación que aparece en una posición superior en la figura 37 posee un contenido en GC aproximado entre el 50 y 60% y la cobertura relativa en ocasiones supera el 10x. Este cúmulo de secuencias parece representar fragmentos genómicos pertenecientes a proteobacterias, en concreto se puede observar la presencia mayoritaria de genes esenciales clasificados como pertenecientes a Gammaproteobacteria y algunos otros clasificados como Betaproteobacteria. En el cúmulo con una posición central en la figura 37 (con un contenido en GC aproximado de 45-50%) se puede observar la presencia mayoritaria de genes esenciales de Actinobacteria, sin embargo la cobertura media es menor que en el cúmulo superior. El cúmulo de la parte inferior de la figura, con un contenido en GC aproximado entre 30-40% parece contener genes esenciales clasificados

como pertenecientes a arqueas del phylum Thaumarchaeota y su cobertura media también es menor que en el cúmulo superior. Este análisis puede sugerir en primer lugar que no parece haber una bacteria o arquea especialmente dominante, sino que esta comunidad predominante podría estar repartida entre varias especies pertenecientes a los phyla Thaumarchaeota, Actinobacteria y Proteobacteria, siendo la más representada aquella o aquellas proteobacterias pertenecientes a la clase Gammaproteobacteria.

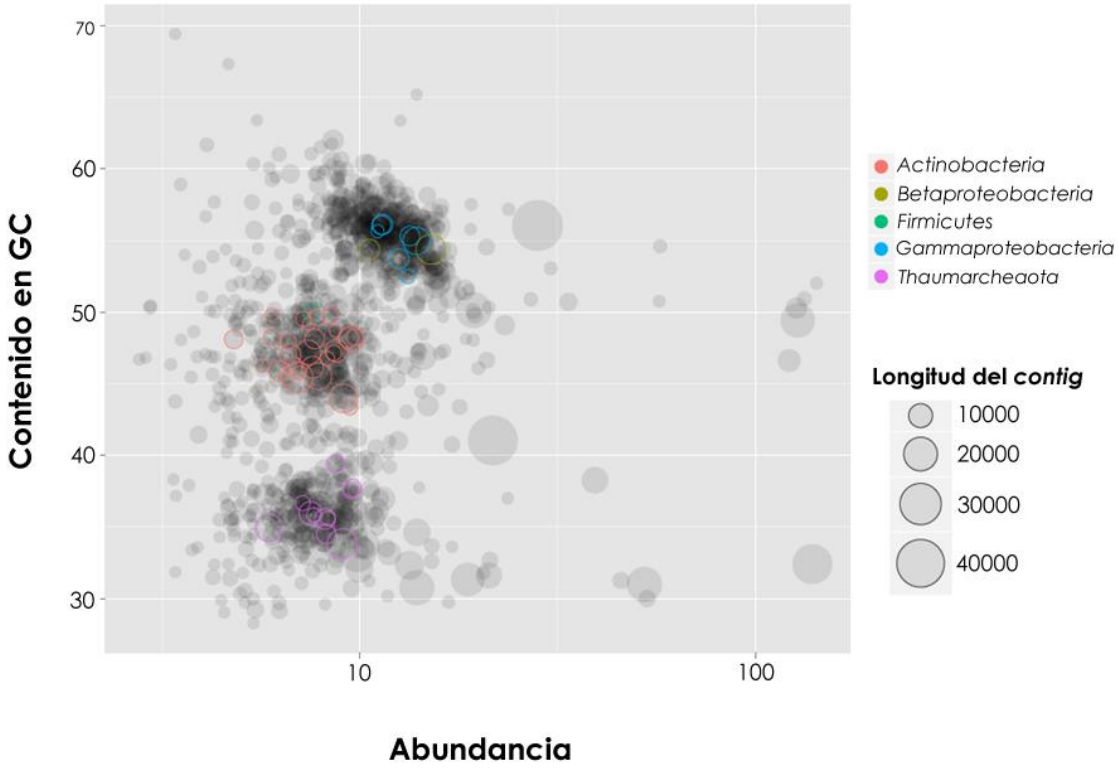

Figura 37 | Distribución de la secuencia ensamblada de *P. littoralis* y análisis taxonómico de genes esenciales. En la distribución se representan los *contigs* mayores de 2500bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). Dependiendo de la clasificación taxonómica obtenida se señalan en distintos colores aquellos *contigs* que contienen genes esenciales.

Para poder estimar el número aproximado de genomas individuales en la secuencia obtenida del metagenoma se necesitaría un análisis más exhaustivo, para así clasificarlos e intentar comprobar si dicho genoma se encuentra potencialmente completo. Para ello el siguiente análisis que se realizó fue una clasificación taxonómica de las secuencias metagenómicas completas mediante algoritmos basados en la comparación. Para llevar a cabo este trabajo se necesita una capacidad de computación de la que no se disponía en la realización de esta Tesis. Por lo tanto se decidió realizar este análisis utilizando algoritmos de clasificación basados en composición de secuencia (ver Materiales y Métodos). En concreto utilizando los datos generados por el método PhyloPythiaS+ se consiguió clasificar el conjunto de la secuencia (Fig. 38).

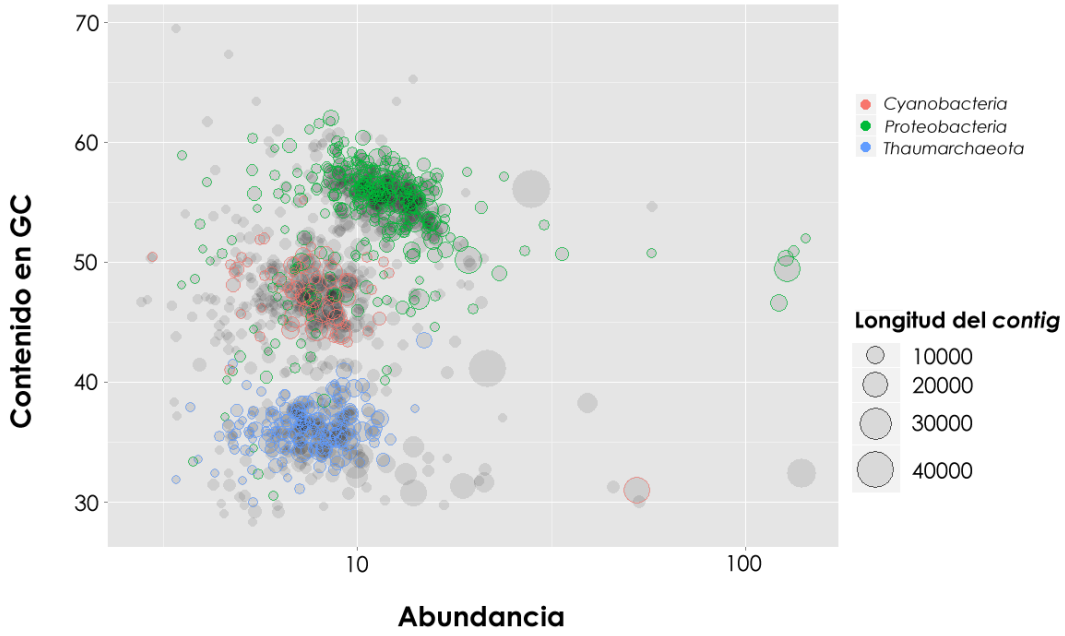

Figura 38 | Distribución de la secuencia ensamblada de *P. littoralis* y análisis taxonómico mediante PhyloPythiaS+. En la distribución se representan los *contigs* mayores de 2500 bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). La clasificación taxonómica obtenida para cada *contig* se señala en distintos colores. Aquellos *contigs* que aparecen sin asignación taxonómica no pudieron ser clasificados por la herramienta PhyloPythiaS+.

Como se puede valorar en la figura, cada uno de las tres acumulaciones se clasificó de forma distinta, y llama la atención que también algunos de los *contigs* más abundantes fueron incluidos en esta clasificación. Esto puede deberse a la identificación de los fragmentos de fagos como fragmentos genómicos comparables a posibles profagos que se encuentren en la secuencia genómica de bacterias con esa clasificación taxonómica. Los cúmulos superior e inferior fueron clasificados como secuencias de *Proteobacteria* y *Thaumarchaeota*, al igual que los genes esenciales que se detectaron en estas posiciones. Sin embargo, el cúmulo central no se clasificó igual que los genes esenciales que contiene. Esto puede deberse a estimaciones erróneas por parte del algoritmo de clasificación estructural, porque los fragmentos genómicos en esa acumulación no sean fácilmente comparables con las bases de datos que utiliza el algoritmo. Aunque la clasificación taxonómica propuesta no sea del todo precisa, este tipo de aproximaciones ayuda a separar las secuencias que probablemente pertenezcan a las mismas especies concretas.

2.1.2. | Búsqueda de secuencias de PKSs y NRPSs en el metagenoma

Para explicar la síntesis del compuesto antitumoral previamente detectado e identificado en los laboratorios de PharmaMar, se realizó una búsqueda especializada centrada en aquellos componentes típicos de los clústeres de síntesis de PKSs y NRPSs. De este modo se pretendía comprobar la efectividad de las herramientas metagenómicas para detectar este tipo de clústeres dependiendo de las características de abundancia de la población bacteriana.

2.1.2.1. | Búsqueda de secuencias de PKSs y NRPSs en el total de la muestra metagenómica

Inicialmente se decidió hacer un recuento total de funciones relacionadas con PKS y NRPS en el total de las secuencias del metagenoma. Para ello se realizaron búsquedas de PKS y NRPS con HMMER utilizando los modelos de la base de datos Pfam para una batería de dominios seleccionados. En concreto para realizar esta búsqueda se decidió utilizar los modelos de Pfam con los dominios Acil transferasa (AT), de unión a AMP (incluido el específico del C-terminal), condensación, ceto-reductasa (KR), cétido sintasa (KS) (N-terminal y C-terminal), de sitio de unión a fosfopanteteína (PP), dehidratasa (DH) y tioesterasa (TE).Todos ellos representan una gran parte de los dominios predominantes en este tipo de clústeres de producción. Se realizaron estas búsquedas tanto el conjunto de los *contigs* como en el de las lecturas y los resultados se pueden observar en la tabla 1. Además, para intentar eliminar parcialmente el efecto de la redundancia en las lecturas se realizó un proceso de *clusterización* (con un 98% de identidad) sobre todas aquellas lecturas positivas en la búsqueda por HMMER y se volvió a realizar la búsqueda (ver tabla 8).

Dominio	Nombre Pfam	*Contigs*	Lecturas	Cluster (98)
Acil-Transferasa (PF00698.16)	Acyl_transf_1	3	24	18
Unión a AMP (PF00501.23)	AMP-binding	60	748	495
Unión a AMP C-terminal (PF13193.1)	AMP-binding_C	23	130	101
Condensación (PF00668.15)	Condensation	1	12	14
KS N-terminal (PF00109.21)	ketoacyl-synt	24	217	143
KS C-terminal (PF02801.17)	Ketoacyl-synt_C	14	109	82
KR (PF08659.5)	KR	40	178	160
Sitio de unión de Fosfopanteteina (PP) (PF00550.20)	PP-binding	13	44	50
Dehidratasa (PF14765.1)	PS-DH	2	21	21
Tioesterasa (PF00975.15)	TE	1	3	3

Tabla 8 | Recuento de la representación de dominios relacionados con clústeres PKS/NRPS en la secuencia de *P. littoralis*. Para cada dominio se señala entre paréntesis el número de acceso correspondiente en la base de datos Pfam. Se representan los recuentos para la secuencia ensamblada (*contigs*) para las lecturas y para las lecturas *clusterizadas* con un 98% de identidad.

Como era de esperar, el número de positivos en el conjunto de las lecturas así como en las secuencias *clusterizadas* fue mayor que en los *contigs*. También se observa claramente que al *clusterizar* se consigue eliminar la redundancia de las lecturas obteniéndose un número más fiable para calcular la proporción de positivos que se encuentran en el total de las secuencias del metagenoma en relación con los *contigs* en cada uno de los casos. En el caso de los dominios de unión a fosfopanteteina y los de condensación, se observa que aumenta el número de positivos detectados. Esto se debe a que al variar el tamaño de la base datos de secuencias tras *clusterizar*, algunos de los positivos que previamente no cumplían los criterios de restricción por *e-value*, ahora sí los cumplen. El dominio de unión a AMP es el que se encuentra más representado probablemente debido a que es posible encontrarlo realizando funciones no directamente

relacionadas con los clústeres de tipo NRPS. Así, se puede concluir que aparecen más positivos para los dominios buscados en el conjunto total de la secuencia y que aparentemente la secuencia considerada como perteneciente a microorganismos no mayoritarios también contiene este tipo de dominios. Sin embargo esta fracción de secuencias positivas no ensambladas no son lo suficientemente largas para poder extraer la información necesaria y así poder situarlas en el contexto de un clúster de síntesis de metabolitos secundarios.

2.1.2.2. | Búsqueda de secuencias de PKSs y NRPSs en la muestra metagenómica ensamblada

La búsqueda de clústeres PKS/NRPS en las secuencias ensambladas posee la principal ventaja de que en estos casos las secuencias son más largas por lo que normalmente es posible extraer información del entorno para determinar si el fragmento en cuestión puede pertenecer o no al clúster de interés. Además, al haberse conseguido detectar inicialmente una molécula de interés en la esponja, una suposición lógica podría ser que el microorganismo productor debe hallarse en gran cantidad siendo mayoritario en la población y por lo tanto podría detectarse con más facilidad utilizando este tipo de herramientas.

De este modo, en primer lugar se identificaron todos aquellos *contigs* con una longitud mayor a 1000 bp en las secuencias resultantes de la búsqueda de dominios específicos realizada en el apartado anterior (ver tabla 8). A continuación se analizaron los contextos genómicos de cada secuencia para así seleccionar todos aquellos que pudieran pertenecer a un clúster PKS/NRPS. Para ello se realizaron búsquedas en Pfam y se utilizó la herramienta BLAST en el NCBI (ver Materiales y Métodos), por lo que finalmente se identificaron manualmente aquellos fragmentos que tienen un entorno génico que coincide con una posible pertenencia a un clúster PKS/NRPS. El criterio principal para seleccionar las secuencias se basó en que estas pudieran pertenecer a ORFs multifuncionales. Por ejemplo, en esta búsqueda se desecharon una gran cantidad secuencias que contenían dominios de unión a AMP por el hecho de que este se encontrase aislado en una única ORF.

Una vez seleccionadas, se intentó extender la longitud de cada una de las secuencias realizando un ensamblaje manual haciendo BLAST con los extremos 5’ y 3’ contra el conjunto de las lecturas. Como resultado de este proceso se seleccionaron 8 secuencias a las que se añadió una novena (secuencia 9), que se identificó manualmente como secuencia de interés por portar un dominio tioesterasa (TE), el cual resultó ser muy escaso entre los positivos de la búsqueda. Aunque esta secuencia proviene de un *contig* menor de 1000 bp, el fragmento fue ensamblado y extendido hasta 4343 bp (ver tabla 9).

Secuencia	*Contigs*	Dominios	Longitud (bp)
1	2696/2253	Methyltransf_18/Abhydrolase_6/PP-binding/C	4060
2	1298	PS_DH/KR	3281
3	1266	KS/KR/PP/KS/MT	5959
4	2897	PS_DH/KR/PP/KS	2997
5	2038	adh_short/PP/KS	1747
6	1839/2059	Methyltransf_23/PP/KS/KR	4641
7	1082/1767	ECH/PP/KS/PS_DH/KR/PP	5123
8	1146	Acyl_transf_1/E1_dh/Acyl_transf_1/Transket_pyr/ACPS	5688
9	5960	TE/HMG_CoA_synt/ECH	4343

Tabla 9 | Secuencias seleccionadas con dominios relacionados con PKSs y NRPSs. Para cada secuencia se indica los *contigs* que la contienen, la sucesión de dominios encontrada (con una notación correspondiente a la de la base de datos Pfam) y su longitud.

Entre las secuencias seleccionadas, la gran mayoría son de tipo PKS. Únicamente la secuencia 1 podría pertenecer a un clúster de tipo NRPS al tener un dominio de condensación, mientras que la secuencia 9 podría pertenecer a ambos tipos. Al comparar el tamaño de las secuencias obtenidas con aquellas de mayor longitud se puede pensar que es probable que ninguno de estos fragmentos pudiera pertenecer a un clúster de alguno de los microorganismos más mayoritarios.

Para comprobar si varias de estas secuencias pertenecen al mismo microorganismo o bien a microorganismos diferentes, se realizó un análisis similar al que se utilizó para detectar la presencia de genes esenciales. Para ello se seleccionaron aquellos *contigs* que aparecen en la tabla 9 y se clasificaron taxonómicamente por comparación (BLAST). A continuación se representaron junto al resto de *contigs* con la excepción del *contig* 5960, cuyo pequeño tamaño impide la correcta visualización de los datos (Fig. 39).

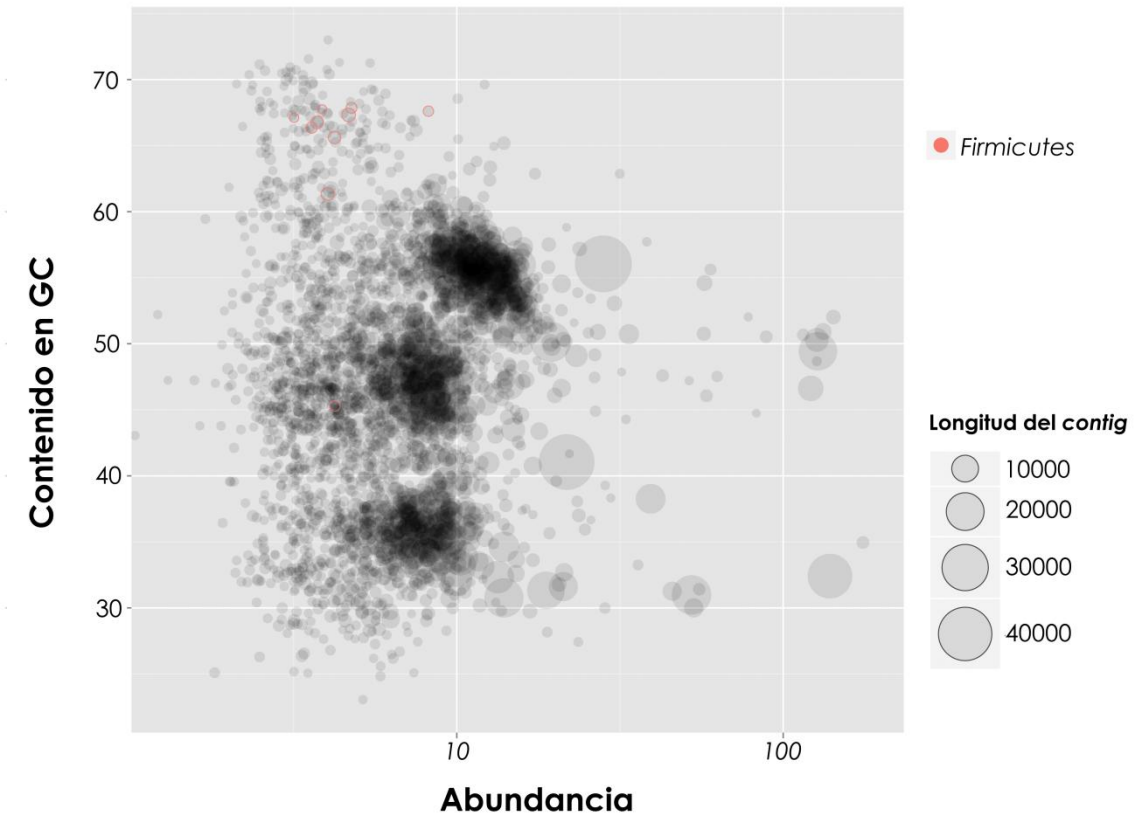

Figura 39 | Distribución de los *contigs* seleccionados con fragmentos con dominios PKS/NRPS en el conjunto de la secuencia ensamblada de *P. littoralis*. En la distribución se representan los *contigs* mayores de 1000 bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). La clasificación taxonómica obtenida para los contigs que contienen dominios PKS/NRPS se señala con colores.

Como se puede observar en la representación de la figura 39, las secuencias seleccionadas poseen cobertura relativa similar. Además la mayoría de ellas poseen un contenido en GC similar de entre 60-70%. Únicamente la secuencia más pequeña, la correspondiente al *contig* 2038, parece tener un contenido en GC entre 50 y 60%. Sin embargo, si atendemos a la clasificación taxonómica, todas ellas pertenecerían al phylum *Firmicutes*. Esto y el hecho de que la mayoría de las secuencias se agrupen en el mismo lugar de la representación, hace pensar que al menos las que se encuentran localizadas de forma similar puedan pertenecer al mismo microorganismo, y por lo tanto, posiblemente al mismo clúster biosintético. Se observa también como dicho clúster se encuentra menos representado que las secuencias de los tres cúmulos principales, por lo que si además tenemos en cuenta que no hemos podido ensamblar la secuencia de los fragmentos de clúster manualmente, podríamos pensar que estas secuencias pertenecen a un microorganismo mayoritario que se encuentra en un nivel inferior en lo que a abundancia se refiere.

2.1.3. | Análisis de otros elementos de interés encontrados en el metagenoma

Como se ha expuesto en los apartados anteriores, la secuencia obtenida del metagenoma de la esponja *P. littoralis* ha generado una gran cantidad de datos. Algunos de los elementos más llamativos que se han encontrado de forma paralela en el desarrollo de

este trabajo han sido también estudiados. En concreto, resultó de interés el análisis la secuencia mitocondrial de la esponja hospedadora, obtenida también del conjunto del metagenoma.

2.1.3.1. | Secuencia mitocondrial de la esponja *P. littoralis*

El DNA mitocondrial de la esponja *P. littoralis* fue ensamblado de forma manual partiendo del *contig* 8 y del conjunto de las lecturas. El DNA mitocondrial al completo se representó como una única molécula lineal de 21 719 bp. El genoma fue refinado de forma manual partiendo de la anotación automática generada por la herramienta web MITOS (Bernt *et al*., 2013) y la organización de los genes mitocondriales se recoge en la tabla 10 (Fig. 40). Todos ellos están codificados en la hebra pesada. Como se puede observar existen 2 genes solapados, que son *trnE* y *nad6* (ver tabla 10). Además de las secuencias génicas de las subunidades grande y pequeña del RNA ribosomal y los 25 RNAs de transferencia (tRNA), el genoma mitocondrial contiene 14 secuencias que codifican proteínas, incluyendo la secuencia del gen de la subunidad 9 de la ATPasa (*atp9*), que no está presente en la mayoría de las secuencias mitocondriales animales, pero si se ha observado en otras demosponjas (Lavrov *et al*., 2008). La composición en GC del total de la secuencia es del 31%. El conjunto de las secuencias intergénicas varían en longitud de 1-610 bp.

Gen	Posición de inicio	Posición de stop	Longitud	Codón de inicio	Codón de stop	Anticodón	Nucleótidos intergénicos
rrnL	36	2699	2664				35
trnY(gta)	2948	3018	71			GUA	248
trnM(cat)	3062	3133	72			CAU	43
cox2	3229	3960	732	ATG	TAA		95
trnK(ttt)	4057	4129	73			UUU	96
atp8	4131	4400	270	ATG	TAA		1
atp6	4460	5194	735	ATG	TAG		59
trnR(tct)	5386	5459	74			UCU	191
cox3	5516	6304	789	ATG	TAG		56
trnQ(ttg)	6367	6438	72			UUG	62
trnW(tca)	6482	6552	71				43
trnN(gtt)	6627	6697	71			GUU	74
trnL1(tag)	6728	6801	74			UAG	30
cob	6803	7957	1155	ATG	TAA		1
trnT(tgt)	8016	8089	74			UGU	58
atp9	8183	8419	237	ATG	TAA		93
trnS1(gct)	8524	8609	86			GCU	104

trnP(tgg)	8648	8720	73			UGG	38
nad4	8784	10235	1452	ATG	TAA		63
trnH(gtg)	10287	10359	73			GUG	51
trnE(ttc)	10423	10494	72			UUC	63
nad6	10492	11091	600	ATG	TAA		-3
nad3	11116	11472	357	ATG	TAA		24
trnR(tcg)	11598	11668	71			UCG	125
nad4L	11669	11968	300	ATG	TAG		0
cox1	12136	13698	1563	ATG	TAG		167
trnS2(tga)	13731	13814	84			UGA	32
trnD(gtc)	13857	13928	72			GUC	42
trnC(gca)	14539	14606	68			GCA	610
nad1	14698	15693	996	ATG	TAG		91
trnL2(taa)	15767	15850	84			UAA	73
trnI(gat)	15899	15971	73			GAU	48
trnM(cat)	15980	16050	71			CAU	8
nad2	16137	17549	1413	ATG	TAA		86
nad5	17646	19538	1893	ATG	TAG		96
trnA(tgc)	19612	19684	73			UGC	73
trnM(cat)	19776	19847	72			CAU	91
trnF(gaa)	19960	20032	73			GAA	112
rrnS	20033	21525	1493				0
trnG(tcc)	21526	21597	72			UCC	0
trnV(tac)	21647	21719	73			UAC	49

Tabla 10 | Organización del genoma mitocondrial de *P. littoralis*. En la relación de elementos del genoma se especifica para cada uno de ellos la posición de inicio y de STOP así como la longitud y los nucleótidos intergénicos para esa posición. En el caso de los genes de tRNA se indica la secuencia del anticodón y para los genes proteicos se indica el codón de inicio y de STOP.

La mayoría de los tRNA pueden ser plegados en la típica estructura secundaria en forma de trébol, sin embargo *trnS2*, *trnL2* and *trnS1* poseen bucles variables más extendidos. El conjunto total de la secuencia es similar al de *Geodia neptuni* (NC_006990) (Lavrov *et al*., 2005), pero difiere de este en los tRNA encontrados y en la posición del gen *trnY*. Además, existe una zona intergénica de 600 bp aproximadamente entre los genes *trnD* y *trnC*, pero esta región no parece contener ninguna secuencia palindrómica o repetida como sucede en otras mitocondrias. Estas secuencias ricas en AT no codificantes se han detectado en otros genomas mitocondriales de esponjas, y el hecho de presentar secuencias palindrómicas repetidas se ha asociado de forma especulativa a funciones de control (Erpenbeck *et al*., 2009).

