Universitat
de les Illes Balears

TESIS DOCTORAL
2017

IMPLICACIONES ECOLÓGICAS Y CLÍNICAS DEL ANÁLISIS COMPARATIVO DE GENOMAS DE MICOBACTERIAS AMBIENTALES Y PATÓGENAS

Daniel Jaén Luchoro

Universitat de les Illes Balears

TESIS DOCTORAL
2017

Programa de Doctorado de Microbiología Ambiental y Biomédica

IMPLICACIONES ECOLÓGICAS Y CLÍNICAS DEL ANÁLISIS COMPARATIVO DE GENOMAS DE MICOBACTERIAS AMBIENTALES Y PATÓGENAS

Daniel Jaén Luchoro

Director: Antoni Bennasar Figueras
Tutor: Antoni Bennasar Figueras

Doctor por la Universitat de les Illes Balears

Universitat de les Illes Balears

Dr. Antoni Bennasar Figueras, de Universidad de las Islas Baleares

DECLARO:

Que la tesis doctoral que lleva por título *Implicaciones ecológicas y clínicas del análisis comparativo de genomas de micobacterias ambientales y patógenas,* presentada por Daniel Jaén Luchoro para la obtención del título de doctor, ha sido dirigida bajo mi supervisión.

Y para que quede constancia de ello firmo este documento.

Firma

Palma de Mallorca, 9-10-2017

A mis padres y mi hermana

Agradecimientos

No resulta nada fácil llegar al final de una etapa, mirar atrás y resumir en pocas palabras lo que has vivido a lo largo de unos cuantos años. Parece que fue ayer cuando empecé mi aventura en la Universidad, lleno de ilusión y con ganas de intentar aventurarme en el mundo de la investigación. No tenía muy claro en aquel momento en que campo me quería involucrar, ni siquiera si podría conseguirlo; lo único que sabía es que tenía ganas de intentarlo. Y casi sin darme cuenta aquí estoy, doce años después, sentado frente al portátil escribiendo las que probablemente sean las últimas palabras que me permitan cerrar este importante capítulo de mi vida.

Han sido unos años intensos de alegrías y decepciones, donde el esfuerzo y la dedicación me han permitido avanzar y aprender sin parar. Pero debo cerrar la puerta a esta etapa y este hecho despierta en mi un sentimiento de profunda satisfacción por haber llegado hasta el final, pero a la vez de tristeza teñida de nostalgia, ya que son años que se van para no volver. En este tiempo he vivido un sinfín de experiencias profesionales y personales y he atesorado recuerdos (algunos buenos y otros no tanto) que se convertirán, sin lugar a dudas, en las típicas historietas que se contarán en las cenas para echarse unas risas. Con todo lo expuesto, y sin extenderme más, me gustaría destacar que nada de esto habría sido posible sin todas y cada una de las personas que me han acompañado hasta el día de hoy. Por este motivo me gustaría dedicar unas palabras a ellas (¡aunque espero no olvidarme de nadie!).

En primer lugar, me gustaría dar las gracias a mi director de tesis, el Doctor Toni Bennasar Figueras por la oportunidad que me brindaste en 2011 de empezar mi andadura por el laboratorio de microbiología de la UIB en el marco de un TFM, y por el voto de confianza al apostar