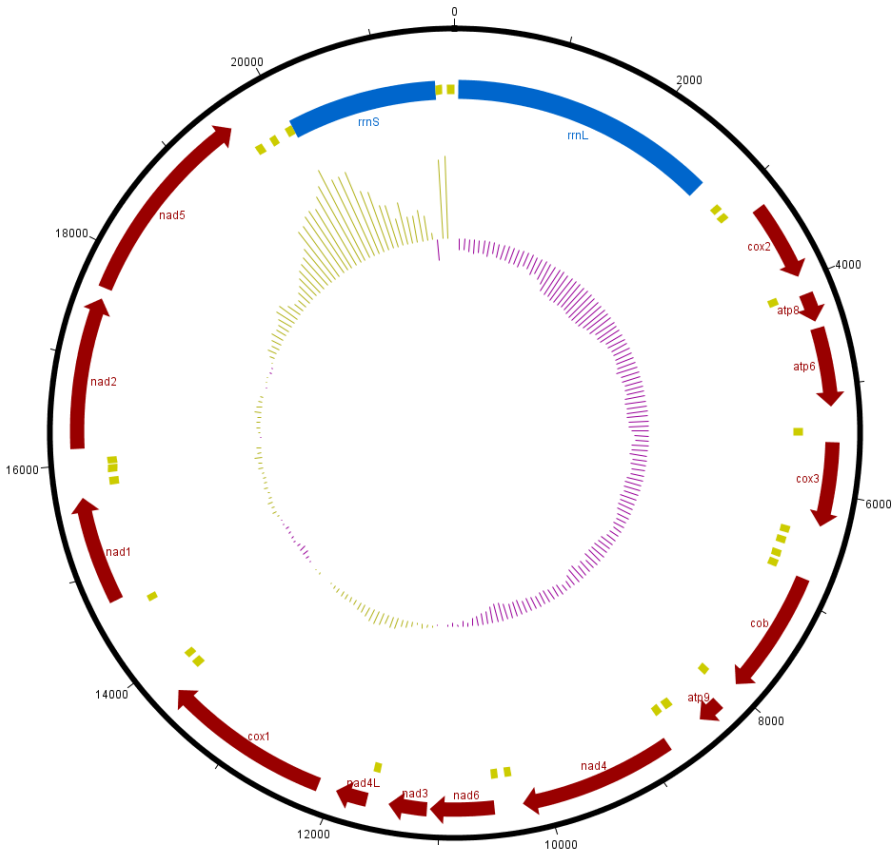

Figura 40 | Esquema de la secuencia genómica mitocondrial de *P. littoralis*. En rojo se representan los genes proteicos, en dorado los tRNAy en azul el RNA ribosomal. En el centro se aprecia la variación del contenido en GC para cada zona (en morado, menor de la media y en dorado mayor).

2.2. | Secuenciación y análisis del metagenoma de PMLT01 para la obtención secuencias de interés que codifican PKSs y NRPSs

2.2.1. | Análisis metagenómico del microbioma

La presencia de un compuesto con estructuras relacionadas con PM050489 (Martín *et al*., 2012, (Patente US 8324406 B2); Martín *et al*., 2013), que posee actividad antitumoral fue detectada por PharmaMar en la esponja lithistida de especie desconocida PMLT01. En concreto, la muestra con la que se realizó este trabajo fue recolectada en el océano Indo-Pacífico y se congeló inmediatamente a -20 °C. La identificación de las características estructurales del compuesto antitumoral mediante HPLC-MS hizo pensar que posiblemente esta molécula estaba siendo sintetizada por un clúster PKS/NRPS bacteriano. Al igual que con la esponja *P. littoralis*, se decidió analizar la fracción microbiana de PMLT01 utilizando herramientas metagenómicas para realizar la búsqueda de clústeres PKS/NRPS.

2.2.1.1. | Aislamiento de la fracción microbiana

Para tratar de extraer la información metagenómica sobre la presencia de clústeres PKS/NRPS en las bacterias asociadas a la esponja se realizó el aislamiento de la fracción microbiana (ver Materiales y Métodos). Debido a la morfología del tejido de esta esponja, en este caso se tuvo en cuenta el incluir partes iguales de cada una de las zonas que se consideraron diferentes entre sí. Estas zonas se identificaron como una zona interior, una zona exterior y la zona de la base. Se realizaron dos aislamientos distintos seguidos de sus sucesivas extracciones de DNA (ver Materiales y Métodos). Para la primera extracción se utilizaron 1,4 g de la esponja y se obtuvo un total de 15 µg de DNA, mientras que la segunda extracción se realizó a partir 4,6 g de esponja de los que se obtuvieron 29,7 µg de DNA. Ambas muestras de DNA de la fracción microbiana se mezclaron para obtener un total de 44,7 µg, con los cuales se pudo proceder a secuenciar con técnicas de secuenciación masiva.

2.2.1.2. | Secuenciación de la fracción microbiana

Tras la experiencia obtenida al secuenciar el metagenoma microbiano de *P. littoralis*, se decidió obtener una mayor cantidad de lecturas a la hora de secuenciar masivamente, dado que se quiso abordar la búsqueda con más profundidad para aumentar las posibilidades de encontrar clústeres PKS/NRPS. Para ello se decidió llevar a cabo una secuenciación mixta entre una aproximación mediante pirosecuenciación y otra utilizando *Ion Torrent*. Las estadísticas de ambos procesos de secuenciación masiva se pueden observar en la tabla 11, siendo el total de lecturas obtenidas de 4146779 con una media de tamaño de 282 bp (Fig. 41), lo cual abarca un total de 1,17 Gb. El contenido medio en GC del total del DNA metagenómico resultó ser del 55% (Fig. 41).

	Pirosecuenciación	**Ion Torrent**	**Total**
Lecturas	513665	3633114	4146779
Media de las lecturas (bp)	637,6	232,2	282,4
Nucleótidos totales (bp)	327526575	843489415	1171015990

Tabla 11 | Características de los procesos de secuenciación masiva del metagenoma de PMLT01

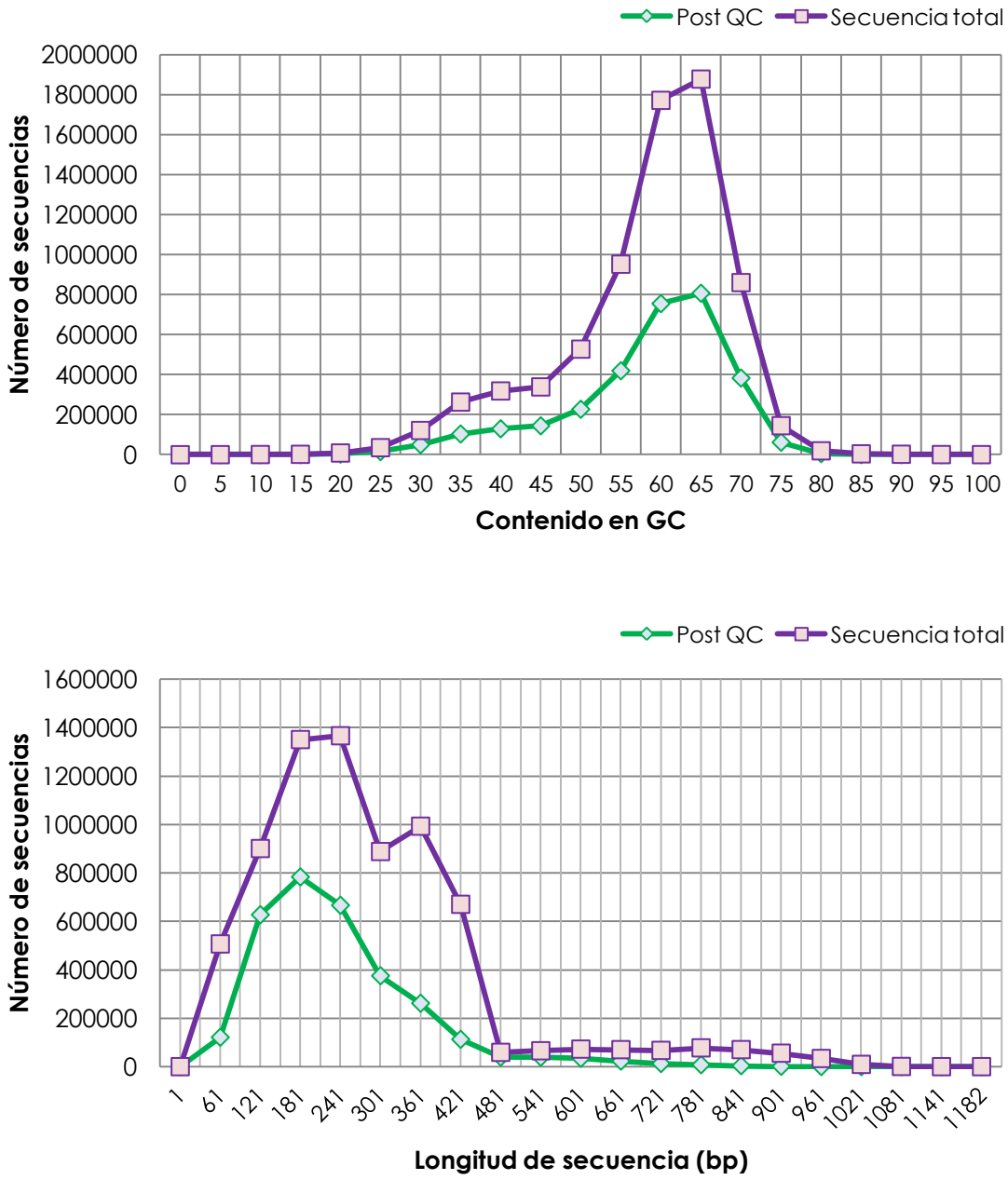

Figura 41 | Distribución de las secuencias del metagenoma de *P. littoralis* según su longitud y contenido en GC. En morado se muestra la secuencia obtenida directamente de los dos procesos de secuenciación masiva mientras que en verde las secuencias que han pasado el control de calidad (QC) del servidor MG-RAST. En el gráfico superior se observa la distribución del número de secuencias según su porcentaje del contenido en G+C mientras que en el inferior se puede ver la distribución del número de secuencias obtenidas según la longitud.

2.2.1.3. | Análisis de la secuencia del metagenoma microbiano en MG-RAST

Como se hizo en el caso de *P. littoralis*, el total de las lecturas se analizaron de forma preliminar utilizando el servidor especializado MG-RAST (Meyer *et al.*, 2008). De esta forma se obtuvieron las características generales del metagenoma microbiano de PMLT01.

2.2.1.3.1. | Resultados de la asignación funcional de l total de la muestra

Del total de las lecturas, únicamente un 75% pasó los controles de calidad de MG-RAST (3109965 secuencias). El 91% de estas secuencias filtradas dieron lugar a 1923198 regiones codificantes, de las cuales únicamente un 27,7% (531767 regiones) pudieron anotarse utilizando las bases de datos de proteínas del servidor (M5NR), mientras que el 72,3% restante no tenían similitud suficiente. El 76% de estas características anotadas pudieron finalmente asignarse a las categorías funcionales definidas por el servidor (Fig. 42).

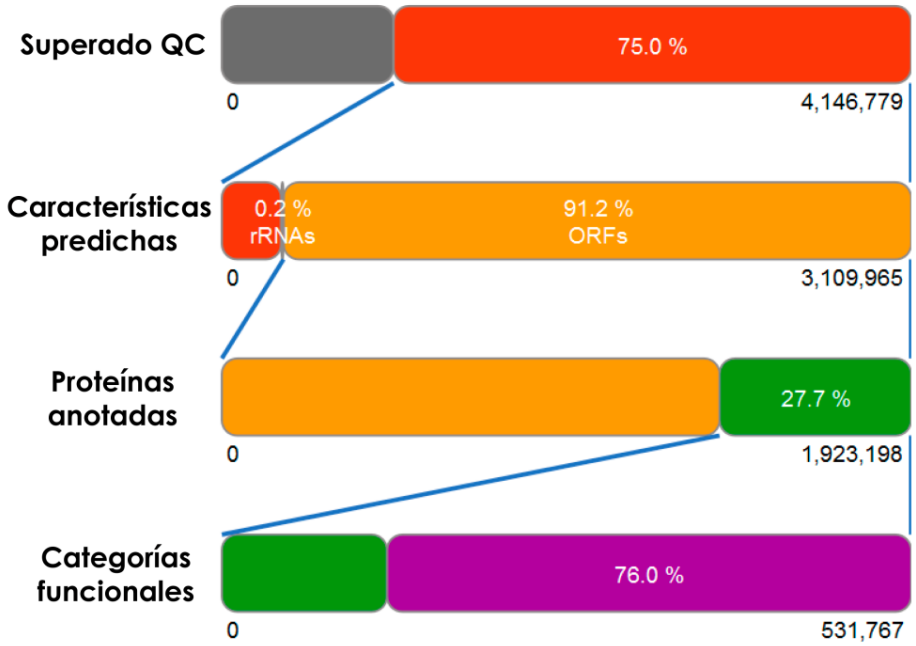

Figura 42 | Esquema del proceso de anotación en MG-RAST. En el esquema se puede observar la cantidad de secuencias procesadas en los pasos de control de calidad (QC), predicción de características, anotación de proteínas y asignación de categorías funcionales.

Como se puede observar en la distribución de funciones por subsistemas, los más representados, sin contar los grupos que recopilan categorías sin determinar, son aquellos relacionados con el metabolismo de los carbohidratos, los aminoácidos y el metabolismo de las proteínas (Fig. 43).

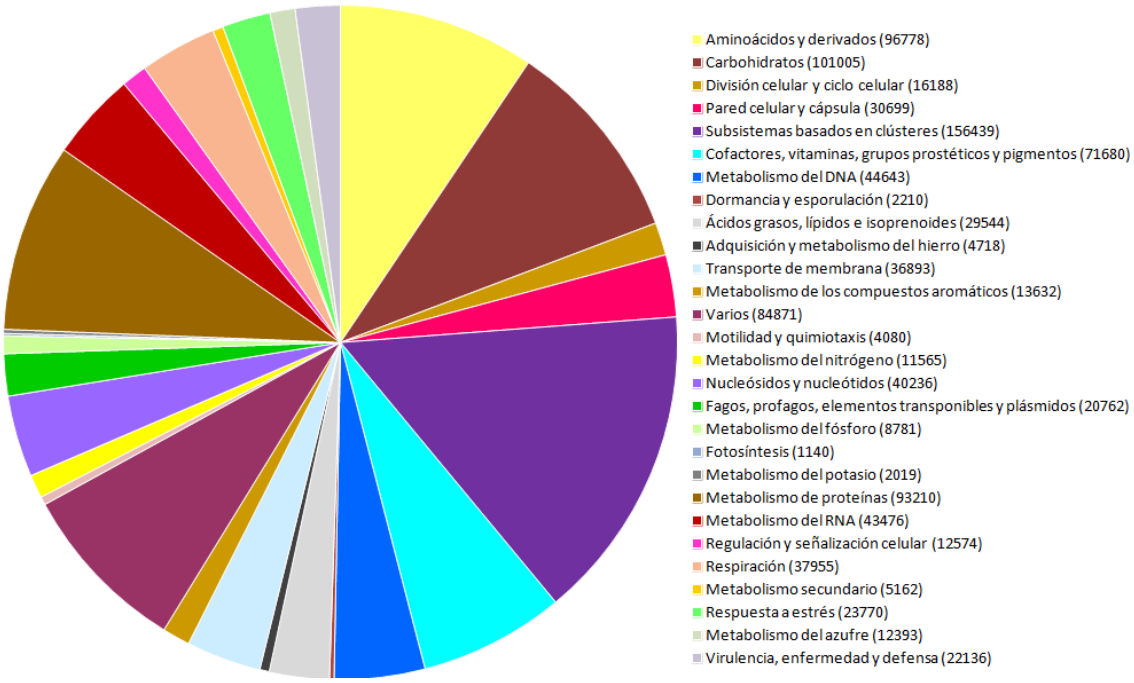

Figura 43 | Gráfica de la distribución por subsistemas de la anotación del metagenoma de PMLT01 En el gráfico proporcionado por la herramienta MG-RAST se representan el número de positivos para cada uno de los subsistemas agrupados en categorías generales tras el proceso de anotación automática.

2.2.1.3.2. | Resultados de la asignación taxonómica del metagenoma del total de la muestra

A continuación se analizaron los datos de asignación taxonómica de la herramienta MG-RAST. De este modo se obtuvo la distribución de las secuencias por dominios. En concreto, se puede observar como la mayoría de la secuencia se clasificó como de origen bacteriano (Fig. 44) seguido por una proporción del 3,8% de DNA de arqueas. Al profundizar más en el siguiente nivel de clasificación taxonómica se obtuvo la distribución de las secuencias según su pertenencia a los distintos phyla (Fig. 45), siendo el phylum Proteobacteria el más abundante, seguido de Actinobacteria, Planctomycetes, Firmicutes, Chloroflexi y Bacteriodetes.

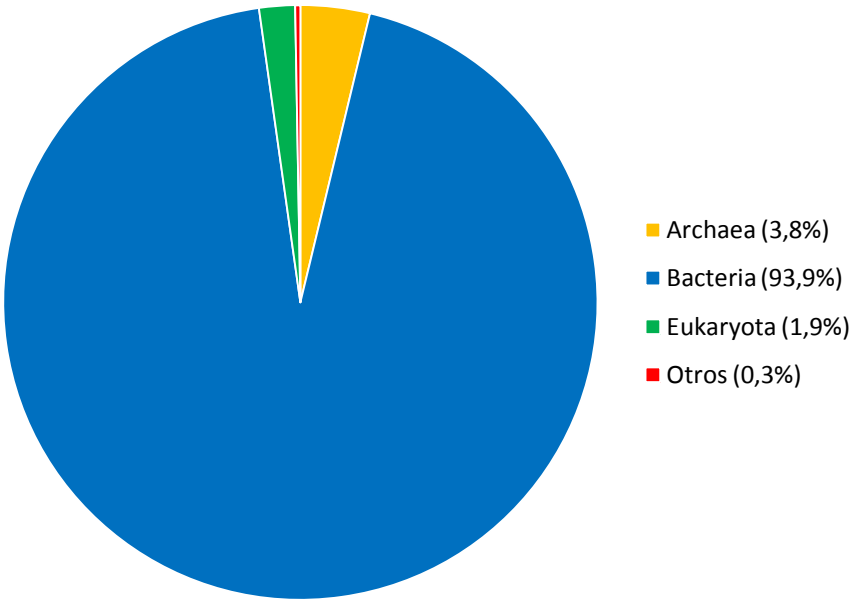

Figura 44 | Distribución por dominios de las secuencias del metagenoma de PMLT01

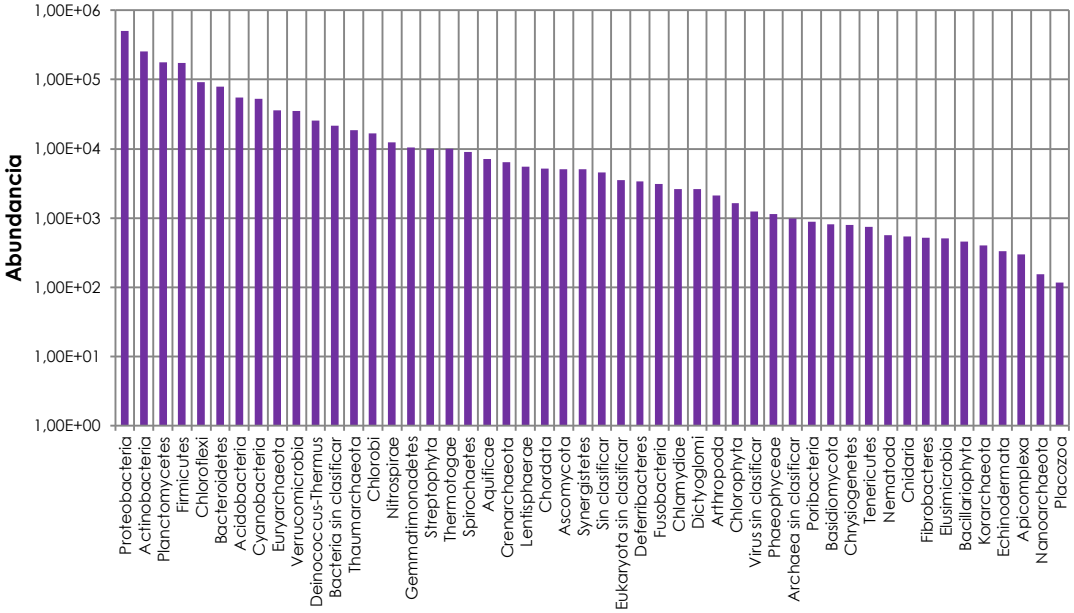

Figura 45| Abundancia de cada phylum en las secuencias clasificadas de PMLT01

La curva de rarefacción generada de forma automática por MG-RAST (Fig. 46) indica el nivel de saturación de la secuencia del metagenoma. Como se muestra en la figura 46 para el conjunto de las secuencias utilizado el número de especies distintas sería aproximadamente de 4700.

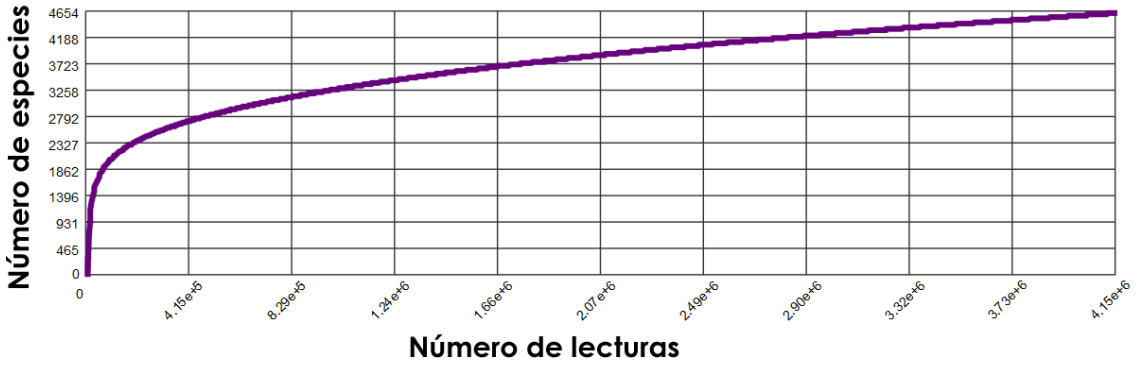

Figura 46 | Curva de rarefacción de las secuencias de PMLT01 En la gráfica obtenida de MG-RAST se muestra el grado de saturación en especies de la muestra identificando el número de especies distintas conforme se van analizando las lecturas.

2.2.1.4. | Ensamblaje de la secuencia metagenómica de la fracción microbiana de la esponja PMLT01

A continuación se procedió a realizar un ensamblaje *de novo* de todas las secuencias para así obtener fragmentos con más información de aquellas secuencias presumiblemente más representadas en el metagenoma. Tal y como se hizo en el proceso de ensamblaje de *P. littoralis* y tras realizar pruebas preliminares de ensamblaje, se decidió utilizar el software Newbler (ver Materiales y Métodos). Sin embargo al disponerse en este caso de mucha más cantidad de secuencia que en el caso del metagenoma de *Polymastia*, no se poseía la capacidad de computación necesaria para realizar un ensamblaje con los parámetros por defecto del ensamblador. Para evitar este problema se realizaron ensayos de *clusterización* de las lecturas utilizando el software CD-HIT con distintas restricciones de identidad (de 0,98 a 1) y se propuso subdividir las secuencias y hacer ensamblajes por separado. Finalmente el ensamblaje que proporcionó el mejor resultado atendiendo a la cantidad de secuencia ensamblada y a la longitud final de los *contigs* resultó ser el procesamiento de la secuencia obtenida con Ion Torrent, sin *clusterizar* y utilizando Newbler en modo optimizado para grandes cantidades de secuencia (opción *large*). El resto de la secuencia obtenida por pirosecuenciación se utilizó en el desarrollo de experimentos posteriores para completar manualmente la secuencia ensamblada.

Finalmente el ensamblaje obtenido que mejor resultó poseía un total de 96558 *contigs* que abarcan 74073039 bp teniendo el mayor *contig* 423453 bp de longitud. A continuación se realizó un mapeo de las lecturas totales frente al ensamblaje realizado utilizando bowtie2 y se observó que un 52,11% de las lecturas estaban representadas en la secuencia ensamblada. Esta fracción estaría formada por un número relativamente pequeño de individuos probablemente más abundantes.

2.2.1.5. | Asignación taxonómica de la secuencia metagenómica ensamblada

Una vez obtenidos fragmentos de secuencia más largos fue posible realizar una asignación taxonómica más precisa. Como se hizo en el caso de *Polymastia* se realizaron variaciones sobre el protocolo *in silico* de Albertsen *et al*. (2013) (ver Materiales y Métodos). De este modo se pudieron separar y agrupar las secuencias según sus características estructurales previamente extraídas.

Una vez extraídos los datos del mismo modo que en el caso del metagenoma de *P. littoralis* (ver Resultados anteriores) en primer lugar se realizó una representación de los *contigs* según sus valores de abundancia y contenido en GC (Fig. 47) utilizando únicamente los *contigs* muy representados. En la figura se pueden observar de forma clara 5 acumulaciones que se nombraron con letras de A-E. Además se realizó un análisis taxonómico con herramientas de comparación (HMMER y Blast) de todos aquellos *contigs* mayores de 8000 bp que tenían valores de abundancia por encima de 10 y no se encontraban en ninguno de los clúster mencionados. También se analizó del mismo modo todas aquellas secuencias que aun siendo más pequeñas se encontraban muy representadas y tenían un valor de abundancia de 25 o superior.

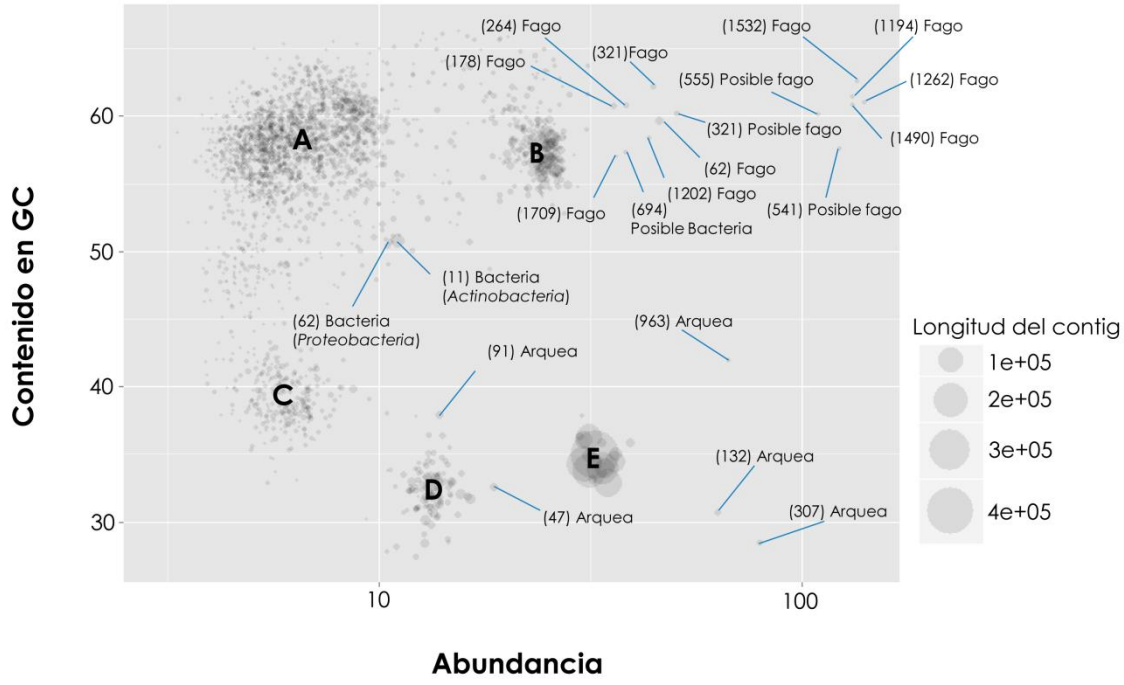

Figura 47 | Distribución de la secuencia ensamblada de PMLT01 y análisis de los *contigs* individuales y representativos. En la distribución se representan los *contigs* mayores de 2500 bp según su contenido en GC y abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). Se muestran los resultados de asignaciones taxonómicas manuales con HMMER, y BLAST para aquellas secuencias no agrupadas en cúmulos principales, representadas más de un 10x y mayores de 8000 bp. Los cúmulos de secuencias más significativos se señalan con letras de la A-E.

De los agrupamientos señalados, únicamente el B, el D y el E parecen tener una abundancia suficiente como para poder sugerir que se pueda tratar de genomas individuales. Los cúmulos A y C sin embargo pueden representar *contigs* de varios genomas distintos, debido a que se encuentran en una zona de poca abundancia y parecen estar más dispersos.

Al analizar la figura 7 se puede observar la presencia de fragmentos muy representados de procedencia fágica en la esquina superior derecha del diagrama, que probablemente podrían pertenecer al mismo fago. Sin embargo los fragmentos más

representados que tienen un contenido en GC más bajo se asocian a arqueas y pueden ser fragmentos pertenecientes a los genomas de los agrupamientos más cercanos ya sea el D o el E.

A continuación se realizó la clasificación taxonómica mediante comparación de los *contigs* en los que se detectaron genes esenciales y los datos obtenidos se incluyeron en la representación (Fig. 48). En concreto el cúmulo A, parece tener una zona con más abundancia que agrupa secuencias de Actinobacteria y otra zona en la que predominan las del phylum Plantomyces. El grupo B, generalmente posee genes esenciales clasificados como pertenecientes a Actinobacteria, mientras que el grupo C no posee una clasificación homogénea, ya que existen secuencias clasificadas pertenecientes a arqueas y a Deltaproteobacterias. Sorprendentemente no aparecen muchos *contigs* que contengan genes esenciales en el cúmulo D, sin embargo, se debe tener en cuenta que en estos diagramas solo se representan aquellos grupos de genes esenciales que poseen más de 5 secuencias clasificadas en el mismo phylum (ver Materiales y Métodos), por lo que volviendo a analizar el contenido en genes esenciales sin esta restricción para el grupo D, se observa la presencia de algunas secuencias más que contienen genes esenciales (datos no mostrados). La clasificación de estas secuencias resultó ser dispar aunque siempre indicando phyla de arqueas, por lo que probablemente la secuencia de este grupo pertenezca a una arquea que no tenga análogos similares en las bases de datos. Aun así, el bajo número de secuencias con genes esenciales parece indicar que el genoma no está completo. Por último, el agrupamiento E contiene varias secuencias muy largas que parecen estar clasificadas como pertenecientes al phylum Thaumarchaeota, por lo que probablemente se trate de un genoma completo muy abundante en el metagenoma.

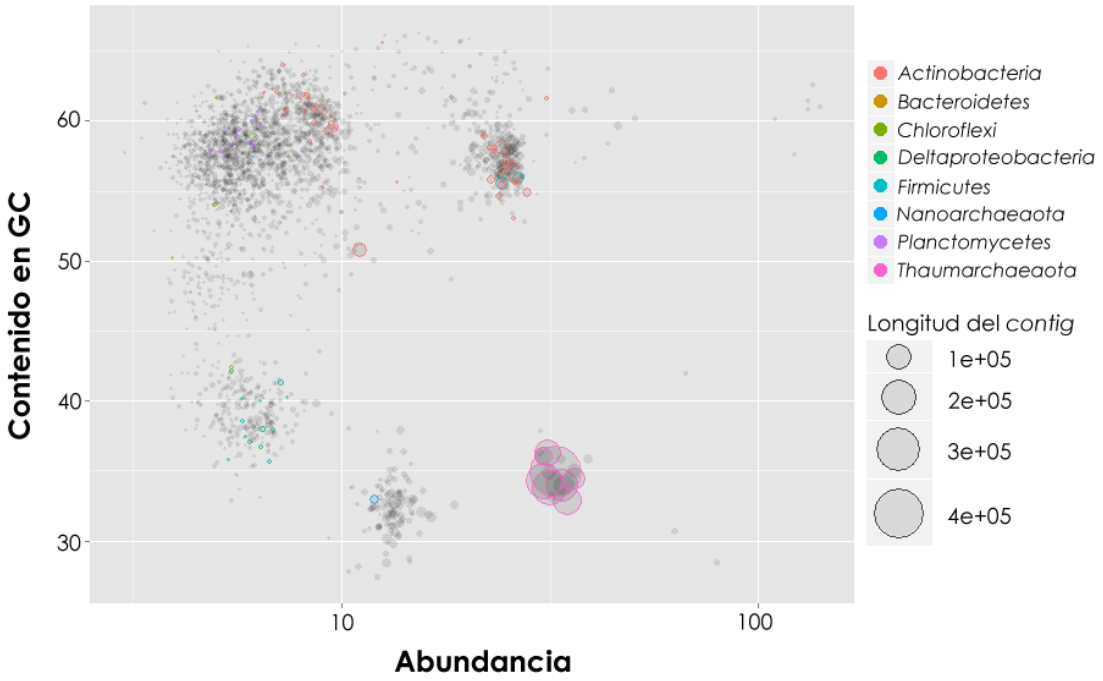

Figura 48 | Distribución de la secuencia ensamblada de PMLT01 y análisis taxonómico de genes esenciales. En la distribución se representan los *contigs* mayores de 2500 bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). Dependiendo de la clasificación taxonómica obtenida se señalan en distintos colores aquellos *contigs* que contienen genes esenciales.

A continuación se procedió a realizar el análisis de asignación taxonómica mediante el algoritmo de clasificación PhyloPythiaS+ y se representó esta clasificación sobre el diagrama antes mostrado (Fig. 49).

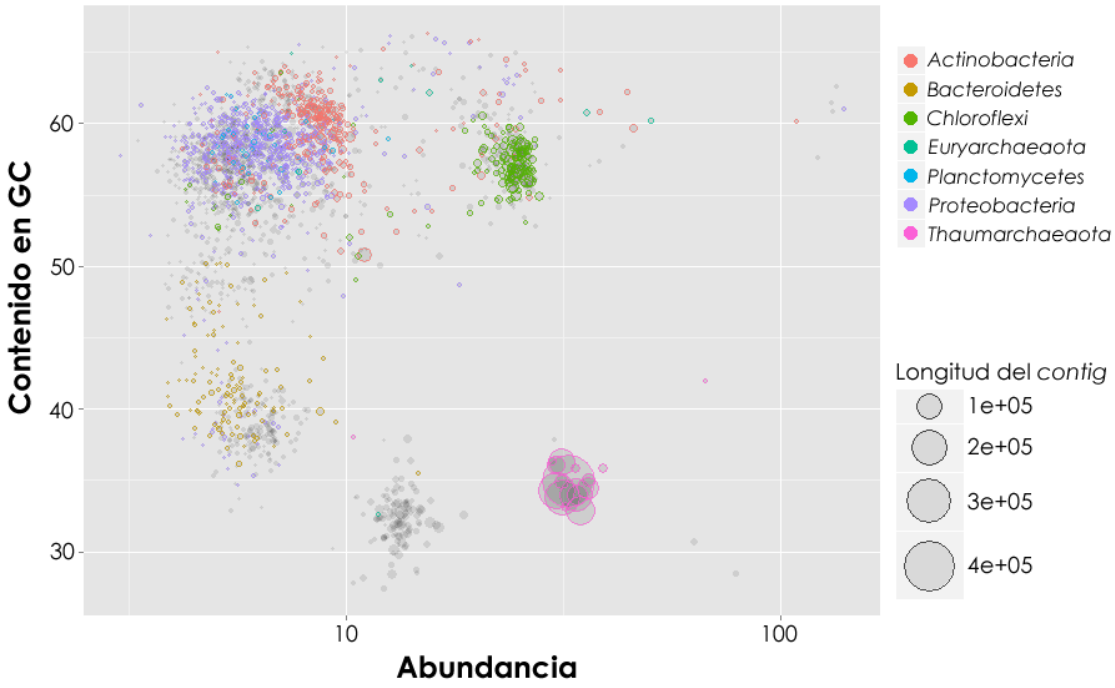

Figura 49 | Distribución de la secuencia ensamblada de PMLT01 y análisis taxonómico mediante PhyloPythiaS+. En la distribución se representan los *contigs* mayores de 2500 bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). La clasificación taxonómica obtenida para cada *contig* se señala en distintos colores. Aquellos *contigs* que aparecen sin asignación taxonómica no pudieron ser clasificados por la herramienta PhyloPythiaS+.

En la figura se puede observar como aquellos grupos que están más representados tienden a agruparse bajo la misma clasificación, lo que sugiere que pueden contener genomas individuales. En concreto los grupos B y E aparecen clasificados como Chloroflexi y Thaumarchaeota respectivamente. El grupo B difiere en su clasificación con la de sus genes esenciales, sin embargo, el grupo E coincide con la clasificación de las secuencias con genes esenciales. El grupo A tiene una zona más representada con presencia de Actinobacterias y otra zona compuesta por secuencias de procedencia variada. Algo parecido ocurre con el grupo C, el cual tiene una zona con presencia mayoritaria de lo que PhyloPythiaS+ clasifica como Bacteroidetes y otra zona sin clasificar. Llama la atención el caso del cúmulo D, cuyas secuencias apenas han sido clasificadas por el algoritmo, algo que concuerda con la poca cobertura de la clasificación de genes esenciales.

Tras analizar las clasificaciones taxonómicas todo parecía indicar que las secuencias agrupadas en el cúmulo E podían pertenecer a un mismo individuo, que correspondería a una arquea. Además, si se observa la longitud de los *contigs* de este agrupamiento, se puede apreciar que estos son de gran longitud, lo cual es un buen indicativo de que la secuencia de este genoma pudiera estar completa.

2.2.2. | Búsqueda de secuencias de PKSs y NRPSs en la secuencia metagenómica

Como se hizo con el metagenoma de *P. littoralis*, se realizó un análisis de la presencia de genes que codifican PKSs y NRPSs para comprobar la eficacia de las herramientas metagenómicas a la hora de encontrar clústeres responsables de la síntesis de este tipo de metabolitos secundarios.

2.2.2.1. | Búsqueda de secuencias de PKSs y NRPSs en el total de la secuencia metagenómica

Para realizar un recuento total de las funciones relacionadas con clústeres PKS/NRPS, se realizaron búsquedas con HMMER para los dominios acil transferasa (AT), de unión a AMP (incluido el específico del C-terminal), condensación (C), ceto-reductasa (KR), cétido sintasa (KS) (N-terminal y C-terminal), de sitio de unión a fosfopanteteina (PP), dehidratasa (DH) y tioesterasa (TE) tomados de la base de datos Pfam. Se realizaron estas búsquedas sobre el total de las lecturas obtenidas y sobre los *contigs* (ver tabla 12). A continuación, para eliminar parcialmente el efecto de la redundancia se realizó un proceso de *clusterización* de las lecturas (con un 98% de identidad) que habían dado positivo con HMMER y se realizó la búsqueda de nuevo.

Dominio	Nombre Pfam	*Contigs*	Lecturas	Clustering
Acil-Transferasa (PF00698.16)	Acyl_transf_1	20	110	102
Unión a AMP (PF00501.23)	AMP-binding	301	3441	2560
Unión a AMP C-terminal (PF13193.1)	AMP-binding_C	60	411	365
Condensación (PF00668.15)	Condensation	2	35	29
KS N-terminal (PF00109.21)	ketoacyl-synt	49	381	301
KS C-terminal (PF02801.17)	Ketoacyl-synt_C	37	384	269
KR (PF08659.5)	KR	74	267	393
Sitio de unión de Fosfopanteteina (PP) (PF00550.20)	PP-binding	16	67	107
Dehidratasa (PF14765.1)	PS-DH	2	21	19
Tioesterasa (PF00975.15)	TE	2	12	6

Tabla 12 | Recuento de la representación de dominios relacionados con clústeres PKS/NRPS en la secuencia de PMLT01 Para cada dominio se señala entre paréntesis el número de acceso correspondiente en la base de datos Pfam. Se representan los recuentos para la secuencia ensamblada (*contigs*) para las lecturas y para las lecturas *clusterizadas* con un 98% de identidad.

Como resulta lógico, el número de positivos en el conjunto de las lecturas así como en las secuencias *clusterizadas* fue mayor que en los *contigs*. Se observó que el proceso de *clusterización* reduce el número de positivos en las lecturas en la mayoría de los casos. Llama la atención el caso de los dominios KR dado que aparecen más dominios al *clusterizar*. Esto se debe a que al reducir el tamaño de la base de datos tras la *clusterización* la gran cantidad de positivos que se encontraban por debajo, pero muy cerca del valor límite del *e-value,* pasaron los criterios de restricción tras este procesamiento. El dominio más abundante en todos los casos es el de unión a AMP, el cual está mucho más representado probablemente debido a que también está asociado a otras funciones no relacionadas con la producción de metabolitos secundarios. Aunque el resultado indica que existen más positivos en la secuencia total que pueden no haber sido

ensamblados, debido a su longitud, estas secuencias no aportan la información suficiente para discernir si se trata de un dominio incluido en un clúster de interés PKS/NRPS.

2.2.2.2. | Búsqueda de secuencias de PKSs y NRPSs en la secuencia metagenómica ensamblada

El trabajo con secuencias ensambladas hace posible contar con la suficiente información para poder clasificar la pertenencia de estos fragmentos a clústeres PKS/NRPS. De este modo y partiendo de nuevo de la hipótesis de que un compuesto que resulta detectable en una muestra marina es generado con mayor probabilidad por un microorganismo que se encuentra de forma más abundante en el microbioma de la esponja, se utilizaron herramientas *in silico* para detectar estas posibles secuencias de interés.

Del conjunto de los positivos detectados en las secuencias ensambladas se seleccionaron todos aquellos *contigs* de más de 1000 bp de longitud. A continuación con estas secuencias se utilizaron herramientas como BLAST y búsquedas en la base de datos Pfam (ver Materiales y Métodos) para identificar los componentes del contexto génico de cada uno de los fragmentos. Para aquellos dominios que resultan muy informativos a la hora de identificar la presencia de un clúster de interés PKS/NRPS (como por ejemplo el caso de los dominios C o DH), se analizaron también *contigs* con un tamaño menor al establecido previamente. A continuación se desecharon manualmente todos aquellos fragmentos en los que el resultado de la búsqueda no dio indicios de pertenencia a un clúster de interés. Uno de los criterios principales en el cribado manual fue que las secuencias pudieran formar parte de ORFs multifuncionales que codifican PKSs y/o NRPSs, algo que no ocurre la mayoría de las veces con la gran cantidad de positivos para los dominios de unión a AMP, los cuales suelen aparecer de forma individual confinados en una única ORF.

Tras utilizar estas herramientas y para extender la longitud de las secuencias extraídas se realizó un ensamblaje manual realizando alineamientos con BLAST de los extremos 5' y 3' correspondientes frente al conjunto total de las lecturas. Es importante señalar que la base de datos de lecturas que se utilizó para realizar estos ensamblajes, contenía las lecturas procedentes de ambos procesos de secuenciación, tanto de la aproximación con pirosecuenciación como de la de Ion Torrent. Como resultado de este proceso se obtuvieron un total de 5 secuencias (ver tabla 13).

Secuencia	Contigs	Dominios de Pfam	Longitud (bp)	Phylum
1	14633	ADH_N/ADH_zinc_N/ PS_DH/Acyl_transf_1	3115	*Planctomycetes*
2	2606/31633	PP/AMP/KS/KS	4934	*Actinobacteria*
3	35957	KR/Condensation	2633	*Proteobacteria*
4	8613	KR/PP/TE/Epimerase	4120	*Planctomycetes*
5	45746	AMP/PP/KS/AT/DH	7726	*Actinobacteria*

Tabla 13| Secuencias seleccionadas con dominios PKS/NRPS. Para cada secuencia se indica los *contigs* que la contienen, la sucesión de dominios encontrada (con la notación correspondiente a la de la base de datos Pfam) y su longitud.

En esta ocasión se encontraron menos secuencias relacionadas con clústeres PKS/NRPS que en el caso de *Polymastia*. El tamaño de las secuencias ensambladas hace pensar que es probable que ninguno de estos fragmentos pueda pertenecer a un clúster de alguno de los microorganismos mayoritarios. En algunos casos como en la secuencia 2, no se detectan dominios esperables en una NRPS convencional o en una PKS de tipo I, como por ejemplo los dominios de condensación o PP, por lo que puede que este fragmento se pueda clasificar como otro tipo de clúster.

Para comprobar si varias de estas secuencias podrían pertenecer al mismo microorganismo o no, se realizó un análisis similar al de presencia de genes esenciales del apartado de asignación taxonómica. De este modo se seleccionaron los *contigs* presentes en la tabla 13 y se clasificaron taxonómicamente por comparación (BLAST). A continuación se representaron los *contigs* con la excepción de los fragmentos menores de 1000 bp, cuyo tamaño impide la correcta visualización de los datos debido al ruido generado por los más pequeños (Fig. 50).

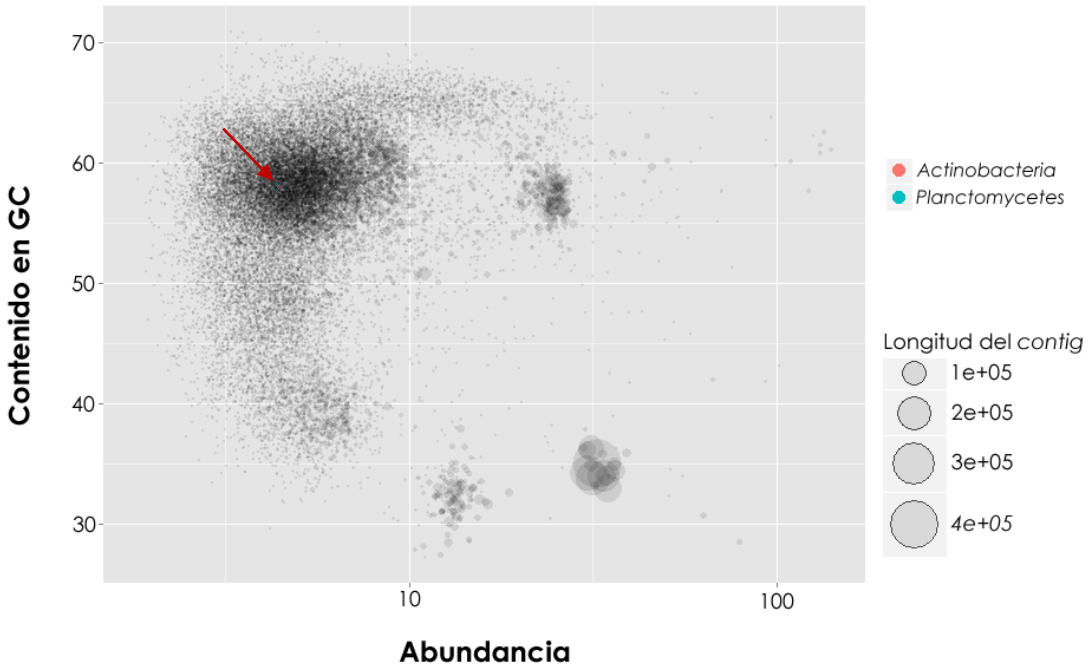

Figura 50 | Distribución de los *contigs* seleccionados con fragmentos con dominios PKS/NRPS en el conjunto de la secuencia ensamblada de PMLT01 En la distribución se representan los *contigs* mayores de 1000 bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). La clasificación taxonómica obtenida para los *contigs* que contienen dominios PKS/NRPS se señala con colores. La flecha de color rojo indica la posición de los *contigs* seleccionados.

Como se puede apreciar en la figura 10 los *contigs* representados se identifican como pertenecientes a los phylum Actinobacteria y Planctomycetes. La zona en la que aparecen es lo suficientemente cercana como para que perteneciesen al mismo microorganismo, sin embargo existe una alta densidad de *contigs* que corresponden a secuencias poco representadas, por lo cual es posible que se trate de una coincidencia, ya

que además están clasificadas de forma distinta. En este caso, los parámetros obtenidos de las secuencias de interés no permiten realizar una asignación fiable de varias de las secuencias a un mismo clúster biosintético. Todos estos datos sugieren que el microorganismo productor de la molécula de interés no se encuentra entre las secuencias mayoritarias del metagenoma obtenido.

2.2.3. | Análisis de otros elementos de interés encontrados en el metagenoma

La gran cantidad de datos generada de la secuenciación del metagenoma de PMLT01 ha proporcionado suficiente información para encontrar algunos elementos llamativos de forma paralela al desarrollo de este trabajo. En concreto en este metagenoma se ha podido analizar la secuencia completa de una arquea simbionte que aparece muy representada en el metagenoma.

2.2.3.1. | Identificación del genoma de la arquea simbionte

Para extraer el genoma completo de la arquea simbionte se aplicó el protocolo desarrollado por Albertsen et al. (Albertsen *et al.*, 2013). En primer lugar se localizó el cúmulo donde se encontraba la principal representación de secuencias pertenecientes a esta arquea y se seleccionaron de forma manual aquellos *contigs* mayores de 1000 bp que se encontraban en un entorno aproximado en la figura 51.

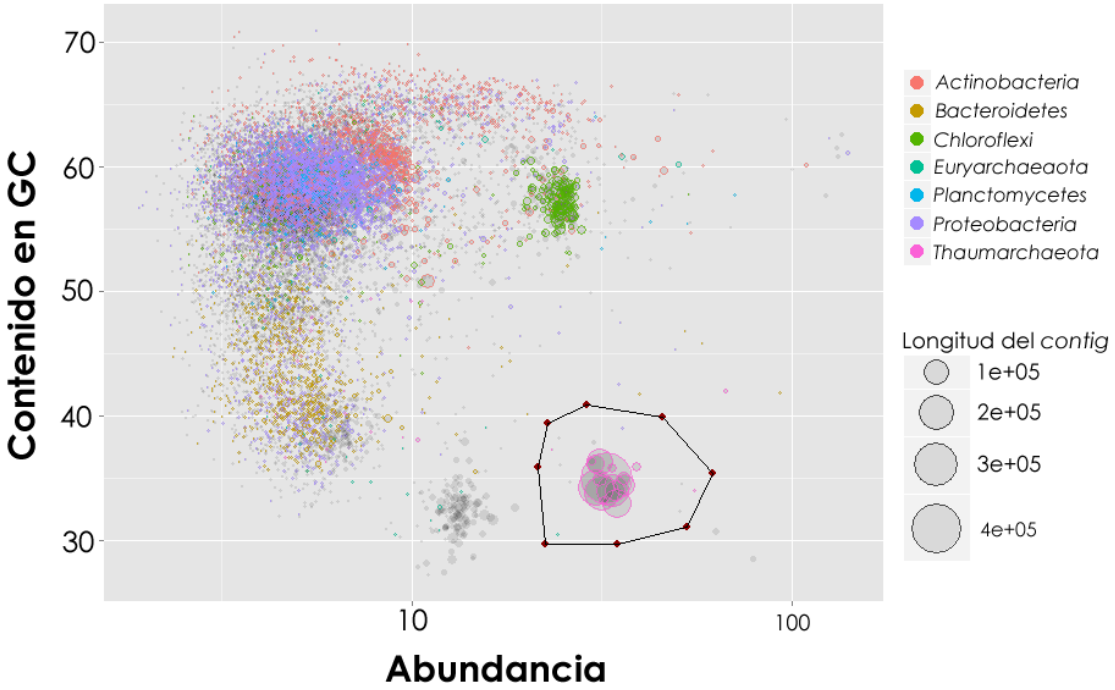

Figura 51 | Distribución de los *contigs* clasificados taxonómicamente para el aislamiento de la arquea simbionte seleccionada. En la distribución se representan los *contigs* mayores de 1000 bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). La clasificación taxonómica fue realizada con la herramienta PhyloPythiaS+ y se señala en distintos colores. La zona seleccionada contiene las secuencias extraídas correspondientes al cúmulo de la arquea simbionte y sus secuencias adyacentes.

Para eliminar posibles secuencias contaminantes se representaron los *contigs* seleccionados en función de distintas características (ver Materiales y Métodos) (Fig. 52) para así volver a agrupar las secuencias teniendo en cuenta otros parámetros distintos a la abundancia y el contenido en GC. De este modo, al representar utilizando los nuevos parámetros se consigue la agrupación de aquellos fragmentos pertenecientes al genoma individual que por alguna razón poseían un valor excepcional de abundancia o contenido en GC y en principio no aparecen en el cúmulo de interés. Como se puede observar en la figura 34 son varias las representaciones que consiguen agrupar de forma más efectiva la secuencia perteneciente al microorganismo de interés separándola de posibles contaminantes. Por ejemplo, en la representación del contenido en GC y PC2 (componente principal del análisis de k-meros), la secuencia se agrupa coincidiendo con la asignación taxonómica total previamente realizada. Consiguiendo evitar la presencia de contaminantes, se seleccionó y se extrajo la secuencia del cúmulo formado.

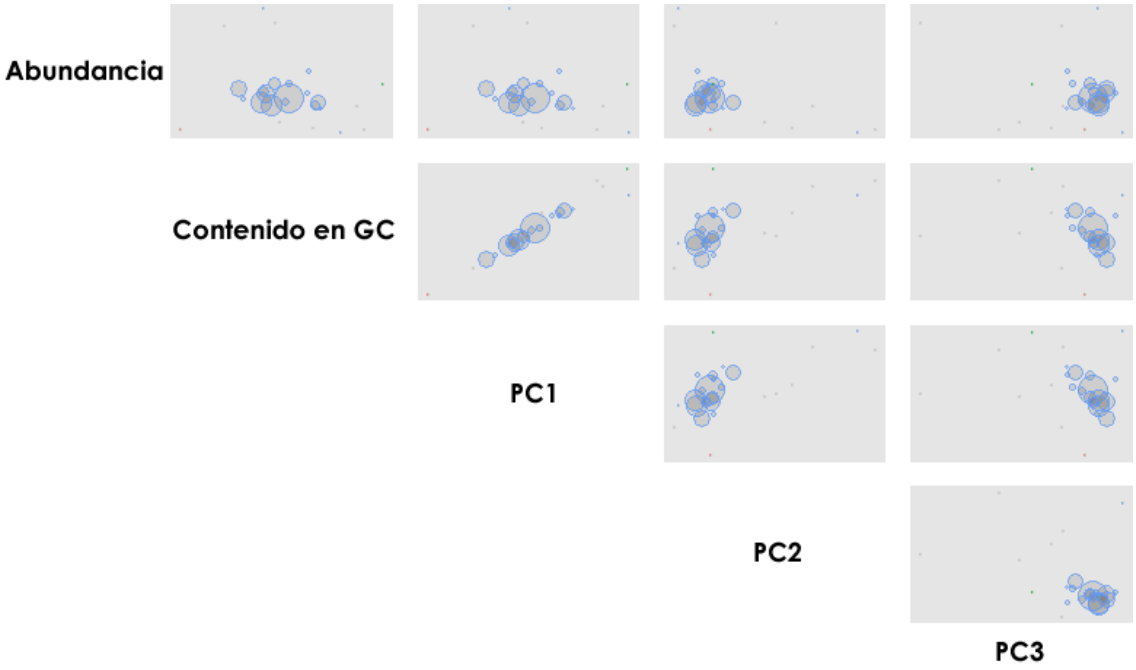

Figura 52 | Grafico de comparación para el subespacio de secuencia seleccionado correspondiente al genoma de la arquea simbionte. Los *contigs* pertenecientes a la arquea simbionte, seleccionados en el paso previo se representan en diversos gráficos comparando de dos en dos las variables de abundancia, contenido en GC y el resultado del análisis de componentes principales (PC) 1, 2, y 3. En azul se señala la secuencia clasificada como *Thaumarchaeota* mediante PhyloPythiaS+.

2.2.3.2. | Curación manual de la secuencia genómica de la arquea identificada

Una vez se obtuvo la secuencia se realizaron búsquedas con secuencias aleatorias de cada uno de los 16 *contigs* extraídos utilizando BLAST. Los resultaros indicaron que todos los fragmentos pertenecían a secuencias de arqueas, y se observó reiteración en la coincidencia de esta secuencia con el genoma de la arquea *Nitrosopumilus maritimus* (Walker *et al*., 2010). Para completar el borrador del genoma se realizó un análisis manual mediante búsquedas con BLAST de los *contigs* circundantes al cúmulo de la arquea (Fig. 51), para comprobar su pertenencia o no a la secuencia de su genoma. Para ello se tuvo en cuenta el porcentaje de identidad de cada uno de los *contigs* con arqueas del phylum Thaumarchaeota y en especial con *N. maritimus*. De este modo se añadieron 4 contigs más al borrador del genoma.

A continuación se extrajo una secuencia del gen del 16S rRNA de 1453 bp y se analizó utilizando la herramienta RDP Classifier (Wang *et al*., 2007). El resultado asignó taxonómicamente la secuencia al phylum *Archaea* y más concretamente al género *Nitrosopumilus*, que se corresponde con los resultados obtenidos en las búsquedas parciales de la secuencia mediante BLAST.

2.2.3.3. | Anotación del borrador final de la secuencia del genoma de la arquea simbionte

Una vez refinada manualmente la secuencia de la arquea se obtuvieron 20 *contigs* que abarcan 1,6 Mb de secuencia (ver tabla 14). El *contig* de mayor tamaño que se consiguió ensamblar tiene un total de 423453 bp y además, existe una gran proporción de la secuencia en los 6 fragmentos de más longitud, representando estos casi un 80% del total del genoma obtenido.

Tamaño del genoma (bp)	1590958
Contigs	20
Media de *contigs (bp)*	79548
Contenido en GC (%)	34,5
Tamaño del mayor contig (bp)	423453

Tabla 14 | Características de la secuencia extraída de la arquea simbionte de PMLT01

El conjunto de 20 *contigs* resultante del proceso de ensamblaje se introdujo en el servicio de anotación automática RAST (Aziz *et al*., 2008). El resultado se visualizó con SEED (*theseed.org*) (Overbeek *et al*., 2014) obteniéndose un total de 49 RNAs y 2079 secuencias codificantes en las cuales se representaban 147 de los subsistemas definidos por SEED.

Al analizar la anotación y realizar posteriores comparaciones manuales se observó que la arquea simbonte muestra características similares a otras arqueas ya secuenciadas, entre las que se encuentran *N. maritimus* (1,65 Mb) (Walker *et al*., 2010), *Candidatus Nitrosoarchaeum limnia* BG20 (1,77 Mb) (Mosier *et al*., 2012) o *Candidatus Nitrosoarchaeum koreensis* MY1 (1,61 Mb) (Kim *et al*., 2011). Un dato interesante es que el genoma incluye secuencias codificantes de una posible ruta de oxidación de amonio exclusiva de arqueas, similar a la propuesta para *N. maritimus* así como genes homólogos a los implicados en la ruta del 3-hidroxiproprionato/4-hidroxibutirato, para la asimilación de carbono, y elementos que pertenecen los sistemas de replicación FtsZ y CvdABC (Walker *et al*., 2010).

2.3. | Secuenciación y análisis del metagenoma de *Lithoplocamia lithistoides*

2.3.1. | Análisis metagenómico del microbioma

El género de esponjas marinas *Lithoplocamia* fue descrito por primera vez por Arthur Dendy en 1922 (Dendy, 1922) y cuenta actualmente con 6 especies según la base de datos de especies marinas WoRMS (http://www.marinespecies.org/).

Este trabajo se ha llevado a cabo a partir de una muestra de la esponja *L. lithistoides* recolectada por PharmaMar en la que se detectó la presencia de un compuesto

con actividad antitumoral. Esta muestra se recogió en Madagascar y fue conservada inmediatamente a -20°C.

En esta muestra PharmaMar detectó e identificó el compuesto antitumoral PM050489 (Martín *et al*., 2012, (Patente US 8324406 B2); Martín *et al*., 2013) y el análisis de las características permitió hipotetizar que su síntesis debería ser mediada por un clúster PKS/NRPS con aproximadamente 10 módulos de síntesis.

Como en los casos de las esponjas anteriores, se decidió analizar la fracción microbiana de *L. lithistoides* haciendo uso de herramientas metagenómicas, para de este modo, realizar la búsqueda de clústeres PKS/NRPS sobre un tercer ejemplo distinto a los estudios previamente descritos.

2.3.1.1. | Aislamiento de la fracción microbiana

En este caso concreto la *L. lithistoides* no presentaba una clara diferenciación de tejidos evidente como sucedía en la esponja PMLT01, por lo que no se separaron fracciones para realizar el proceso de homogenización. Debido a las características de la muestra que poseía una elevada viscosidad, no se realizó un fraccionamiento por filtración como en los casos de *Polymastia* y PMLT01 y se abordó una extracción de DNA total (ver Materiales y Métodos). La cantidad de DNA obtenido fue de 644 µg y además este cumplió con los mínimos de calidad exigidos para llevar a cabo el proceso de secuenciación masiva.

2.3.1.2. | Secuenciación de la fracción microbiana

Como en el caso de PMLT01, en el caso de *L. lithistoides* se utilizó la tecnología de secuenciación masiva *Ion Torrent*, para así combinar un tamaño de lectura medio y una mayor profundidad de secuencia que cuando se secuenciaron los metagenomas mediante pirosecuenciación.

Las estadísticas relativas al resultado de la secuenciación masiva del metagenoma mediante *Ion Torrent* se muestran en la tabla 15, siendo el total de lecturas obtenidas de 6 086 661 con una media de longitud de 233 bp, lo cual supone un total de 1,4 Gb de secuencia. El contenido medio en GC del total del DNA metagenómico fue del 59%.

	Ion Torrent
Lecturas	6086661
Media de las lecturas (bp)	233
Nucleótidos totales (bp)	1420872091

Tabla 15 | Características del proceso de secuenciación masiva del metagenoma de PMLT01

2.3.1.3. | Análisis de la secuencia metagenómica microbiana en MG-RAST

Todas las lecturas obtenidas se analizaron de forma preliminar utilizando el servidor especializado MG-RAST (Meyer *et al*., 2008). De esta forma se obtuvieron las características generales del metagenoma microbiano de *L. lithistoides*, siendo el total de lecturas obtenidas de 6 086 661 con una media de tamaño de 233 bp (Fig. 53), lo cual abarca un total de 1,4 Gb. El contenido medio en GC del total del DNA metagenómico resultó ser del 59 % (Fig. 53).

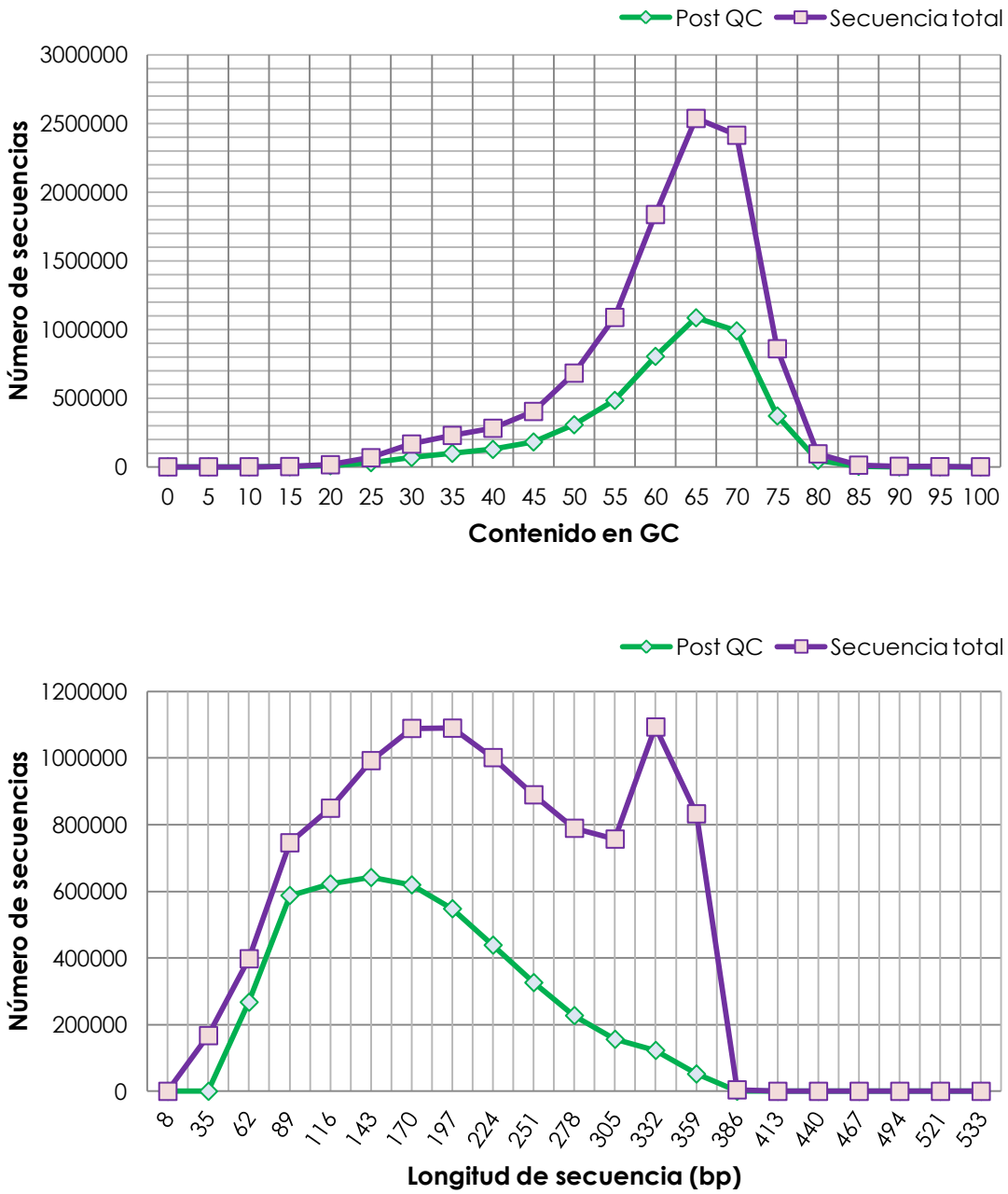

Figura 53 | Distribución de las secuencias del metagenoma de *L. lithistoides* según su longitud y contenido en GC. En morado se muestra la secuencia obtenida directamente de los dos procesos de secuenciación masiva mientras que en verde las secuencias que han pasado el control de calidad (QC) del servidor MG-RAST. En el gráfico superior se observa la distribución del número de secuencias según su porcentaje del contenido en G+C mientras que en el inferior se puede ver la distribución del número de secuencias obtenidas según la longitud.

2.3.1.3.1. | Resultados de la asignación funcional del total de la muestra

Del total de lecturas, únicamente un 75,7% superaron los controles de calidad de MG-RAST (4607643 secuencias). El 86,9% de estas secuencias filtradas dieron lugar a 2117024 regiones codificantes, de las cuales únicamente un 25,3% (535328 regiones) pudieron anotarse utilizando las bases de datos de proteínas del servidor (M5NR), mientras que el 74,7% restante no tenían similitud suficiente. El 76% de estas

características anotadas pudieron finalmente asignarse a las categorías funcionales definidas por el servidor (Fig. 54).

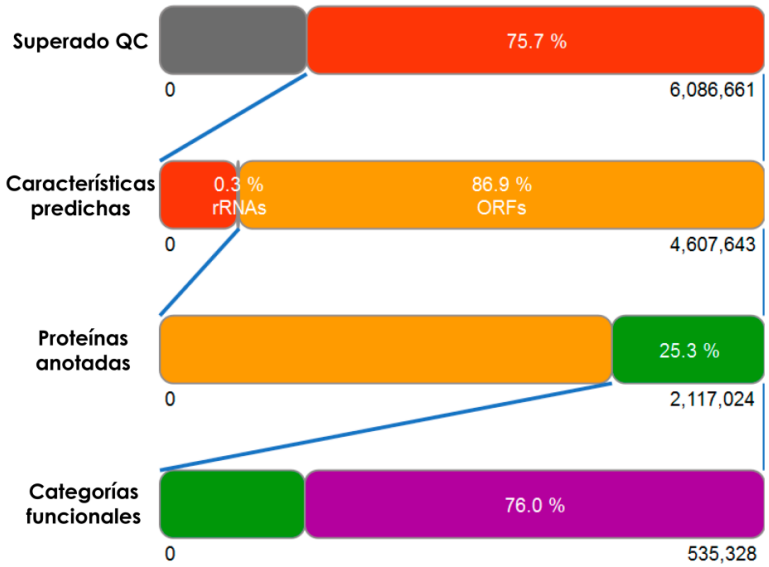

Figura 54 | Esquema del proceso de anotación en MG-RAST. En el esquema se puede observar la cantidad de secuencias procesadas en los pasos de control de calidad (QC), predicción de características, anotación de proteínas y asignación de categorías funcionales.

Tal y como se muestra en la figura 55 la distribución de funciones por subsistemas generada por MG-RAST, las más representadas excluyendo los grupos que recopilan categorías sin determinar, son aquellas relacionadas con el metabolismo de los carbohidratos, los aminoácidos y el metabolismo de las proteínas (Fig. 55).

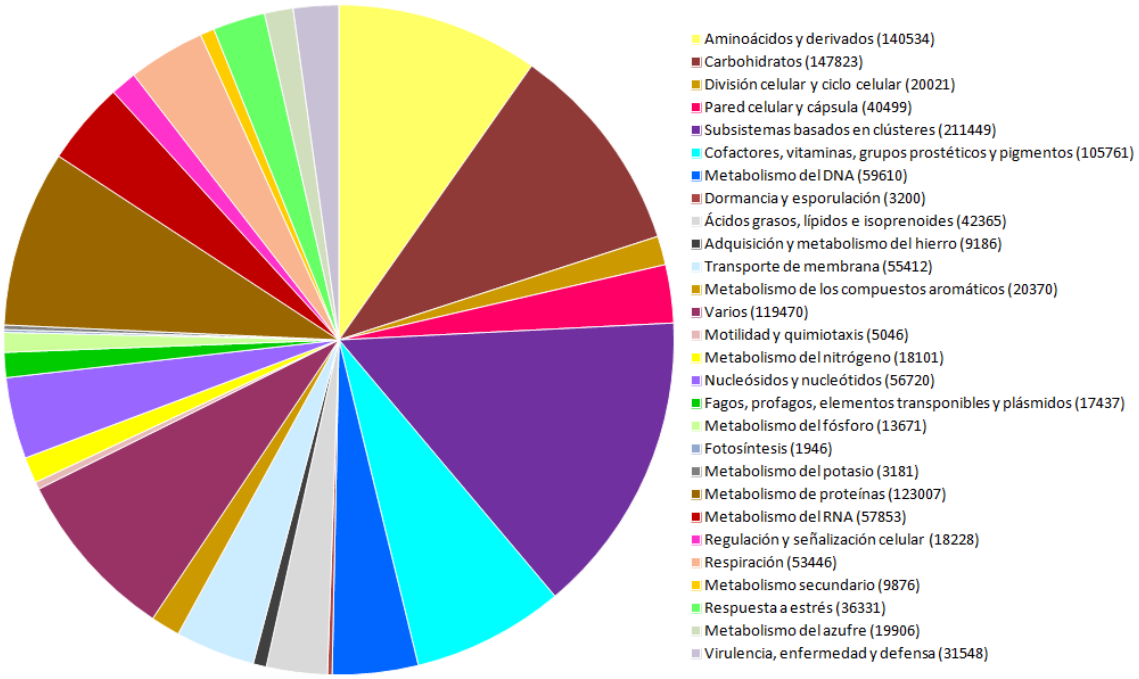

Figura 55 | Gráfica de la distribución por subsistemas de la anotación del metagenoma de *L. lithistoides* En el gráfico proporcionado por la herramienta MG-RAST se representan el número de positivos para cada uno de los subsistemas agrupados en categorías generales tras el proceso de anotación automática.

2.3.1.3.2. | Resultados de la asignación taxonómica del total de la muestra

A continuación se procedió a analizar los datos de asignación taxonómica generados por MG-RAST. De este modo se pudo obtener la clasificación taxonómica del total de las lecturas por dominios (Fig. 56). La mayoría de la secuencia resultó ser de origen bacteriano, existiendo también un 4,9% de secuencia de arquea y un 2,6% de eucariotas. Al profundizar más en la clasificación taxonómica se pudo obtener la distribución de las secuencias según su pertenencia a distintos phyla (Fig. 57), siendo el phylum Proteobacteria el más abundante tras otros phyla mayoritarios como Actinobacteria, Firmicutes, Chloroflexi y Acidobacteria.

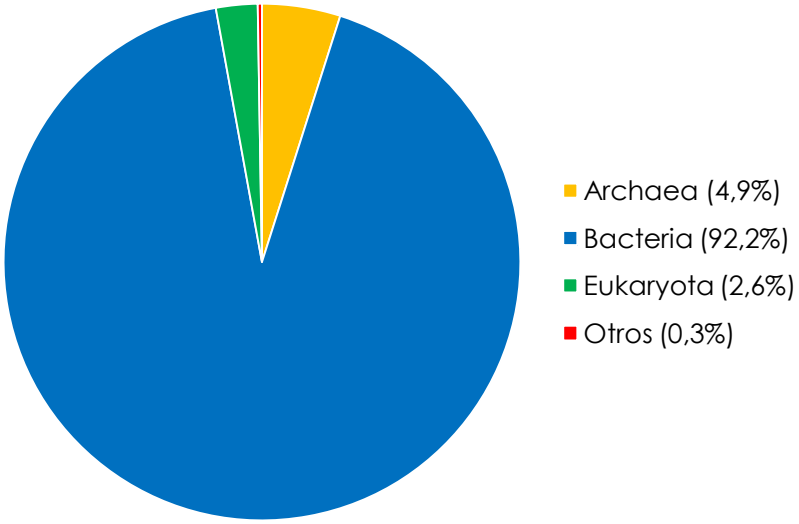

Figura 56 | Distribución por dominios de las secuencias del metagenoma de *L. lithistoides*

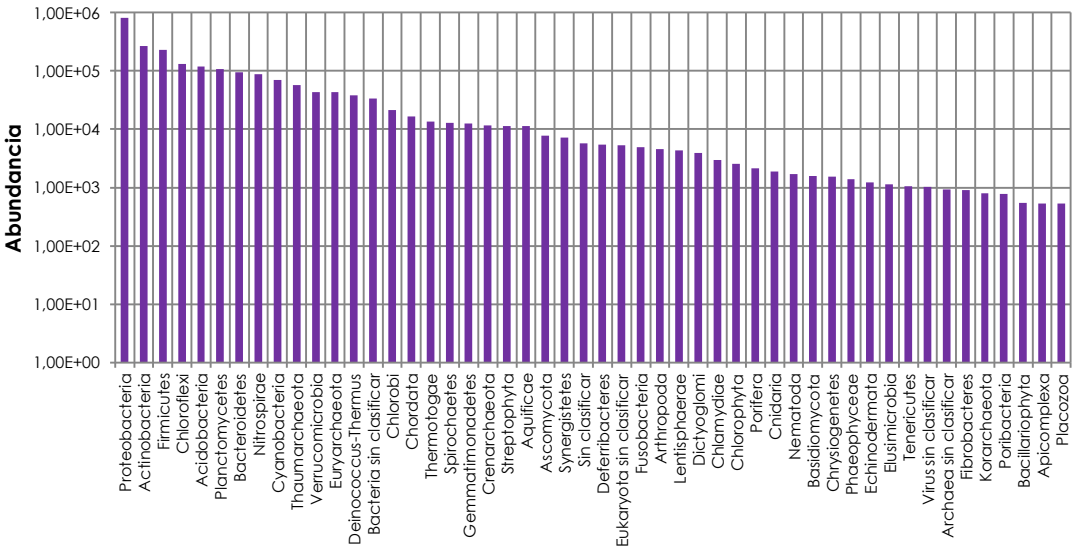

Figura 57 | Abundancia de cada phylum en las secuencias clasificadas de PMLT01

La curva de rarefacción generada de forma automática por MG-RAST (Fig. 58) indica el nivel de saturación de la secuencia metagenómica. Como se muestra en la figura 6 para el conjunto de las secuencias utilizado el número de especies distintas sería aproximadamente de 5100.

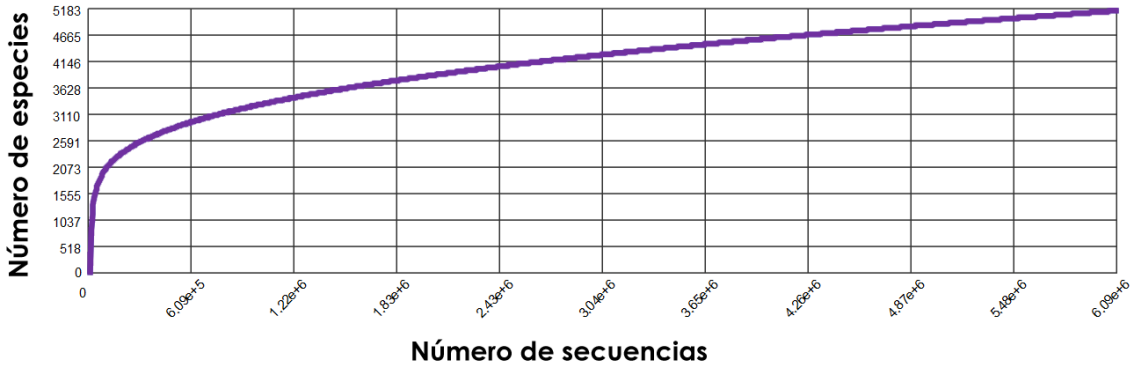

Figura 58 | Curva de rarefacción de las secuencias de *L. lithistoides* En la gráfica obtenida de MG-RAST se muestra el grado de saturación en especies de la muestra identificando el número de especies distintas conforme se van analizando las lecturas.

2.3.1.4. | Ensamblaje de la secuencia metagenómica de la fracción microbiana

Tras analizar el conjunto de las lecturas obtenidas se procedió a realizar un ensamblaje *de novo* de todas las secuencias para generar de este modo fragmentos más largos con mayor información y que con más probabilidad representaran la fracción más abundante. A diferencia de los ensamblajes anteriores, el ensamblaje de *Lithoplocamia* fue llevado a cabo por Lifesequencing S.L. mediante el software Newbler y utilizando toda la secuencia generada.

El ensamblaje obtenido poseía un total de 75120 *contigs* que abarcan 85494039 bp teniendo el *contig* de mayor tamaño una longitud de 245331 bp. A continuación se realizó un mapeo de las lecturas totales frente al ensamblaje realizado utilizando bowtie2 y se observó que un 73,79% de las lecturas estaban representadas en la secuencia ensamblada.

2.3.1.5. | Asignación taxonómica de la secuencia metagenómica ensamblada

Una vez ensamblados los fragmentos en secuencias más largas, fue posible realizar una asignación taxonómica más precisa. Al igual que en el análisis de las anteriores esponjas, en este caso se pudo separar y agrupar las secuencias según las distintas características estructurales previamente extraídas utilizando variaciones sobre el protocolo *in silico* de Albertsen *et al*. (2013) (ver Materiales y Métodos).

Una vez se extrajeron los datos de cada uno de los *contigs* ensamblados, se realizó una representación en la cual los más representados se agruparon según sus valores de abundancia y contenido en GC (Fig. 59). En esta figura destacan 7 agrupaciones nombradas con letras de A-G. Aquellos *contigs* que no parecen formar parte de ninguno de los cúmulos y además poseen valores de abundancia por encima 10 y de longitud por encima 8000 bp, se analizaron de forma manual utilizando herramientas de comparación

(HMMER y Blast). Del mismo modo también se analizaron *contigs* más pequeños con valores de abundancia de 25 o más.

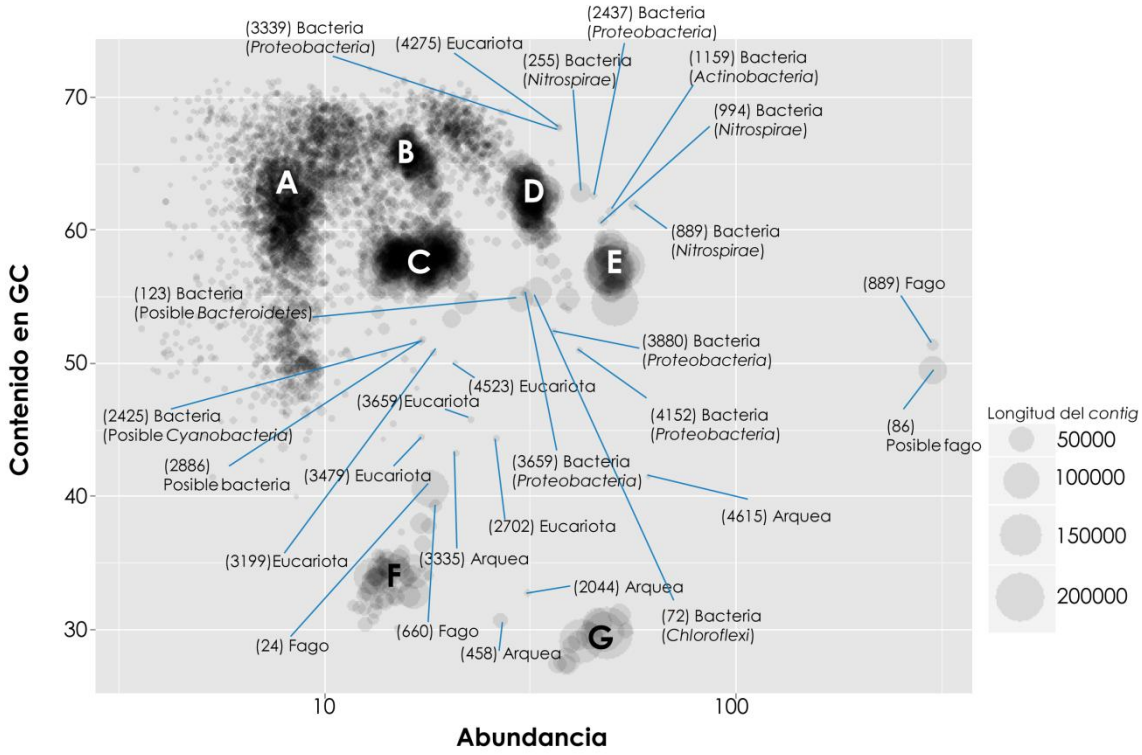

Figura 59 | Distribución de la secuencia ensamblada de *L. lithistoides* y análisis de los *contigs* individuales y representativos. En la distribución se representan los *contigs* mayores de 2500 bp según su contenido en GC y abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). Se muestran los resultados de asignaciones taxonómicas manuales con HMMER, y BLAST para aquellas secuencias no agrupadas en cúmulos principales, representadas más de un 10x y mayores de 8000 bp además de para *contigs* más pequeños con valores de abundancia de más de 25. Los cúmulos de secuencias más significativos se señalan con letras de la A-G.

Del conjunto de los agrupamientos señalados, del B-G tienen unos valores de abundancia suficientes como para poder sugerir que se traten de genomas individuales. Sin embargo, el cúmulo A se encuentra en una zona donde podría existir la presencia de *contigs* de genomas distintos ya que posee poca abundancia y se encuentra más disperso.

Del análisis de aquellos fragmentos que no parecen formar parte de ninguno de los cúmulos, se observó la presencia de dos secuencias fágicas muy representadas. También se apreció la existencia de secuencias pertenecientes a distintos phyla bacterianos, arqueas, eucariotas (posiblemente pertenecientes a la esponja hospedadora) y otras secuencias de fago con menos representación que las anteriormente comentadas.

A continuación se realizó la clasificación taxonómica mediante comparación de aquellos *contigs* en los que se detectó la presencia de genes esenciales. Estos datos generados se incluyeron en la representación de los *contigs* según su contenido en GC y su abundancia (Fig. 60). Si se observa en detalle, el cúmulo A tiene menos densidad de *contigs* con genes esenciales, además estos parecen ser de origen taxonómico muy diverso. En los cúmulos B, D y E, sin embargo, se aprecia más concentración de *contigs* con genes

esenciales y además la clasificación taxonómica es más uniforme, siendo clasificadas estas agrupaciones como posibles fragmentos de *Actinobacteria, Acidobacteria* y *Nitrospirae*, respectivamente. El cúmulo C posee una gran concentración de *contigs* con genes esenciales pero la clasificación taxonómica de estos no es homogénea, lo que podría indicar que contiene de más de un genoma. Finalmente, los cúmulos F y G poseen secuencias de arqueas, sugiriendo la posibilidad de que haya 2 tipos de arqueas distintas muy representadas, siendo el del cúmulo G mayoritario.

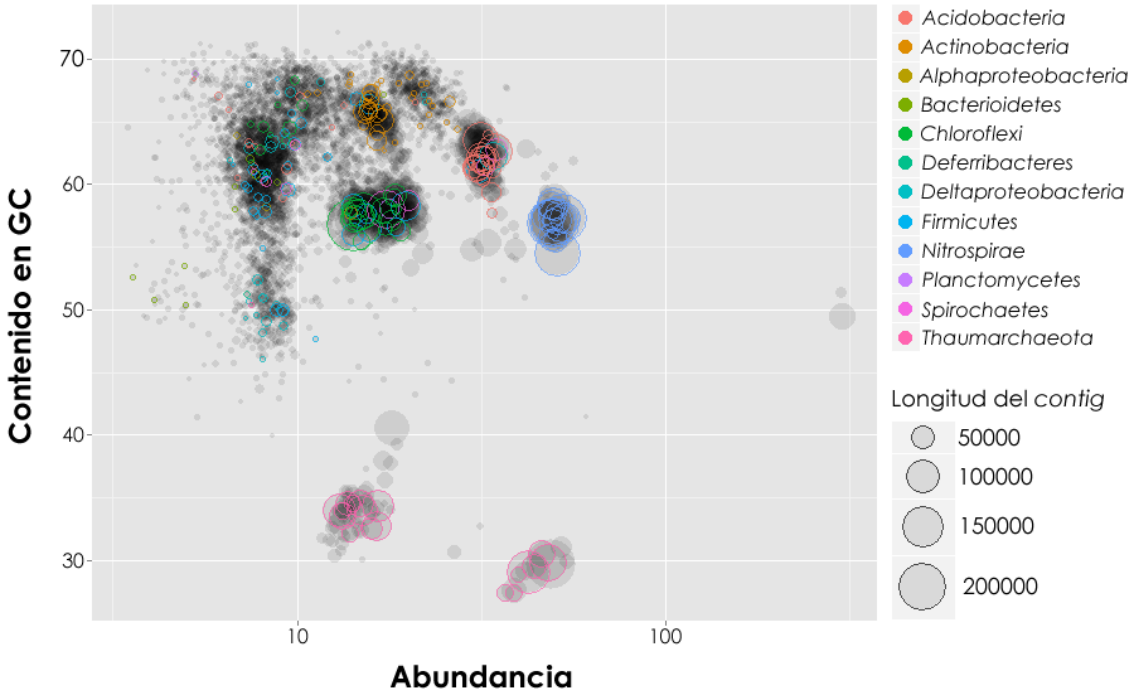

Figura 60 | Distribución de la secuencia ensamblada de *L. lithistoides* y análisis taxonómico de genes esenciales. En la distribución se representan los *contigs* mayores de 2500 bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). Dependiendo de la clasificación taxonómica obtenida se señalan en distintos colores aquellos *contigs* que contienen genes esenciales.

Además se realizó una clasificación taxonómica a nivel de phylum realizada con el algoritmo PhyloPythiaS+ del total de los *contigs* (Fig. 70).

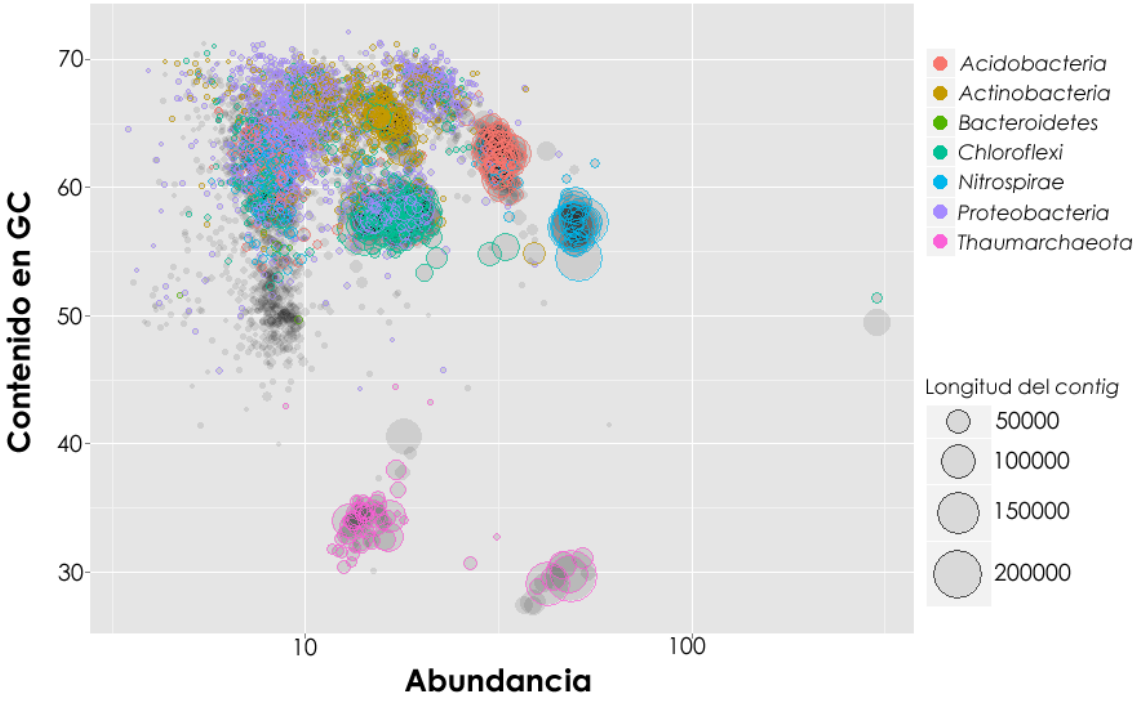

Figura 61 | Distribución de la secuencia ensamblada de *L. lithistoides* y análisis taxonómico mediante PhyloPythiaS+. En la distribución se representan los *contigs* mayores de 2500 bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). La clasificación taxonómica obtenida para cada *contig* se señala en distintos colores. Aquellos *contigs* que aparecen sin asignación taxonómica no pudieron ser clasificados por la herramienta PhyloPythiaS+.

Como se puede observar en la fig. 61, las agrupaciones B, D y E aparecieron definidas de forma homogénea a nivel taxonómico, siendo clasificadas como posibles Actinobacteria, Acidobacteria y Nitrospirae, respectivamente, coincidiendo con la predicción de genes esenciales (Fig. 60). El cúmulo A se mostró muy diverso desde el punto de vista taxonómico, lo que probablemente es otra evidencia de que contiene numerosos fragmentos de genomas distintos. El cúmulo C no fue homogéneo, sin embargo, puede observarse que sus secuencias se clasificaron principalmente como *Chloroflexi* y *Proteobacteria*, lo cual indica que posiblemente esta agrupación contenga en su mayoría, la mezcla de dos genomas distintos con un contenido en GC parecido y un nivel de representación similar. De este modo, se puede hipotetizar que con mucha probabilidad los grupos B, D, E, F y G representan un único genoma individual y que por el nivel de abundancia media que poseen, estas secuencias podrían estar completas. El cúmulo C contendría 2 genomas completos que no se pueden aislar separando por criterios de contenido en GC y abundancia, siendo necesario aplicar otros procedimientos para intentar aislar ambos conjuntos.

2.3.2. | Búsqueda de secuencias de PKSs y NRPSs en el metagenoma

Para identificar la presencia de genes que codifican PKSs y NRPSs que podrían estar implicados en la síntesis de policétidos se realizaron dos análisis, uno más general basado en la presencia de funciones de interés en el total de la secuencia metagenómica y otro en el cual se realizaron las búsquedas únicamente en la secuencia ensamblada

2.3.2.1. | Búsqueda secuencias de PKSs y NRPSs en el total de la muestra metagenómica

En primer lugar se realizó un recuento total de las funciones de interés relacionadas con clústeres PKS y NRPS, por lo que se llevaron a cabo búsquedas utilizando HMMER para los dominios acil transferasa (AT), de unión a AMP (incluido el específico del C-terminal), condensación (C), ceto-reductasa (KR), cétido sintasa (KS) (N-terminal y C-terminal), de sitio de unión a fosfopanteteina (PP), dehidratasa (DH) y tioesterasa (TE) tomados de la base de datos Pfam. Estas búsquedas se realizaron sobre el conjunto de la secuencia ensamblada, sobre las lecturas, y para evitar parcialmente la redundancia en dichas lecturas, también se hizo la búsqueda sobre el resultado de un proceso de *clusterización* (con un 98% de identidad) (ver tabla 16).

	Nombre Pfam	*Contigs*	Lecturas	Clustering
Acil-transferasa (PF00698.16)	Acyl_transf_1	68	404	345
Unión a AMP (PF00501.23)	AMP-binding	396	3833	3383
Unión a AMP C-terminal (PF13193.1)	AMP-binding_C	123	451	445
Condensación (PF00668.15)	Condensation	0	0	4
KS N-terminal (PF00109.21)	ketoacyl-synt	98	1237	867
KS C-terminal (PF02801.17)	Ketoacyl-synt_C	67	819	571
KR (PF08659.5)	KR	195	465	357
Sitio de unión de fosfopanteteina (PP) (PF00550.20)	PP-binding	24	115	162
Deshidratasa (PF14765.1)	PS-DH	31	198	180
Tioesterasa (PF00975.15)	TE	1	0	0

Tabla 16 | Recuento de la representación de dominios relacionados con clústeres PKS/NRPS en la secuencia de *L. lithistoides* Para cada dominio se señala entre paréntesis el número de acceso correspondiente en la base de datos Pfam. Se representan los recuentos para la secuencia ensamblada (*contigs*) para las lecturas y para las lecturas clusteriadas con un 98% de identidad.

Como se puede observar, el número de lecturas positivas en el conjunto, se encuentren *clusterizadas* o no, es mayor en todos los casos salvo en el único dominio TE encontrado, el cual se identifica únicamente tras realizar el ensamblaje. La tendencia en las muestras de las esponjas con las que se ha trabajado en esta Tesis Doctoral ha sido la de encontrar más positivos en el conjunto total de las lecturas que en el grupo de las secuencias *clusterizadas*, sin embargo, en el caso de los dominios C y PP, se encontraron más positivos en las lecturas *clusterizadas*. Esto puede deberse a que la reducción del tamaño de la base de datos permitió que los positivos, que se encontraban por debajo pero muy cerca del valor límite del *e-value,* pasaran los criterios de restricción tras este procesamiento.

El dominio más abundante en todos los casos es el de unión a AMP, el cual, por estar asociado a otras funciones, aparece más representado. Aunque el resultado indica que existen más positivos en el total de las secuencias que no han sido ensamblados debido a su longitud, estas secuencias no aportan la información suficiente para discernir si se trata de un dominio incluido en un clúster PKS/NRPS.

2.3.2.2. | Búsqueda de secuencias de PKSs y NRPSs en la secuencia metagenómica ensamblada

Para realizar el análisis se tuvieron en cuenta los *contigs* positivos detectados previamente mediante HMMER de más de 1000 bp de longitud. Seguidamente, se realizaron búsquedas mediante BLAST y la base de datos Pfam (ver Materiales y Métodos) para identificar de forma manual los componentes del contexto génico de cada uno de los fragmentos. Para aquellos dominios que resultan muy informativos a la hora de identificar la presencia de un clúster PKS/NRPS (como por ejemplo el caso de los dominios C o DH), se analizaron también *contigs* con un tamaño menor al establecido previamente. A diferencia de las otras esponjas, en este caso la secuencia ensamblada generó más cantidad de secuencias largas aptas para el análisis mediante la herramienta antiSMASH, por lo que se utilizó dicho recurso para analizar en paralelo de forma automatizada estos *contigs* de más longitud. Los fragmentos ensamblados en los que las búsquedas no ofrecieron indicios de pertenencia a un clúster de interés se desecharon manualmente. Todos los dominios de unión a AMP cuyos positivos eran mayoritarios fueron desechados por no aparecer formando parte de ORFs multifuncionales de interés.

Una vez seleccionadas las secuencias ensambladas de interés se procedió a realizar un ensamblaje manual realizando alineamientos mediante BLAST de los extremos 5' y 3' con el objetivo de aumentar la longitud de dichas secuencias lo máximo posible. La base de datos utilizada para realizar esta serie de ensamblajes manuales contenía la totalidad de las lecturas generadas del proceso de secuenciación masiva. Como resultado de este proceso se obtuvieron finalmente un total de 12 secuencias de interés (tabla 17).

Secuencia	Contigs	Dominios	Longitud
1	1159	KS/AT/DH/ER/KR/PP	7106
2	3280/4905/4971/5255	KS/AT/DH/ER/MT/KR/PP/KS/AT	10347
3	3901	AT/DH	3058
4	347/78	KS/AT/DH/MT/ER/KR/PP/KS/AT/DH/ MT/ER/KR/PP/KS/AT/ACPS	12468
5	255	KS/AT/DH/MT/ER/KR/PP/KS/AT	8556
6	6572	AT/DH/MT	4012
7	624/1633	KS/AT/DH/MT/ER/KR/PP/KS/AT	10350
8	125	KS/AT/DH/ER/KR/PP	7047
9	4091	KS/AT/PP/ACPS	3109
10	2555/6141/2195	AT/DH/MT/ER/KR/PP/KS	8872
11	3326	KR/PP/KS	3558
12	6181	KR/PP/KS	2553

Tabla 17 | Secuencias seleccionadas con dominios relacionados con PKSs y NRPSs. Para cada secuencia se indican los *contigs* que la contienen, la sucesión de dominios encontrada (con la notación correspondiente a la de la base de datos Pfam) y su longitud.

En esta ocasión algunas de las secuencias contenían lo que aparentemente podría ser identificado como clústeres completos. En concreto, las secuencias 1, 2, 4, 5, 7 y 8 tenían una estructura completa debido a que los dominios que aparecían en ambos extremos ya no estaban relacionados con este tipo de clústeres de interés.

Para situar en su contexto las 12 secuencias seleccionadas y comprobar si existía la posibilidad de que pertenecieran al mismo microorganismo se realizó un análisis de los

contigs involucrados en el cual se clasificaron taxonómicamente por comparación (BLAST). A continuación, estas secuencias de interés se representaron junto al resto de *contigs* atendiendo a su abundancia y contenido en GC (Fig. 62).

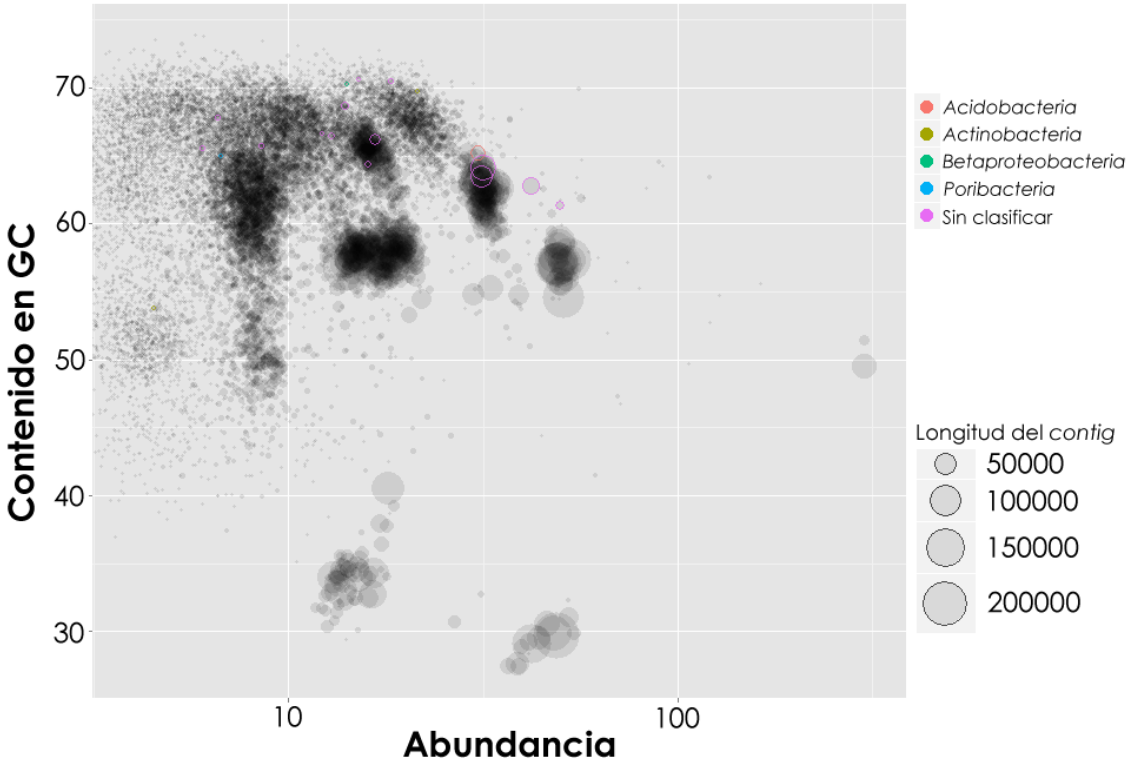

Figura 62 | Distribución de los *contigs* seleccionados con fragmentos con dominios PKS/NRPS en el conjunto de la secuencia ensamblada de *L. lithistoides* En la distribución se representan los *contigs* mayores de 1000 bp atendiendo a su contenido en GC y su abundancia. El tamaño del punto es relativo a la longitud del *contig* (nts). La clasificación taxonómica obtenida para los contigs que contienen dominios PKS/NRPS se señala con colores.

Como se puede comprobar en la figura, la mayoría de las 12 secuencias se situaron con un contenido en GC por encima de 60%. Al realizar la asignación taxonómica, todas las secuencias de interés resultaron tener un origen bacteriano, sin embargo, la procedencia de la mayoría de ellas no pudo asignarse por comparación, y no pudieron ser clasificadas. Al realizar BLASTx con la mayoría de estos fragmentos en la base de datos del NCBI (datos no mostrados), los resultados con mejor puntuación resultaron ser en casi todos los casos entradas de fragmentos clonados de simbiontes de esponjas como *Theonella swinhoei* y *Aplysina aerophoba* pertenecientes al trabajo de Fieseler *et al*. (2007).

La situación de estos fragmentos en la figura 62 respecto a los cúmulos de secuencias definidos anteriormente, hace pensar que la mayoría de las secuencias de interés no pertenecen a un mismo microorganismo ni tampoco a un mismo clúster biosintético. Únicamente en el caso del grupo D, en el cual vemos 3 secuencias de tamaño intermedio, y el B, el cual posee 2 pequeños fragmentos, podrían pertenecer a un mismo individuo.

VII. Discusión

1. Utilización de las herramientas genómicas para la obtención de la secuencia de un clúster PKS/NRPS de interés de un microorganismo cultivable

Para poder identificar la secuencia génica de un clúster de interés, contenido en un microorganismo cultivable, en esta Tesis se han utilizado herramientas genómicas relacionadas con la secuenciación masiva de DNA. Así, se ha secuenciado el genoma completo del microorganismo simbionte cultivable *T. mobilis* MES-10-09-028 y se ha llevado a cabo la identificación del clúster responsable de la síntesis de didemninas. Como ya se ha comentado en la Introducción los clústeres PKS/NRPS poseen una estructura constituida por módulos individuales de síntesis que hace posible realizar pronósticos acertados sobre la estructura general que debe poseer dicho clúster. Debido principalmente a la consistencia de estas propuestas de síntesis teóricas es factible detectar clústeres de producción aplicando herramientas bioinformáticas especializadas de análisis y anotación.

1.1. | Estructura del clúster de síntesis de didemninas

Tras analizar la estructura génica resultante del análisis de la secuencia obtenida de la secuenciación de *T. mobilis* MES-10-09-028 se pudo observar la distribución del conjunto de los dominios en los distintos módulos de síntesis (ver Resultados). Aunque existen coincidencias con la propuesta de síntesis previa (ver Resultados), la profundidad de nuestro análisis nos ha permitido identificar ciertos elementos que difieren claramente con el pronóstico realizado previamente. Estas diferencias radican principalmente en los módulos iniciales de la síntesis y en la ausencia de los dominios AT en los módulos PKS (ver Resultados, apartado 1.1.9). La presencia de los dos módulos iniciales sugiere que la estructura del clúster no concordaría con la síntesis de la molécula de didemnina B, y sí con la producción de de didemnina X e Y, lo que haría necesario un procesamiento para obtener la molécula de didemnina B. Según se postula, este tipo de mecanismos pueden ser ventajosos para las bacterias productoras, ya que se sintetiza primero un precursor no tóxico, que se procesa posteriormente, activándose en el exterior celular, donde por ejemplo cumpliría la función de eliminar otras bacterias competidoras en el hospedador (Riemer y Bode, 2014). Por lo tanto, para discernir la función biológica de las didemninas en *Tistrella* y en el posible organismo hospedador, puede resultar interesante el estudio de estos fenómenos de procesamiento de la molécula inicial.

En la fecha de inicio de realización de esta Tesis, existían pocos ejemplos de activación de moléculas sintetizadas por PKSs y NRPSs, siendo uno de los más paradigmáticos el caso del antibiótico xenocoumacina (Fig. 63), producido por la bacteria simbionte *Xenorhabdus nematophila* (Riemer *et al*., 2011). En esta bacteria, las moléculas de las prexenocoumacinas son producidas y posteriormente procesadas por la peptidasa codificada por el gen *xcnG*, la cual se une a la membrana interna del periplasma y procesa la molécula de prexenocoumacina eliminando la parte de la molécula que contiene un ácido graso para generar la molécula final que es exportada al exterior celular. Además de la xenocoumacina existen otros ejemplos descritos que se activan por un mecanismo similar basado en XcnG como son la zwittermicina (Kevany *et al*., 2009), la amicoumacina (Li *et al*., 2015) o la colibactina (Dubois *et al*., 2011) (revisado en Riemer y Bode, 2014). La

síntesis de xenocoumacina presenta similitudes con la síntesis de didemnina B, ya que se da el procesamiento de un lipopéptido fruto de una biosíntesis mixta PKS/NRPS, sin embargo en el caso del clúster de síntesis de didemninas, no existe un gen que codifique un homólogo de la peptidasa XcnG lo que sugiere un mecanismo diferente al de la xenocoumacina. Además, el hecho de que el clúster productor de didemninas posea 2 módulos iniciales que pueden añadir de forma iterativa residuos de glutamina al ácido graso, y que esta iteración pueda ser variable según se añadan 3 o 4 aminoácidos (ver Introducción), indica que posiblemente el mecanismo de hidrólisis encargado de procesar la didemnina B debe ser más permisivo que el mecanismo de activación de la xenocoumacina (Riemer y Bode, 2014). Así, las diferencias encontradas en el caso de la biosíntesis de didemnina pone de manifiesto la existencia de un mecanismo nuevo todavía no descrito.

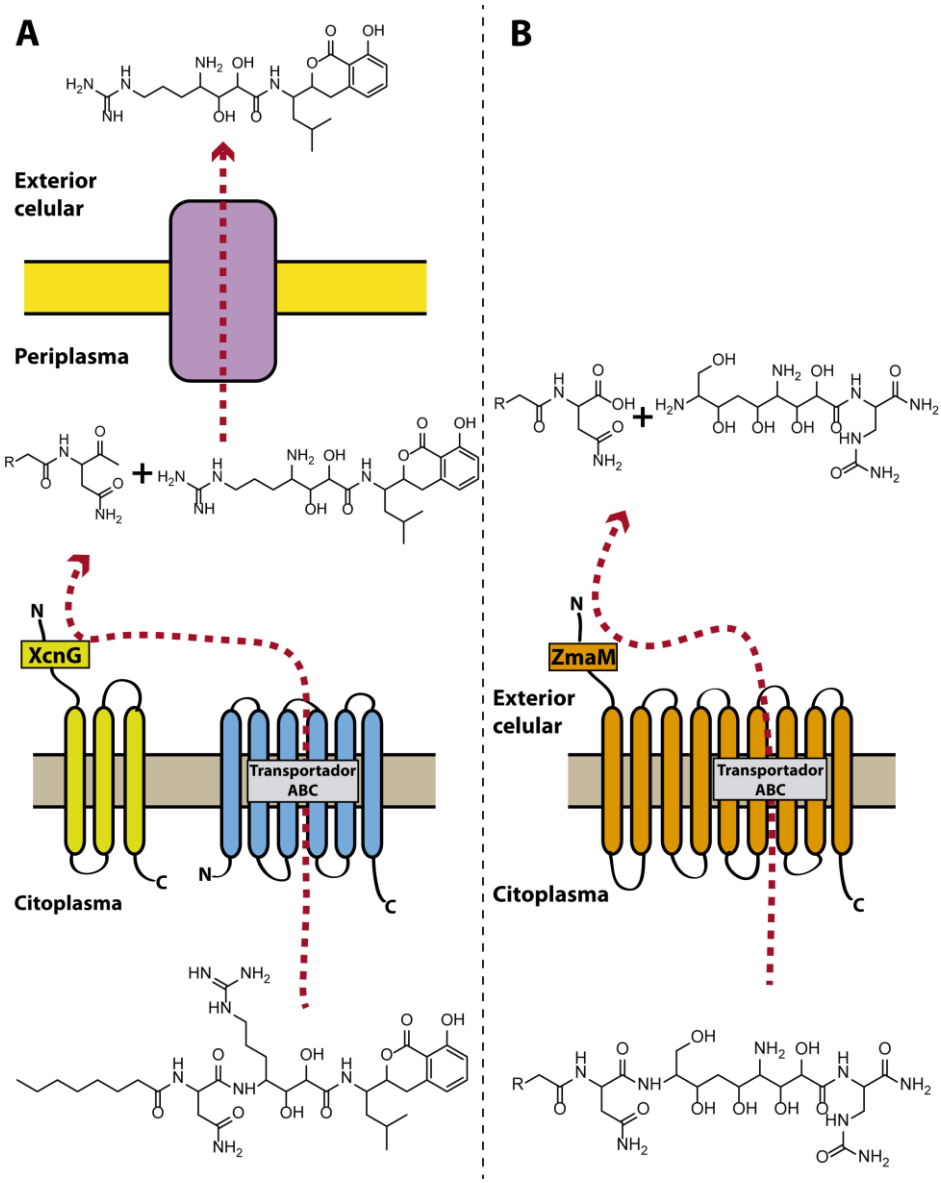

Figura 63 | Otros modelos de procesamiento de péptidos no ribosomales. A | Mecanismo de activación de la molécula de prexenocoumacina a xenocumaciona por una enzima XcnG independiente del transportador ABC propuesto. B | Mecanismo de activación de la molécula de prezwittermicina (con ácido graso (R) desconocido) a zwittermicina a por una enzima ZmaM fusionada a un transportador ABC. Modificado de Riemer y Bode (2014).

Tal y como proponen Xu *et al*. (2012), algunos de los genes adyacentes al núcleo del clúster de *T. mobilis* KA081020-065 podrían poseer las funciones necesarias para el procesamiento de las didemninas X e Y. Algunos de estos productos génicos tendrían funciones relacionadas con proteínas de membrana y con enzimas hidrolíticas. En el caso de *T. mobilis* MES-10-09-028 también aparecen genes similares en las inmediaciones del clúster (ver Resultados apartado 1.1.10.). Xu *et al*. (2012) se realizaron experimentos de localización de las moléculas mediante espectrometría de masas de imágenes MALDI (MALDI-imaging MS) sobre colonias de *T. mobilis*, los cuales sugirieron que existe un periodo inicial en el cual las moléculas precursoras, didemnina X y didemnnina Y, se excretan para posteriormente ser transformadas en didemnina B en el espacio extracelular. Por tanto, basándose en los resultados obtenidos, los autores proponen un modelo en el que la transformación ocurre en el exterior de la célula. En nuestro caso sin embargo, no se detecta producción extracelular de las didemninas precursoras en los caldos de cultivo de *T. mobillis* en medio de producción LB. Por otro lado nuestros resultados indican que la presencia de didemnina B en el interior de la célula es muy superior que en el ambiente extracelular (ver figura 17 en Resultados). Además, la concentración de didemninas a lo largo del tiempo en el interior de la célula, es consistente con el modelo que propone la transformación de las didemninas X e Y en didemnina B, ya que las primeras van desapareciendo para acabar detectándose únicamente didemnina B. Esto parece indicar que la transformación en didemnina B no se produciría únicamente en el exterior de la célula. Por lo tanto, los datos obtenidos en este estudio contradicen en cierto modo la propuesta realizada por Xu *et al*. (2012), indicando que el modelo de procesamiento propuesto es probablemente más complejo.

2. Análisis de los mutantes de *T. mobilis* en el clúster productor de didemninas

Frecuentemente, los microorganismos productores naturales no resultan aptos a la hora de producir un metabolito secundario de interés. Esto puede deberse principalmente a que pueden presentar dificultad en la manipulación genética, crecimientos lentos o en el caso de las adaptaciones industriales, pueden poseer características determinadas que no hagan factible el escalado de los procesos de fermentación (Baltz, 2006). Por lo tanto, para tratar de paliar estas limitaciones, la aproximación más utilizada es la producción heteróloga. Sin embargo, en ocasiones el productor nativo posee las características necesarias para ser manipulado en el laboratorio, como por ejemplo en el caso de la producción de daptomicina en *Streptomyces roseosporus* (Baltz *et al.*, 2004; Baltz *et al.*, 2006), aunque la mayoría de las veces estos microorganismos se encuentran filogenéticamente relacionados con los microorganismos más comunes utilizados en el laboratorio, y por lo tanto ya se disponen de herramientas cuyo uso se puede trasladar con relativa facilidad. En este estudio, para el caso de *T. mobilis* MES-10-09-028, aun perteneciendo a un taxón novedoso, se han obtenido unos niveles de crecimiento rápidos en medio rico y se ha conseguido modificar genéticamente mediante el uso de herramientas de biología molecular. Esto ha supuesto una gran ventaja, ya que ha generado la posibilidad de modificar la producción y además, estudiar con más detalle el

mecanismo de síntesis en un clúster PKS/NRPS complejo, como es el de producción de didemninas.

2.1. | La mutación KR3 en *T. mobilis*

El primer mutante generado de *T. mobilis* fue la cepa KR3 en la cual se ha realizado una sustitución en la tirosina catalítica del dominio KR del módulo 3 utilizando el vector suicida pK18KR3. El efecto de esta mutación redujo la producción de didemnina B y no se detectó la presencia de los precursores de didemninas X e Y. Este resultado sugiere que la función del dominio KR del módulo 3 es necesaria para la incorporación a la molécula del producto generado por los dos primeros módulos de síntesis. Esto se debe a que la reacción de ceto-reducción llevada a cabo por el dominio KR consigue que la molécula de piruvato que se incorpora en el módulo 3 pase a ser una molécula de lactato, la cual posee un grupo hidroxilo esencial para la condensación de esta nueva molécula incorporada con el resto de la cadena de síntesis (ver figura 24 en Resultados). Por tanto, el mutante resultante con una sustitución del residuo de tirosina catalítico por un residuo de fenilalanina en el módulo KR, debería ignorar el fragmento de cadena generado por los dos primeros módulos de síntesis y como consecuencia, produciría directamente la molécula de aplidina, la cual además, corresponde con la estructura teórica que se obtendría al eliminar el dominio KR del módulo 3 del clúster. El hecho de observar una reducción considerable, pero no total, de la producción de didemnina B, indica que la función del dominio KR no ha sido totalmente anulada. Al estar parcialmente inactivado el dominio KR, lo que teóricamente podría esperarse sería además de didemnina la aparición del análogo correspondiente generado por la falta de función del dominio KR. Sin embargo esto no ocurre, por lo que se puede concluir que la síntesis de una molécula alternativa no puede llevarse a cabo ignorando los módulos iniciadores, sugiriendo que la reacción de condensación en el módulo 3 es esencial para la síntesis.

Los resultados no coinciden con los obtenidos en los trabajos de Reid *et al.* (2003) con *E. coli* y de Power *et al*. (2008) con *Streptomyces nodosus*. En estos trabajos, se realizaron mutaciones puntuales en el sitio activo del dominio KR de los clústeres de síntesis, de este modo se sintetizaban nuevos análogos de 6-deoxieritronolida B y de anfotericina respectivamente. La mutación puntual en el sitio activo del dominio KR había sido suficiente en estos casos para causar un desorden estructural que inactivara dicho dominio. De hecho, en el caso del análogo 3-oxo de la 6-deoxieritronolida B en *E. coli*, se obtuvieron buenos rendimientos de producción al anular la actividad de uno de los dominios KR (Reid *et al.*, 2003). Sin embargo, en estos casos, al tratarse de módulos de PKSs, el grupo OH formado no intervenía en la reacción de condensación.

Otro resultado interesante obtenido apartir del estudio del mutante KR3 es, que a diferencia de la cepa silvestre, no se detecta la presencia de las didemninas X e Y. Sin embargo, la producción residual de didemnina B indica que forzosamente estas didemninas deben estar siendo sintetizadas aunque no sea posible su detección. La medida de la producción de didemninas fue tomada a las 96 h de cultivo, lo que quiere decir que la proporción de didemninas X e Y en ese estadio de crecimiento es mucho menor en relación con la didemnina B. Por lo tanto una de las posibilidades que se podrían plantear para explicar este hecho, es que al reducir drásticamente la cantidad de didemninas producida, la proporción de didemninas precursoras a las 96 h sea ínfima e

indetectable. Además la maquinaria que se encarga del procesamiento de las didemninas precursoras debería estar funcionando con unos niveles de actividad similares a los de la cepa silvestre, por lo que al haber menos didemninas X e Y estas podrían procesarse de forma inmediata, lo que quiere decir que este cuello de botella en la cepa KR3 se desplazaría a la acción del dominio KR mutante del módulo 3, el cual es más ineficiente que en la cepa silvestre.

2.2. | La mutación DidA en *T. mobilis*

Para seguir profundizando en el mecanismo de síntesis de didemnina B, se realizó una deleción del primer gen *ddnA*, que contenía los dos primeros módulos de síntesis, tanto en la cepa silvestre como en la cepa KR3 generando las cepas DidA y DidA-KR3 respectivamente (ver figura 27 en Resultados). El análisis de ambas cepas indica que a tiempos relativamente largos de producción (96 y 144 h) aun existen pequeñas trazas de didemnina B lo que sugiere que aunque sea con una eficiencia muy baja, el módulo 3 de síntesis del clúster puede actuar como módulo iniciador. En ninguna de las dos cepas se aprecia la presencia de análogos esperados como la aplidina lo que podría ser debido al hecho de que el dominio KR aún continua parcialmente activo y no permitiría derivar la síntesis a la de ningún posible análogo.

2.3. | Mecanismo de corrección en el clúster de síntesis de didemninas

Existen otros fenómenos que podrían explicar la falta de análogos en los extractos de los mutantes que estarían relacionados con los mecanismos de control internos del clúster. En ocasiones pueden ocurrir errores en las líneas de ensamblaje de PKSs y NPRSs que pueden formar cuellos de botella de intermediarios que no se han terminado de procesar. Como se comentó en la Introducción apartado 3.2.1.2, existen enzimas que poseen dominios TE de tipo II cuyos genes en ocasiones se encuentran en las inmediaciones de los clústeres PKS/NRPS (Du y Lou, 2010). Estas enzimas pueden actuar en *trans* corrigiendo fallos en las líneas de ensamblaje y liberando aquellas moléculas aberrantes que pueden llegar a bloquear el proceso de síntesis (Schwarzer *et al*., 2002). Por lo tanto, en módulos de NRPSs puede ocurrir que en ocasiones el dominio de adenilación falle al incorporar el aminoácido y el correspondiente dominio de condensación sea incapaz de reconocer el residuo que debe condensar (Yeh *et al*., 2004). En estos casos, la línea de ensamblaje quedaría bloqueada impidiéndose la síntesis y afectando el rendimiento total de producción de la bacteria (Fig. 64).

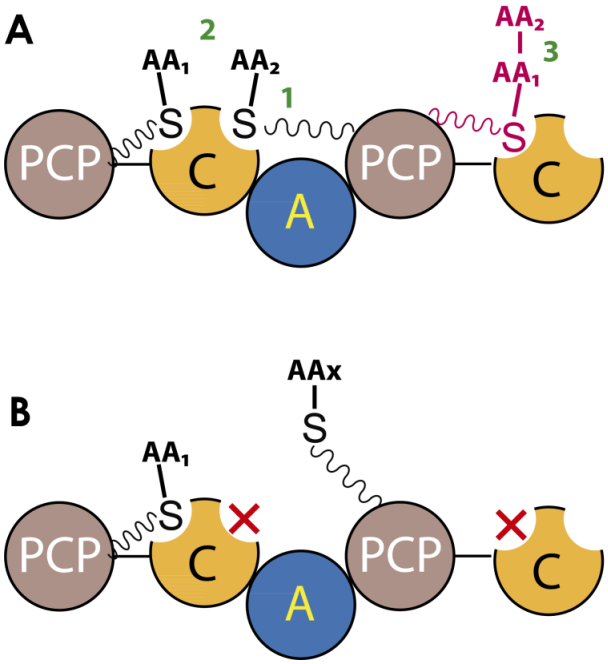

Figura 64 | Modelo del procesamiento de aminoácidos correctos e incorrectos por un módulo NRPS. A | Procesamiento normal en el que se lleva a cabo la condensación de un aminoácido incorporado (AA_2) con el aminoácido que proviene del módulo anterior (AA_1). El proceso incluye los pasos: 1. Selección e incorporación inicial del aminoácido. 2. Condensación de ambos aminoácidos. 3. Incorporación de la molécula resultante al siguiente paso de condensación. **B |** Bloqueo del módulo de síntesis por la incorporación de una molécula incorrecta (AA_X). "X" indica la incapacidad del dominio de condensación de reconocer la molécula incorporada.

En el caso concreto del clúster de síntesis de didemninas en *Tistrella*, en el análisis de las zonas adyacentes aparece una ORF (*orf4*) que contiene un único dominio TE de tipo II (ver Resultados apartado 1.1.10.), lo que sugiere que el clúster podría disponer de este tipo de mecanismos de corrección. Por lo tanto, la mutación KR3, al provocar que no se transforme la molécula de piruvato en lactato, podría repercutir en el reconocimiento de dicha molécula por el dominio de condensación que se encarga de unir esta nueva subunidad a la cadena que proviene de los dos primeros módulos. En el caso de los mutantes DidA, esto ocurriría directamente en el dominio de condensación que se encuentra a continuación del módulo, ya que como se puede ver en Resultados, el primer dominio de condensación del módulo 3 también fue eliminado en este mutante. De este modo, la molécula de piruvato sin el grupo hidroxilo permanecería bloqueando el proceso de síntesis hasta que fuese o bien metabolizada o retirada directamente de la síntesis por el mecanismo de corrección intrínseco del clúster.

2.4. | Desde la didemnina B a la aplidina

Como ya se ha comentado previamente (ver Resultados) además de intentar comprender mejor los mecanismos de síntesis de didemninas en *Tistrella*, los mutantes generados en este sentido tenían como objetivo secundario intentar generar una cepa de *Tistrella* capaz de generar un suministro de aplidina, compuesto con un valor clínico muy superior al de la didemnina B (ver Introducción, apartado 3.3).

Hasta la fecha de realización de esta Tesis, se desconoce qué organismo es el responsable de la producción de aplidina y en la naturaleza únicamente se ha detectado en

presencia del tunicado *A. albicans* (ver Introducción, apartado 3.3). Debido al gran parecido entre las moléculas de aplidina y didemnina B, cabe suponer que su síntesis pueda depender de clústeres génicos muy parecidos entre sí y que por lo tanto, el productor natural de esta molécula pueda tener relación con el género *Tistrella*. Sin embargo, aunque existen otros miembros del género *Tistrella* que pueden sintetizar didemninas, en ningún caso se ha detectado la presencia de aplidina. Por lo tanto, se propuso que modificaciones como las que se han desarrollado en esta Tesis y que paralelamente se sugieren en Xu *et al*. (2012), pudieran redirigir la síntesis de didemnina B a aplidina. Sin embargo, con los mutantes obtenidos, no ha sido posible obtener el suministro de aplidina deseado, por lo que o bien existen limitaciones que no se han tenido en cuenta aún para poder diseñar este proceso o bien podrían existir diferencias significativas entre los clústeres de producción no esperadas *a priori* dado el gran parecido de los dos compuestos. Una de las posibilidades consistiría en la existencia de diferencias significativas en los primeros módulos del clúster respecto a los módulos análogos del putativo clúster de síntesis de aplidina, por lo que con las modificaciones que se han propuesto en esta Tesis, tomando como punto de partida la síntesis de didemnina B, el rediseño del clúster con el objetivo de obtener aplidina no sería posible. Otra posibilidad implicaría que el mecanismo de síntesis de aplidina fuera muy distinto al de didemnina B y que pudiera encontrarse en un organismo taxonómicamente distante a *Tistrella* no identificado hasta la fecha. Una tercera opción consistiría en que la didemnina B fuese un precursor de la aplidina y que este paso de procesamiento se encontrara ausente en las cepas productoras de didemnina B que se han aislado. Partiendo de esta propuesta, caben dos posibilidades, una de ellas es que esta función se encuentre en el organismo productor, o bien, que la función esté presente en el ambiente donde habita este organismo. Esta última suposición podría ser también factible en el caso de *A. albicans*, ya que en este caso, podría existir una cepa de *Tistrella* productora de didemnina B y que alguna enzima del entorno, específica del ambiente de *A. albicans* podría actuar generando aplidina. Esta enzima, por ejemplo, podría ser un citocromo capaz de actuar de forma selectiva sobre el grupo hidroxilo del residuo de lactato, eliminándolo y generando un enlace de tipo éster en su lugar (Bernhardt, 2006). Un hecho que parece sustentar esta hipótesis es que al analizar el perfil de producción de didemninas en *A. albicans* se detecta de forma simultánea didemnina B y aplidina (PharmaMar, comunicación personal), por lo que probablemente la síntesis de aplidina podría estar basada en modificaciones de la molécula de didemnina y no directamente en organizaciones alternativas del clúster génico.

3. Optimización de herramientas para el clonaje y la producción heteróloga utilizando el ejemplo de la ruta completa de producción de didemninas

En esta Tesis se han generado genotecas en BACs a partir del DNA genómico de MES-10-09-028 y se han optimizando herramientas de cribado basadas en la amplificación de sondas, mediante las cuales, se ha conseguido identificar un clon que contiene el núcleo PKS/NRPS y la mayoría de genes colindantes del clúster de producción de didemninas (ver Resultados apartado 1.3.1.4). Este BAC denominado 13A10 fue trasladado al hospedador

heterólogo *E. coli* BAC-Optimized Replicator v2.0 observándose expresión transcripcional, lo que indicó, que no era necesario modificar las características reguladoras del sistema. Aunque los genes del clúster de la didemnina se expresan en *E. coli*, no se detectó la presencia de didemninas ni de ningún análogo de éstas en ninguna de las condiciones ensayadas en esta bacteria (ver Resultados apartado 1.3.2.2.1).

La producción heteróloga de clústeres PKS/NRPS de gran longitud presenta multitud de limitaciones (ver Introducción apartado 4.1.2.3.). Aunque existan varios ejemplos exitosos (revisado en Ongley *et al.*, 2013b), los resultados obtenidos del clonaje heterólogo de la ruta de producción de didemninas ponen de manifiesto este hecho.

Existen múltiples cuellos de botella que suelen dar como resultado la ausencia de producción. Uno de los principales es la ausencia de la regulación adecuada de la expresión en el hospedador heterólogo. Estos fenómenos suelen solucionarse con el intercambio del promotor nativo por otro cuyas condiciones de expresión se conocen en el hospedador heterólogo, lo cual es una práctica muy extendida a la hora de realizar intentos de producción heteróloga (Fu *et al*., 2008; Ongley *et al*., 2013a). De este modo se elimina el posible efecto de los reguladores ya contenidos en el clúster y se controlan las condiciones determinadas de expresión. Sin embargo, en el caso del clúster productor de didemninas, se decidió en primera instancia mantener el promotor nativo y posteriormente se detectó la expresión de genes del clúster en condiciones de producción, por lo que en este caso se puede presuponer que la ausencia de producción no se debe a la presencia de fenómenos de regulación.

El microorganismo utilizado en este estudio para abordar la producción heteróloga de didemninas fue *E. coli*. Este microorganismo fue elegido porque presenta una serie de características óptimas que proporcionaron ventajas en la puesta a punto de las herramientas que se presentan en esta Tesis. Entre estas características se encuentra la gran disponibilidad de herramientas de biología molecular que existen para este microorganismo (entre las que se encuentran una gran variedad de vectores desarrollados), la disponibilidad de protocolos de cultivo, la facilidad de su uso en el laboratorio o el hecho de que sea un organismo modelo en ingeniería genética y metabólica (revisado en Rodríguez *et al*., 2009).

Otros fenómenos que en este caso podrían estar afectando a la producción de didemninas son los siguientes:

3.1. | Especificidad de las PPTasas del microorganismo hospedador en la activación heteróloga de clústeres PKS/NRPS

Como ya se ha comentado en la Introducción, la función de las enzimas PPTasas es esencial para la activación de los clústeres biosintéticos PKS/NRPS. En el caso de *E. coli*, se conocen 3 PPTasas codificadas por los genes *acpS*, *entD* y *acpT* (formalmente *yhhU*) (Polacco y Cronan, 1981; Lambalot *et al*., 1996; Flugel *et al*., 2000). *EntD* es una PPTasa de tipo *Sfp* que interacciona con enzimas PKSs y NRPSs (Beld *et al*., 2014; Nakano *et al*., 1988). Sin embargo, muestra generalmente unos niveles funcionales insuficientes en los procesos de expresión heteróloga de moléculas sintetizadas por PKSs y/o NRPSs, ya que no es capaz de activar aquellos dominios ACP y PCP no nativos de forma eficiente (Sunbul *et al*., 2009). Por esta razón, para lograr la producción heteróloga en *E. coli*, es común

expresar a la misma vez PPTasas más promiscuas que sean capaces de interaccionar con un abanico mayor de dominios ACP y PCP, como por ejemplo *Sfp* (Watanabe *et al*., 2006) o *MtaA* (Gaitatzis *et al*., 2001; Bian *et al*., 2009).

En el caso concreto de la síntesis de didemninas en *E. coli*, se realizó el clonaje en un vector de la posible PPTasa localizada en las inmediaciones del clúster nativo de *Tistrella*. Este vector llamado pSEVAPPT (que posee un sistema de expresión inducible) se trasladó a la cepa de *E. coli* B13A10 y se coexpresó de forma heteróloga el gen que codifica la PPTasa junto con los genes del clúster productor de didemninas contenidos en el BAC 13A10 (ver Resultados, apartado 1.3.2.3.). Sin embargo, tampoco se detectó producción heteróloga de didemninas en esta construcción. Este resultado sugiere que la falta de producción de didemninas en *E. coli* no estaría relacionada con la ausencia de actividad PPTasa. Sin embargo, también se ha de tener en cuenta factores como que la enzima coexpresada aún no esté caracterizada como PPTasa y que por lo tanto no se haya demostrado su función.

3.2. | Limitaciones metabólicas del organismo hospedador

Las capacidades metabólicas del microorganismo productor son determinantes para la producción heteróloga de este tipo de compuestos. En concreto, para producir didemninas, desde el punto de vista metabólico, principalmente es necesario el aporte de algunos aminoácidos que teóricamente están disponibles en la cepa *E. coli* hospedadora. Sin embargo, los precursores X e Y de la síntesis de didemnina B son lipopéptidos. Esto se debe a la incorporación de un β-hidroxiácido graso (β-hidroxidecanoico) en las etapas iniciales de la síntesis. Al realizar un análisis de las capacidades metabólicas de *E. coli* en las bases de datos KEGG y Biocyc, se observó la presencia teórica de este metabolito intermediario en el metabolismo de los ácidos grasos. Aún así, el hecho de que *E. coli* pueda llegar a metabolizar este metabolito no significa que la concentración existente en el interior de la célula sea suficiente para iniciar la producción de didemninas a niveles detectables, por lo que la disponibilidad de este precursor, podría estar limitada en el microorganismo hospedador.

Este tipo de aproximaciones basadas en biología de sistemas resultan ser una herramienta útil para estudiar la capacidad metabólica de los microorganismos hospedadores a la hora de realizar producciones heterólogas de metabolitos secundarios de interés, o mejorar dichas producciones según intereses biotecnológicos (Breitling *et al*., 2013).

3.3. | Otras incompatibilidades con la maquinaria celular del hospedador

Tal y como se ha comentado anteriormente (ver Introducción apartado 3.2.5.), algunos clústeres PKS/NRPS tienen un gran tamaño. Debido a esto, es de suponer que la maquinaria celular de los productores nativos pueda poseer algún tipo de adaptación para tratar con complejos macroproteicos de unas dimensiones excepcionalmente grandes, como por ejemplo la presencia de chaperonas especializadas. Por lo tanto, en un proceso de producción heteróloga, como el que se presenta en esta Tesis, este hecho puede suponer limitaciones en la producción del metabolito de interés. Para tratar de contrarrestar estas deficiencias existen casos de producción heteróloga en los que se han

llegado a coexpresar chaperonas con la finalidad de mejorar estos procesos (Zhang *et al*., 2010).

Además del tratamiento de grandes complejos de proteínas, el microorganismo hospedador también debe enfrentarse a otro tipo de limitaciones. Estas por ejemplo pueden tener que ver con la estabilidad del RNA, el contenido en G+C del organismo o la diferencia en el uso de codones (revisado en Rodríguez *et al.*, 2009). En relación con esto, se conocen casos de traslado de rutas de PKSs y NRPSs en los que el éxito de la aproximación radica en la compatibilidad del microorganismo hospedador y donador en este tipo de factores (Cole Stevens *et al*., 2010). Sin embargo, aunque la relación filogenética no sea del todo cercana, *Tistrella* y *Escherichia* son bacterias Gram negativas y su contenido en GC no es extremadamente dispar (67,6% y 50,8%, respectivamente) por lo que se puede suponer, aunque no asegurar, que este tipo de factores no sean los limitantes a la hora de no obtener producción.

Aunque en esta Tesis no se ha conseguido detectar la producción de didemninas en un microorganismo heterlogo como *E. coli* se han optimizado herramientas para el traslado y la expresión del clúster logrando que la producción heteróloga pueda ser posible en un futuro.

4. *T. mobilis* como posible chasis para la expresión de clústeres PKS/NRPS

Como se ha demostrado en esta Tesis, aunque las características de *E. coli* son las óptimas para la producción de metabolitos secundarios, no ha resultado ser el huésped heterólogo idóneo para obtener producción de didemnina. Por ello, planteamos la posibilidad de desarrollar otros hospedadores heterólogos especializados en la producción de moléculas sintetizadas por PKSs y/o NRPSs.

Las bacterias del género *Streptomyces* han sido consideradas como productoras tradiconales de metabolitos secundarios. Debido a esto sus rutas de síntesis han sido trasladadas a otros organismos y ellos mismos han sido utilizados como hospedadores (Baltz *et al*., 2006a; Penn *et al*., 2006; Eustáquio *et al*. 2004). Sin embargo, existen muchos otros taxones no relacionados con las Actinobacterias que poseen microorganismos productores de moléculas de interés, por lo tanto, es necesario el desarrollo de nuevos sistemas de expresión heteróloga de clústeres PKS/NRPS.

Analizando los resultados de esta Tesis decidimos proponer a *T. mobilis* MES-10-09-028 como un microorganismo interesante para ser utilizado en este tipo de aproximaciones. En concreto, *T. mobilis* MES-10-09-028 posee al menos 2 clústeres complejos de tipque codifican PKSs y/o NRPSs (ver Resultados, apartado 1.1.6), habiéndose estudiado los de producción de didemninas y thalasospiramidas (Ross *et al*., 2013). Esto se demostraría que *Tistrella* posee un metabolismo secundario optimizado para la producción de este tipo de compuestos. Además, se trata de un microorganismo de crecimiento rápido, cuyos protocolos de cultivo y de modificación genética se han puesto a punto en esta Tesis por lo que es apto para el uso en el laboratorio. Por lo tanto, la

utilización de *T. mobilis* MES-10-09-028 como chasis celular para la producción de metabolitos secundarios podría aportar muchas ventajas metabólicas con respecto a *E. coli*, además de permitir el uso de herramientas de ingeniería genética y biología molecular, por lo que de este modo, se podrían poner a punto procesos de producción nativos de organismos no relacionados con las Actinobacterias (especialmente bacterias Gram-negativas). Otra aplicación muy interesante que se derivaría de estas propiedades, es la utilización de *Tistrella* para producir moléculas sintetizadas por PKSs y NRPSs cuyos clústeres de síntesis se encuentran silenciados en los organismos nativos (Bode y Müller, 2005).

Por lo tanto, los resultados obtenidos y los protocolos desarrollados en esta Tesis, nos permiten sugerir a *T. mobilis* MES-10-09-028 como un chasis ideal para abordar la producción heteróloga de metabolitos secundarios.

5. Eficiencia de las herramientas metagenómicas desarrolladas para la búsqueda de clústeres PKS/NRPS en metagenomas de esponja

5.1. | Consideraciones del procesado de las muestras ambientales

Como se ha comentado en cada caso, una vez recolectadas, las muestras de esponja fueron inmediatamente congeladas. En esta Tesis como en otros muchos estudios (Piel *et al*., 2004; Schirmer *et al*., 2005; Kennedy *et al*., 2009), este procedimiento se realizó por defecto para facilitar el procesamiento y el transporte desde localizaciones remotas. Sin embargo, el hecho de congelar la muestra puede afectar a la integridad de la misma provocando sesgos en los resultados obtenidos, ya que algunos microorganismos podrían no resistir este proceso y por lo tanto su material genético podría degradarse. Este tipo de consideraciones deben tenerse en cuenta a la hora de analizar los resultados de las poblaciones microbianas que aquí se presentan.

Otros posibles sesgos que se han cometido en la preparación de las muestras podrían estar relacionados con la naturaleza misma del tejido de la esponja. Como se ha observado anteriormente (verMateriales y Métodos, apartado 4.4.2.1.), las morfologías tisulares de las tres esponjas procesadas son muy diferentes entre sí, por lo tanto, esta morfología podría afectar al proceso de obtención de la fracción microbiana y a la extracción posterior de DNA. En concreto, el tejido de la esponja *P. littoralis* parecía tener pequeñas espículas calcáreas, las cuales, en el proceso de homogenización mecánica del tejido, se disgregaron, actuando probablemente, como un agente extra de lisis que pudo afectar a la integridad de algunas células microbianas. Otro ejemplo está presente en la extracción de *Lithoplocamia*, ya que tras la homogenización mecánica, la viscosidad de la extracción impidió el proceso de filtrado, la centrifugación y afectó a la calidad de la extracción de DNA, haciendo necesario un paso extra de ultracentrifugación (ver Materiales y Métodos, apartado 4.4.2.2.). Por lo tanto, este tipo de características intrínsecas han provocado que sea necesario utilizar protocolos de extracción de DNA individualizados para cada una de las muestras de esponja, por lo que para analizar y comparar, por ejemplo, los datos de rendimiento de la extracción del material genético o

las distribuciones de la población microbiana de la muestra, se deberían tener en cuenta estas diferencias en el tratamiento, la cuales podrían actuar a modo de sesgos.

Por lo tanto, teniendo en cuenta estas posibles causas, al comparar el rendimiento de las extracciones de DNA de los metagenomas de las tres esponjas procesadas queda patente que en *Polymastia*, se obtuvo una cantidad de material genético por gramo de tejido mucho menor que en los otros dos casos. De hecho, uno de los posibles indicadores de este fenómeno es la presencia en el metagenoma de la secuencia mitocondrial completa del propio hospedador cuando esto no ocurre en los otros casos. Como se ha mencionado en la Introducción, la proporción de células bacterianas en las esponjas marinas suele ser mucho mayor que la de células de la propia esponja, por lo tanto, que se haya obtenido una representación semejante de la secuencia mitocondrial indicaría que en este caso, en el momento de la lisis, la cantidad de microorganismo sería mucho menor que la esperada.

5.2. | Análisis de la distribución de las poblaciones de las esponjas *P. littoralis*, PMLT01 y *L. lithistoides*

En la presente Tesis y con el objetivo de intentar identificar la presencia de clústeres PKS/NRPS en las secuencias metagenómicas, se utilizaron aproximaciones en las que se analizaban las secuencias que podían ser ensambladas, asumiendo que estas pertenecería a los microorganismos mayoritarios (ver Resultados). De los datos de porcentajes de lecturas ensambladas, se puede deducir que en las tres muestras procesadas, una gran proporción de las secuencias obtenidas pertenecen a unos pocos microorganismos concretos. De estos ensamblajes se ha podido obtener la secuencia prácticamente completa de varios organismos, algo que también se ha conseguido en otros estudios (Hallam *et al*., 2006; Gao *et al*., 2014). El hecho de que estos microorganismos se encuentren en una proporción tan elevada provocaría un fenómeno de enmascaramiento que impediría observar otras poblaciones más minoritarias con la cantidad de secuencia obtenida.

Estos microorganismos mayoritarios parecen ser simbiontes auténticos cuyos niveles de abundancia podrían indicar una conexión metabólica importante con el hospedador. Un ejemplo de esto es el caso de la arquea presente en la esponja PMLT01 (ver Resultados, apartado 2.2.3.), la cual podría estar desempeñando un papel importante en el metabolismo del nitrogeno de la esponja hospedadora metabolizando en este caso el exceso de urea, tal y como se propone para simbiontes parecidos en otras esponjas (Hallam *et al*., 2006; Walker *et al*., 2010).

Si se comparan las clasificaciones taxonómicas proporcionadas por la herramienta MG-RAST para cada una de las tres esponjas analizadas (ver tabla 18) se puede observar como en concreto *P. littoralis* posee unos perfiles de diversidad que se alejan de los de las otras dos esponjas. Entre otras cosas, llama la atención la gran proporción de proteobacterias encontradas en esta esponja así como la presencia relativamente alta de fragmentos de Cyanobacteria, o de secuencias víricas o de arqueas. Otro dato interesante es la relativamente alta proporción de secuencias pertenecientes a organismos eucariotas, algo que puede explicarse por la presencia en este caso de la mitocondria de la propia esponja.

Al comparar la asignación de las secuencias de PMLT01 y *L. lithistoides* aunque son parecidas entre sí, llama la atención la gran proporción de Actinobacteria y Planctomycetes que posee PMLT01 frente a *L. lithistoides*. Sin embargo, *L. lithistoides* posee más secuencias de Acidobacterias y Nitrospirae, algo que se explica al observar la clasificación taxonómica de la secuencia ensamblada (ver figura 61 en Resultados), en la cual se puede observar dos cúmulos de secuencia muy representada con la misma clasificación que probablemente correspondan a dos organismos individuales mayoritarios.

Como se ha podido apreciar, aunque se puedan sacar diversas conclusiones, a la hora de analizar estos datos, se debe tener en cuenta el hecho de que la mayoría de la secuencia pertenecerá a los microorganismos más abundantes del microbioma por lo que la diversidad del metagenoma no se refleja de forma tan realista como utilizando otras aproximaciones como la secuenciación de amplicones del gen 16S rRNA.

Phylum	*P. littoralis*	PMLT01	*L. lithistoides*	Media
Proteobacteria	54,50	30,38	35,16	40,01
Actinobacteria	9,66	15,30	11,53	12,17
Firmicutes	4,46	10,43	9,89	8,26
Planctomycetes	1,46	10,61	4,57	5,55
Chloroflexi	1,04	5,48	5,66	4,06
Cyanobacteria	5,46	3,17	3,04	3,89
Bacteroidetes	2,51	4,78	4,09	3,79
Thaumarchaeota	6,34	1,13	2,47	3,31
Acidobacteria	0,35	3,34	5,08	2,92
Euryarchaeota	1,10	2,16	1,86	1,71
Nitrospirae	0,16	0,75	3,76	1,56
Verrucomicrobia	0,57	2,13	1,88	1,52
Deinococcus-Thermus	0,44	1,54	1,63	1,20
Bacteria sin clasificar	0,56	1,31	1,43	1,10
Chordata	1,57	0,31	0,71	0,86
Virus sin clasificar	1,76	0,08	0,04	0,63
Eukaryota sin clasificar	1,09	0,21	0,23	0,51

Tabla 18 | Comparación de phyla mayoritarios representados en los metagenomas de las tres esponjas estudiadas. En la tabla se indica en porcentaje de las lecturas la representación de aquellos phyla representados al menos un 1% en alguno de los tres metagenomas.

5.3. | Eficiencia de las herramientas *in silico* de detección de PKSs y NRPSs

La herramienta diseñada en esta Tesis en la cual se incorpora la detección y el análisis de fragmentos de genes que codifican PKSs y NRPSs a un sistema espacial de clasificación de secuencias (Albertsen *et al*., 2013) ha permitido distinguir fragmentos pertenecientes a clústeres biosintéticos de metabolitos secundarios y deducir si su procedencia puede ser común. De este modo, se han encontrado en los casos de *Polymastia* y *Lithoplocamia*, fragmentos que probablemente pertenezcan al mismo microorganismo y

en ocasiones al mismo clúster de biosíntesis. Sin embargo, únicamente en el caso del *Polymastia*, los fragmentos encontrados podrían coincidir con la propuesta de síntesis del compuesto de interés. Este hecho, por lo tanto, demuestra que la asunción de que un compuesto detectado esté siendo sintetizado por uno de los microorganismos mayoritarios no es del todo cierta. De hecho, la detección del posible clúster de interés en el metagenoma de *Polymastia* se realiza en un clúster de secuencias incompleto, por lo que faltan fragmentos para obtener una estructura completa (ver Resultados, apartado 2.1.2.2.). En este caso el compuesto de interés estaría siendo sintetizado por un microorganismo en un segundo nivel de abundancia del que no se conoce el genoma completo. Sin embargo, para los otros dos metagenomas, aun disponiendo de más secuencia que en el caso de *Polymastia*, no se detecta ningún clúster de biosíntesis de los compuestos de interés en la totalidad de la secuencia ensamblada. Aun así no se puede descartar la posibilidad de que estas ausencias en la detección se deban a algunos de los sesgos en el procesamiento de la muestra antes comentados.

Cuando se observa el número de positivos para dominios relacionados con clústeres PKS/NRPS en las lecturas obtenidas para cada uno de los metagenomas, se puede observar que en general aparecen más positivos en las lecturas que en las secuencias ensambladas (ver Resultados tablas 8, 12 y 16). Además, es lógico pensar que existan lecturas positivas que no estén representadas en los *contigs*. Estos fragmentos, que podrían contener registros de organismos menos representados no se han tenido en cuenta a la hora de analizar la presencia del clúster de interés. Esto se debe a que las herramientas desarrolladas no pueden trabajar con secuencias de un tamaño tan pequeño. La longitud y la naturaleza repetitiva de los módulos de PKSs y NRPSs dificulta mucho la asignación de fragmentos de pequeño tamaño a una propuesta de síntesis concreta, por lo que en la mayoría de los casos, en una lectura no se dispone de secuencia suficiente para encontrar las señas de identidad del clúster que permitirían la comparación con la propuesta de síntesis.

Por lo tanto, dado que es más difícil y costoso acceder a las poblaciones menos representadas de un metagenoma, las herramientas desarrolladas en este Tesis estarían optimizadas para identificar clústeres de interés pertenecientes a organismos mayoritarios, pero en el caso de que se dispusiese de una profundidad de secuencia suficiente, podrían también ser utilizadas para detectar clústeres de interés pertenecientes a microorganismos menos abundantes. Estrategias similares de prospección en las poblaciones microbianas, como la utilizada en el caso de la ET-743, demuestran que este tipo de identificaciones son aproximaciones válidas con una cantidad de secuencia relativamente similar a la utilizada en este estudio para cada una de las muestras (Rath *et al*., 2011). Por lo tanto el éxito de la identificación de la secuencia del clúster es extremadamente dependiente de la proporción del microorganismo productor en el microbioma y de la profundidad de la prospección metagenómica realizada.

De acuerdo con los resultados obtenidos, la suposición realizada al comienzo de esta Tesis por la cual el hecho de detectar un compuesto de interés implicaría que el microorganismo productor fuese mayoritario, ha resultado no ser cierta en los casos analizados. Excepto en el caso de *Polymastia*, no se han obtenido evidencias de la presencia de estos clústeres génicos. Entre otras causas, esto podría deberse a fenómenos de acumulación del compuesto o al gran nivel de sensibilidad de las herramientas de

detección utilizadas. Por lo tanto, se puede suponer que utilizando las herramientas empleadas en esta Tesis sería posible detectar *in silico* los clústeres de interés en el caso de que se pudiera acceder a esas poblaciones minoritarias, ya sea utilizando una mayor profundidad en la secuenciación masiva, o bien utilizando otras aproximaciones basadas por ejemplo en la secuenciación de célula única, o bien enriqueciendo poblaciones espécificas de bacterias.

VIII. Conclusiones

El trabajo descrito en esta Tesis Doctoral ha dado lugar a las siguientes conclusiones principales:

5. La cepa *T. mobilis* MES-10-09-028 puede ser modificada mediante herramientas de ingeniería genética lo que la convierte en un candidato excelente para ser utilizada como chasis para la producción de moléculas sintetizadas por PKSs y/o NRPSs.

6. Se ha demostrado que el clúster *ddn* es el responsable de la síntesis de didemninas y se ha expresado heterólogamente en *E. coli*.

7. Algunos microbiomas de las esponjas marinas analizadas contienen unos pocos simbiontes muy mayoritarios, lo que permite obtener sus genomas completos utilizando profundidades de secuenciación masiva relativamente bajas.

8. El análisis de las secuencias de clústeres génicos que codifican PKSs y/o NRPSs presentes en los metagenomas de esponjas demuestra que las moléculas antitumorales detectadas en las mismas no siempre son producidas por un simbionte mayoritario.

IX. Bibliografía

Abe, T. et al. Construction of a Metagenomic Library for the Marine Sponge *Halichondria okadai*. *Bioscience, Biotechnology, and Biochemistry* **76**, 633-639 (2012).

Ahmadian, A., Ehn, M. & Hober, S. Pyrosequencing: History, biochemistry and future. *Clinica Chimica Acta* **363**, 83-94 (2006).

Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology* **31**, 533-8 (2013).

Austin, M.B. & Noel, J.P. The chalcone synthase superfamily of type III polyketide synthases. *Natural Product Reports* **20**, 79-110 (2003).

Ayuso-Sacido, A. & Genilloud, O. New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: Detection and distribution of these biosynthetic gene sequences in major taxonomic groups. *Microbial Ecology* **49**, 10-24 (2005).

Aziz, R.K. et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

Baker, P.W., Kennedy, J., Dobson, A.D.W. & Marchesi, J.R. Phylogenetic diversity and antimicrobial activities of fungi associated with haliclona simulans isolated from irish coastal waters. *Marine Biotechnology* **11**, 540-547 (2009).

Baltz, R.H. Molecular engineering approaches to peptide, polyketide and other antibiotics. *Nature biotechnology* **24**, 1533-1540 (2006).

Baltz, R.H., Brian, P., Miao, V. & Wrigley, S.K. Combinatorial biosynthesis of lipopeptide antibiotics in *Streptomyces roseosporus*. *Journal of Industrial Microbiology and Biotechnology* **33**, 66-74 (2006).

Baltz, R.H., Miao, V. & Wrigley, S.K. Natural products to drugs: daptomycin and related lipopeptide antibiotics. *Natural product reports* **22**, 717-741 (2005).

Bavestrello, G. et al. Parasitic diatoms inside Antarctic sponges. *Biological Bulletin* **198**, 29-33 (2000).

Beld, J., Sonnenschein, E.C., Vickery, C.R., Noel, J.P. & Burkart, M.D. The phosphopantetheinyl transferases: catalysis of a post-translational modification crucial for life. *Natural Product Reports* **31**, 61-108 (2014).

Bernhardt, R. Cytochromes P450 as versatile biocatalysts. *Journal of Biotechnology* **124**, 128-45 (2006).

Bernt, M. et al. MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution* **69**, 313-319 (2013).

Bian, X. et al. Direct cloning, genetic engineering, and heterologous expression of the syringolin biosynthetic gene cluster in *E. coli* through Red/ET recombineering. *ChemBioChem* **13**, 1946-1952 (2012).

Blin, K. et al. antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Research* **41**, (2013).

Blunt, J.W., Copp, B.R., Munro, M.H.G., Northcote, P.T. & Prinsep, M.R. Marine natural products. *Natural Product Reports* **28**, 196-268 (2011).

Bode, H.B. & Müller, R. The impact of bacterial genomics on natural product research. *Angewandte Chemie - International Edition* **44**, 6828-6846 (2005).

Brady, S.F. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nature Protocols* **2**, 1297-1305 (2007).

Breitling, R., Achcar, F. & Takano, E. Modeling challenges in the synthetic biology of secondary metabolism. *ACS Synthetic Biology* **2**, 373-378 (2013).

Cárdenas, A., Rodríguez-R, L.M., Pizarro, V., Cadavid, L.F. & Arévalo-Ferro, C. Shifts in bacterial communities of two caribbean reef-building coral species affected by white plague disease. *The ISME Journal* **6**, 502-512 (2012).

Cerrano, C. et al. Are diatoms a food source for Antarctic sponges? Chemistry and Ecology 20, 57-64 (2004).

Challis, G.L., Ravel, J. & Townsend, C.A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry and Biology* **7**, 211-224 (2000).

Chevreux, B., Wetter, T. & Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)* 45-56. (1999).

Cole Stevens, D., Henry, M.R., Murphy, K.A. & Boddy, C.N. Heterologous expression of the oxytetracycline biosynthetic pathway in *Myxococcus xanthus*. *Applied and Environmental Microbiology* **76**, 2681-2683 (2010).

Cole, J.R. et al. The Ribosomal Database Project (RDP-II): Sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research* **33**, D294-D296 (2005).

Dendy, A. "1922-Report on the Sigmatotetraxonida collected by HMS" Sealark" in the Indian Ocean." *Trans. Linn. Soc. London, Zoology*. (1922).

Diaz, M.C. & Ward, B.B. Sponge-mediated nitrification in tropical benthic communities. *Marine Ecology Progress Series* **156**, 97-107 (1997).

Dobson, A.D.W., et al. "Marine Sponges–Molecular Biology and Biotechnology." *Hb25_Springer Handbook of Marine Biotechnology*. Springer Berlin Heidelberg, 219-254 (2015).

Dodsworth, J.A. et al. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nature Communications* **4**, 1854 (2013).

Du, L. & Lou, L. PKS and NRPS release mechanisms. *Natural Product Reports* **27**, 255-278 (2010).

Du, Y.L., Dalisay, D.S., Andersen, R.J. & Ryan, K.S. N-carbamoylation of 2,4-diaminobutyrate reroutes the outcome in padanamide biosynthesis. *Chemistry and Biology* **20**, 1002-1011 (2013).

Dubois, D. et al. ClbP is a prototype of a peptidase subgroup involved in biosynthesis of nonribosomal peptides. *Journal of Biological Chemistry* **286**, 35562-35570 (2011).

Dupont, C.L. et al. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *The ISME Journal* **6**, 1186-1199 (2012).

Eddy, S.R. Accelerated profile HMM searches. *PLoS Computational Biology* **7**, (2011).

Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-7 (2004).

Eisen, J.A. Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes. *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches* 157-162 (2011).

Erpenbeck, D., Voigt, O., Wörheide, G. & Lavrov, D.V. The mitochondrial genomes of sponges provide evidence for multiple invasions by Repetitive Hairpin-forming Elements (RHE). *BMC Genomics* **10**, 591 (2009).

Eustáquio, A.S. et al. Production of 8′-halogenated and 8′-unsubstituted novobiocin derivatives in genetically engineered Streptomyces coelicolor strains. *Chemistry and Biology* **11**, 1561-1572 (2004).

Fieseler, L. et al. Widespread occurrence and genomic context of unusually small polyketide synthase genes in microbial consortia associated with marine sponges. *Applied and Environmental Microbiology* **73**, 2144-2155 (2007).

Finking, R. & Marahiel, M.A. Biosynthesis of nonribosomal peptides. *Annual Review of Microbiology* **58**, 453-488 (2004).

Finn, R.D. et al. The Pfam protein families database. *Nucleic Acids Research* **38**, D211-D222 (2010).

Finn, R.D., Clements, J. & Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29-37 (2011).

Firn, R.D. & Jones, C.G. Natural products--a simple model to explain chemical diversity. *Natural Product Reports* **20**, 382-391 (2003).

Firn, R.D. & Jones, C.G. The evolution of secondary metabolism - A unifying model. *Molecular Microbiology* **37**, 989-994 (2000).

Fischbach, M.A. & Walsh, C.T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic machinery, and mechanisms. *Chemical Reviews* **106**, 3468-3496 (2006).

Flugel, R.S., Hwangbo, Y., Lambalot, R.H., Cronan, J.E. & Walsh, C.T. Holo-(acyl carrier protein) synthase and phosphopantetheinyl transfer in *Escherichia coli*. *Journal of Biological Chemistry* **275**, 959-968 (2000).

Frost, T. et al. A yellow-green algal symbiont in the freshwater sponge, *Corvomeyenia everetti*: convergent evolution of symbiotic associations. *Freshwater Biology* **38**, 395-399 (1997).

Fu, J. et al. Efficient transfer of two large secondary metabolite pathway gene clusters into heterologous hosts by transposition. *Nucleic Acids Research* **36**, (2008).

Fujii, I., Watanabe, A., Sankawa, U. & Ebizuka, Y. Identification of Claisen cyclase domain in fungal polyketide synthase WA, a naphthopyrone synthase of *Aspergillus nidulans*. *Chemistry and Biology* **8**, 189-197 (2001).

Gaino, E., Bavestrello, G., Cattaneo-Vietti, R. & Sara, M. Scanning electron microscope evidence for diatom uptake by two Antarctic sponges. *Polar Biology* **14**, (1994).

Gaitatzis, N., Hans, A., Müller, R. & Beyer, S. The *mtaA* gene of the myxothiazol biosynthetic gene cluster from *Stigmatella aurantiaca* DW4/3-1 encodes a phosphopantetheinyl transferase that activates polyketide synthases and polypeptide synthetases. *Journal of Biochemistry* **129**, 119-124 (2001).

Gao, Z.M. et al. Symbiotic adaptation drives genome streamlining of the cyanobacterial sponge symbiont "Candidatus *Synechococcus pongiarum*".*mBio* **5**, (2014).

Garson, M.J., Flowers, A.E., Webb, R.I., Charan, R.D. & McCaffrey, E.J. A sponge/dinoflagellate association in the haplosclerid sponge Haliclona sp.: Cellular origin of cytotoxic alkaloids by Percoll density gradient fractionation. *Cell and Tissue Research* **293**, 365-373 (1998).

Gauthier, M.E.A., Du Pasquier, L. & Degnan, B.M. The genome of the sponge *Amphimedon queenslandica* provides new perspectives into the origin of Toll-like and interleukin 1 receptor pathways. *Evolution and Development* **12**, 519-533 (2010).

Glass, E.M. & Meyer, F. The Metagenomics RAST Server: A public resource for the automatic phylogenetic and functional analysis of metagenomes. *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches* 325-331 (2011).

Gregor, I., Dröge, J., Schirmer, M., Quince, C. & McHardy, A.C. PhyloPythiaS+: A self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *arxiv.org* 1-67 (2014).

Hahn, M. & Stachelhaus, T. Harnessing the potential of communication-mediating domains for the biocombinatorial synthesis of nonribosomal peptides. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 275-280 (2006).

Hahn, M. & Stachelhaus, T. Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15585-15590 (2004).

Hall, N. Advanced sequencing technologies and their wider impact in microbiology. *The Journal of experimental biology* **210**, 1518-1525 (2007).

Hallam, S.J. et al. Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 18296-18301 (2006).

Harrington, C. et al. Evidence of bacteriophage-mediated horizontal transfer of bacterial 16S rRNA genes in the viral metagenome of the marine sponge Hymeniacidon perlevis. *Microbiology (United Kingdom)* **158**, 2789-2795 (2012).

Hentschel, U., Piel, J., Degnan, S.M. & Taylor, M.W. Genomic insights into the marine sponge microbiome. *Nature Reviews Microbiology* **10**, 641-654 (2012).

Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S.C. Integrative analysis of environmental sequences using MEGAN4. *Genome Research* **21**, 1552-1560 (2011).

Hyatt, D., Locascio, P.F., Hauser, L.J. & Uberbacher, E.C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223-2230 (2012).

Irigoien, X., Huisman, J. & Harris, R.P. Global biodiversity patterns of marine phytoplankton and zooplankton. *Nature* **429**, 863-867 (2004).

Juan, C. et al. Construction of a metagenomic DNA library of sponge symbionts and screening of antibacterial metabolites. *Journal of Ocean University of China* **5**, 119-122 (2006).

Kalisky, T., Blainey, P. & Quake, S.R. Genomic analysis at the aingle-cell level. *Annual Review of Genetics* **45**, 431-445 (2011).

Kennedy, J. et al. Isolation and analysis of bacteria with antimicrobial activities from the marine sponge *Haliclona simulans* collected from Irish waters. *Marine Biotechnology* **11**, 384-396 (2009).

Kennedy, J. et al. Marine metagenomics: New tools for the study and exploitation of marine microbial metabolism. *Marine Drugs* **8**, 608-628 (2010).

Kevany, B.M., Rasko, D.A. & Thomas, M.G. Characterization of the complete zwittermicin A biosynthesis gene cluster from *Bacillus cereus*. *Applied and Environmental Microbiology* **75**, 1144-1155 (2009).

Kim, B.K. et al. Genome sequence of an ammonia-oxidizing soil archaeon, "Candidatus *Nitrosoarchaeum koreensis*" MY1. *Journal of Bacteriology* **193**, 5539-5540 (2011).

Kim, S.K & Jayachandran V. "Introduction to Marine Biotechnology." *Hb25_Springer Handbook of Marine Biotechnology*. Springer Berlin Heidelberg, 1-10 (2015).

Kim, S.K & Senthilkumar K. "Introduction to Anticancer Drugs from Marine Origin." *Handbook of Anticancer Drugs from Marine Origin*. Springer International Publishing, 1-13 (2015).

Kim, U.J., Shizuya, H., Jong, P.J. de, Birren, B. & Simon, M.I. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Research* **20**, 1083-1085 (1992).

Kumar, P., Li, Q., Cane, D.E. & Khosla, C. Intermodular communication in modular polyketide synthases: Structural and mutational analysis of linker mediated protein - protein recognition. *Journal of the American Chemical Society* **125**, 4097-4102 (2003).

Lambalot, R.H. et al. A new enzyme superfamily - the phosphopantetheinyl transferases. *Chemistry & Biology* **3**, 923-936 (1996).

Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359 (2012).

Lavrov, D.V., Forget, L., Kelly, M. & Lang, B.F. Mitochondrial genomes of two demosponges provide insights into an early stage of animal evolution. *Molecular Biology and Evolution* **22**, 1231-9 (2005).

Lavrov, D.V., Wang, X. & Kelly, M. Reconstructing ordinal relationships in the Demospongiae using mitochondrial genomic data. *Molecular Phylogenetics and Evolution* **49**, 111-124 (2008).

Lee, J., Currano, J.N., Carroll, P.J. & Joullié, M.M. Didemnins, tamandarins and related natural products. *Natural product reports* **29**, 404 (2012).

Lee, O.O. et al. Pyrosequencing reveals highly diverse and species-specific microbial communities in sponges from the Red Sea. *The ISME journal* 5, 650-664 (2011).

Lee, Y., Lee, J. & Lee, H. Microbial symbiosis in marine sponges. *Journal of Microbiology-Seoul-* **39**, 254-264 (2001).

Li, J. & Neubauer, P. *Escherichia coli* as a cell factory for heterologous production of nonribosomal peptides and polyketides. *New Biotechnology* (2014).

Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

Li, Y. et al. Five new amicoumacins isolated from a marine-derived Bacterium *Bacillus subtilis*. *Marine Drugs* **10**, 319-328 (2012).

Lohr, J.E., Chen, F. & Hill, R.T. Genomic analysis of bacteriophage ΦJL001: Insights into its interaction with a sponge-associated alpha-Proteobacterium. *Applied And Environmental Microbiology* **71**, 1598-1609 (2005).

Maloof, A.C. et al. Possible animal-body fossils in pre-Marinoan limestones from South Australia. *Nature Geoscience* **3**, 653-659 (2010).

Marahiel, M.A. & Essen, L.O. Chapter 13 Nonribosomal Peptide Synthetases. Mechanistic and Structural Aspects of Essential Domains. *Methods in Enzymology* **458**, 337-351 (2009).

Margot, H., Acebal, C., Toril, E., Amils, R. & Fernández Puentes, J.L. Consistent association of crenarchaeal Archaea with sponges of the genus *Axinella*. *Marine Biology* **140**, 739-745 (2002).

Martin, J.L. & McMillan, F.M. SAM (dependent) I AM: The S-adenosylmethionine-dependent methyltransferase fold. *Current Opinion in Structural Biology* **12**, 783-793 (2002).

Martín M.J. *et al*. Antitumoral dihydropyran-2-one compounds. Patent: US 8,324,406 B2. (2012).

Martín, M.J. et al. Isolation and first total synthesis of PM050489 and PM060184, two new marine anticancer compounds. *Journal of the American Chemical Society* **135**, 10164-10171 (2013).

McAlpine, J.B. et al. Microbial genomics as a guide to drug discovery and structural elucidation: ECO-02301, a novel antifungal agent, as an example. *Journal of Natural Products* **68**, 493-496 (2005).

Medema, M. H., & Fischbach, M. A. (2015). Computational approaches to natural product discovery. *Nature chemical Biology*, **11(9)**, 639-648.

Medema, M.H. et al. AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research* **39**, 339-346 (2011).

Metsä-Ketelä, M. et al. An efficient approach for screening minimal PKS genes from Streptomyces. *FEMS Microbiology Letters* **180**, 1-6 (1999).

Meyer, F. et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).

Miller, D.J. et al. Crystal complexes of a predicted S-adenosylmethionine-dependent methyltransferase reveal a typical AdoMet binding domain and a substrate recognition domain. *Protein Science* **12**, 1432-1442 (2003).

Mosier, A.C., Allen, E.E., Kim, M., Ferriera, S. & Francis, C. a Genome sequence of "candidatus *nitrosoarchaeum limnia*" bg20, a low-salinity ammonia-oxidizing archaeon from the San Francisco bay estuary. *Journal of Bacteriology* **194**, 2119-2120 (2012).

Muscholl-Silberhorn, A., Thiel, V. & Imhoff, J.F. Abundance and bioactivity of cultured sponge-associated bacteria from the Mediterranean Sea. *Microbial Ecology* **55**, 94-106 (2008).

Nakano, M.M., Marahiel, M.A. & Zuber, P. Identification of a genetic locus required for biosynthesis of the lipopeptide antibiotic surfactin in *Bacillus subtilis*. *Journal of Bacteriology* **170**, 5662-5668 (1988).

Nielsen, H.B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* **32**, 822–828 (2014).

Nybakken, J.W. Marine Biology: An Ecological Approach. *Ecology* **Fourth ed**, 579 (2005).

Okamura, Y. et al. Isolation and characterization of a GDSL esterase from the metagenome of a marine sponge-associated bacteria. *Marine Biotechnology* **12**, 395-402 (2010).

Olano, C. et al. Biosynthesis of the angiogenesis inhibitor borrelidin by *Streptomyces parvulus* Tü4055: Insights into nitrile formation. *Molecular Microbiology* **52**, 1745-1756 (2004).

Olano, C. et al. Deciphering Biosynthesis of the RNA polymerase inhibitor Ssreptolydigin and generation of glycosylated derivatives. *Chemistry and Biology* **16**, 1031-1044 (2009).

Olano, C. Hutchinson's legacy: keeping on polyketide biosynthesis. *The Journal of Antibiotics* **64**, 51-57 (2011).

Olano, C., Méndez, C. & Salas, J.A. Post-PKS tailoring steps in natural product-producing actinomycetes from the perspective of combinatorial biosynthesis. *Natural Product Reports* **27**, 571-616 (2010).

On, O.L., Pui, Y.C., Yue, H.W., Pawlik, J.R. & Qian, P.Y. Evidence for vertical transmission of bacterial symbionts from adult to embryo in the Caribbean sponge *Svenzea zeai*. *Applied and Environmental Microbiology* **75**, 6147-6156 (2009).

Ongley, S.E. et al. High-titer heterologous production in *E. coli* of lyngbyatoxin, a protein kinase C activator from an uncultured marine cyanobacterium. *ACS Chemical Biology* **8**, 1888-1893 (2013a).

Ongley, S.E., Bian, X., Neilan, B. a & Müller, R. Recent advances in the heterologous expression of microbial natural product biosynthetic pathways. *Natural product reports* **30**, 1121-38 (2013b).

Overbeek, R. et al. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Research* **42**, D206–D214 (2014).

Pape, T. et al. Dense populations of Archaea associated with the demosponge Tentorium semisuberites Schmidt, 1870 from Arctic deep-waters. *Polar Biology* **29**, 662-667 (2006).

Penn, J. et al. Heterologous production of daptomycin in Streptomyces lividans. *Journal of Industrial Microbiology and Biotechnology* **33**, 121-128 (2006).

Piel, J. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14002-14007 (2002).

Piel, J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural Product Reports* **27**, 996-1047 (2010).

Piel, J. et al. Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 16222-16227 (2004).

Polacco, M.L. & Cronan, J.E. A mutant of *Escherichia coli* conditionally defective in the synthesis of holo-[acyl carrier protein]. *Journal of Biological Chemistry* **256**, 5750-5754 (1981).

Power, P. et al. Engineered synthesis of 7-oxo- and 15-deoxy-15-oxo-amphotericins: Insights into structure-activity relationships in polyene antibiotics. *Chemistry and Biology* **15**, 78-86 (2008).

Preston, C.M., Wu, K.Y., Molinski, T.F. & DeLong, E.F. A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 6241-6246 (1996).

Quast, C. et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590–D596 (2013).

Radax, R., Hoffmann, F., Rapp, H.T., Leininger, S. & Schleper, C. Ammonia-oxidizing archaea as main drivers of nitrification in cold-water sponges. *Environmental Microbiology* **14**, 909-923 (2012).

Rath, C.M. et al. Meta-omic characterization of the marine invertebrate microbial consortium that produces the chemotherapeutic natural product ET-743. *ACS Chemical Biology* **6**, 1244-1256 (2011).

Reid, R. et al. A model of structure and catalysis for ketoreductase domains in modular polyketide synthases. *Biochemistry* **42**, 72-79 (2003).

Reimer, D. & Bode, H.B. A natural prodrug activation mechanism in the biosynthesis of nonribosomal peptides. *Natural Product Reports* **31**, 154-9 (2014).

Reimer, D., Pos, K.M., Thines, M., Grün, P. & Bode, H.B. A natural prodrug activation mechanism in nonribosomal peptide synthesis. *Nature Chemical Biology* **7**, 888-890 (2011).

Richardson, C., Hill, M., Marks, C., Runyen-Janecky, L. & Hill, A. Experimental manipulation of sponge/bacterial symbiont community composition with antibiotics: Sponge cell aggregates as a unique tool to study animal/microorganism symbiosis. *FEMS Microbiology Ecology* **81**, 407-418 (2012).

Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 19126-19131 (2009).

Rinehart, K. L., et al. Structures of the didemnins, antiviral and cytotoxic depsipeptides from a Caribbean tunicate. *Journal of the American Chemical Society* **103.7**: 1857-1859 (1981).

Rinehart, K.L. and Lithgow-Bertelloni, A.M. PTC Int. Pat., WO 9104985 A1 (1991).

Rodriguez, E., Menzella, H.G. & Gramajo, H. Chapter 15 Heterologous Production of Polyketides in Bacteria. *Methods in Enzymology* **459**, 339-365 (2009).

Romero, F. et al. Thiocoraline, a new depsipeptide with antitumor activity produced by a marine Micromonospora. I. Taxonomy, fermentation, isolation, and biological activities. *The Journal of Antibiotics* **50**, 734-737 (1997).

Ross, A.C. et al. Biosynthetic multitasking facilitates thalassospiramide structural diversity in marine bacteria. *Journal of the American Chemical Society* **135**, 1155-1162 (2013).

Sambrook, J. & W Russell, D. Molecular Cloning: A Laboratory Manual. *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY* 999 (2001)

Scalera-Liaci, L., Sciscioli, M., Lepore, E. & Gaino, E. Symbiotic zooxanthellae in *Cinachyra tarentina*, a non-boring demosponge. *Endocytobiosis and Cell Research* **13**, 105-114 (1999).

Schäfer, A. et al. Small mobilizable multi-purpose cloning vectors derived from the Escherichia coli plasmids pK18 and pK19: Selection of defined deletions in the chromosome of *Corynebacterium glutamicum*. *Gene* **145**, 69-73 (1994).

Schirmer, A. et al. Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta*. *Applied and Environmental Microbiology* **71**, 4840-4849 (2005).

Schleissner, C., Rodríguez, P., García, J., de la Calle, F. & Cuevas, C. Didemnins production by two different strains of *Tistrella mobilis* isolated from a polychaete and a sponge at PharmaMar. 14th International Symposium on Marine Natural Products / 8th European Conference on Marine Natural Products (Comunicación) (2013).

Schmitt, S., Angermeier, H., Schiller, R., Lindquist, N. & Hentschel, U. Molecular microbial diversity survey of sponge reproductive stages and mechanistic insights into vertical transmission of microbial symbionts. *Applied and Environmental Microbiology* **74**, 7694-7708 (2008).

Scholz-Schroeder, B.K., Soule, J.D. & Gross, D.C. The *sypA*, *sypS*, and *sypC* synthetase genes encode twenty-two modules involved in the nonribosomal peptide synthesis of syringopeptin by *Pseudomonas syringae* pv. syringae B301D. *Molecular Plant-Microbe Interactions* **16**, 271-80 (2003).

Schwarzer, D., Mootz, H.D., Linne, U. & Marahiel, M.A. Regeneration of misprimed nonribosomal peptide synthetases by type II thioesterases. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14083-14088 (2002).

Segata, N. et al. Computational meta'omics for microbial community studies. *Molecular Systems Biology* **9**, 666 (2013).

Selvin, J., Kennedy, J., Lejon, D.P.H., Kiran, S. & Dobson, A.D.W. Isolation identification and biochemical characterization of a novel halo-tolerant lipase from the metagenome of the marine sponge *Haliclona simulans*. *Microbial Cell Factories* **11**, 72 (2012).

Shen, B. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology* **7**, 285-295 (2003).

Shi, B.-H., Arunpairojana, V., Palakawong, S. & Yokota, A. *Tistrella mobilis* gen nov, sp nov, a novel polyhydroxyalkanoate-producing bacterium belonging to alpha-Proteobacteria. *The Journal of General and Applied Microbiology* **48**, 335-343 (2002).

Shizuya, H. & Kouros-Mehr, H. The development and applications of the bacterial artificial chromosome cloning system. *The Keio Journal of Medicine* **50**, 26-30 (2001).

Siegl, A. et al. Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. *The ISME journal* **5**, 61-70 (2011).

Silva-Rocha, R. et al. The Standard European Vector Architecture (SEVA): A coherent platform for the analysis and deployment of complex prokaryotic phenotypes. *Nucleic Acids Research* **41**, (2013).

Simister, R.L., Deines, P., Botté, E.S., Webster, N.S. & Taylor, M.W. Sponge-specific clusters revisited: A comprehensive phylogeny of sponge-associated microorganisms. *Environmental Microbiology* **14**, 517-524 (2012).

Sipkema, D. et al. Multiple approaches to enhance the cultivability of bacteria associated with the marine sponge *Haliclona* (gellius) sp. *Applied and Environmental Microbiology* **77**, 2130-2140 (2011).

Stachelhaus, T., Mootz, H.D. & Marahiel, M.A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry and Biology* **6**, 493-505 (1999).

Steele, J.H. A comparison of terrestrial and marine ecological systems. *Nature* **313**, 355-358 (1985).

Steger, D. et al. Diversity and mode of transmission of ammonia-oxidizing archaea in marine sponges. *Environmental Microbiology* **10**, 1087-94 (2008).

Stephens, J. XV.—Atlantic Sponges collected by the Scottish National Antarctic Expedition. *Transactions of the Royal Society of Edinburgh* **50.02**: 423-467 (1915).

Sunbul, M., Zhang, K. & Yin, J. Chapter 10 Using Phosphopantetheinyl Transferases for Enzyme Posttranslational Activation, Site Specific Protein Labeling and Identification of Natural Product Biosynthetic Gene Clusters from Bacterial Genomes. *Methods in Enzymology* **458**, 255-275 (2009).

Tae, K.K. & Fuerst, J. a Diversity of polyketide synthase genes from bacteria associated with the marine sponge *Pseudoceratina clavata*: Culture-dependent and culture-independent approaches. *Environmental Microbiology* **8**, 1460-1470 (2006).

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30**, 2725-2729 (2013).

Tang, G.L., Cheng, Y.Q. & Shen, B. Leinamycin biosynthesis revealing unprecedented architectural complexity for a hybrid polyketide synthase and nonribosomal peptide synthetase. *Chemistry and Biology* **11**, 33-45 (2004).

Taylor, M.W., Radax, R., Steger, D. & Wagner, M. Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. *Microbiology and Molecular Biology reviews* **71**, 295-347 (2007).

Thomas, T. et al. Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *The ISME journal* **4**, 1557-1567 (2010).

Vacelet, J. & Donadey, C. Electron microscope study of the association between some sponges and bacteria. *Journal of Experimental Marine Biology and Ecology* **30**, 301-314 (1977).

Van Soest, R.W.M., et al. *World Porifera database*. (2015)

Walker, C.B. et al. *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proceedings of the National Academy of Sciences* **107**, 8818-8823 (2010).

Walsh, C.T., Gehring, A.M., Weinreb, P.H., Quadri, L.E. & Flugel, R.S. Post-translational modification of polyketide and nonribosomal peptide synthases. *Current Opinion in Chemical Biology* **1**, 309-315 (1997).

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**, 5261-7 (2007).

Watanabe, K. et al. Total biosynthesis of antitumor nonribosomal peptides in *Escherichia coli*. *Nature Chemical Biology* **2**, 423-428 (2006).

Weber, T. et al. Molecular analysis of the kirromycin biosynthetic gene cluster revealed βalanine as precursor of the pyridone moiety. *Chemistry and Biology* **15**, 175-188 (2008).

Webster, N.S., Negri, A.P., Munro, M.M.H.G. & Battershill, C.N. Diverse microbial communities inhabit Antarctic sponges. *Environmental Microbiology* **6**, 288-300 (2004).

Webster, N.S. et al. Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. *Environmental Microbiology* **12**, 2070-2082 (2010).

Webster, N.S. & Taylor, M.W. Marine sponges and their microbial symbionts: Love and other relationships. *Environmental Microbiology* **14**, 335-346 (2012).

Wehrl, M., Steinert, M. & Hentschel, U. Bacterial uptake by the marine sponge *Aplysina aerophoba*. Microbial Ecology 53, 355-365 (2007).

Wesener, S.R., Potharla, V.Y. & Cheng, Y.Q. Reconstitution of the FK228 biosynthetic pathway reveals cross talk between modular polyketide synthases and fatty acid synthase. *Applied and Environmental Microbiology* **77**, 1501-1507 (2011).

Wiens, M. et al. Toll-like receptors are part of the innate immune defense system of sponges (Demospongiae: Porifera). *Molecular Biology and Evolution* **24**, 792-804 (2007).

Wirth, R., Friesenegger, A. & Fiedler, S. Transformation of various species of gram-negative bacteria belonging to 11 different genera by electroporation. *Molecular and General Genetics* **216**, 175-177 (1989).

Wu, K., Chung, L., Revill, W.P., Katz, L. & Reeves, C.D. The FK520 gene cluster of *Streptomyces hygroscopicus* var. *ascomyceticus* (ATCC 14891) contains genes for biosynthesis of unusual polyketide extender units. *Gene* **251**, 81-90 (2000).

Xi, L., Ruan, J. & Huang, Y. Diversity and biosynthetic potential of culturable actinomycetes associated with marine sponges in the China seas. *International Journal of Molecular Sciences* **13**, 5917-5932 (2012).

Xu, Y. et al. Bacterial biosynthesis and maturation of the didemnin anti-cancer agents. *Journal of the American Chemical Society* **134**, 8625-8632 (2012).

Yeh, E., Kohli, R.M., Bruner, S.D. & Walsh, C.T. Type II thioesterase restores activity of a NRPS module stalled with an aminoacyl-S-enzyme that cannot be elongated. *ChemBioChem* **5**, 1290-1293 (2004).

Zhang, H., Wang, Y., Wu, J., Skalina, K. & Pfeifer, B.A. Complete biosynthesis of erythromycin A and designed analogs using *E. coli* as a heterologous host. *Chemistry and Biology* **17**, 1232-1240 (2010).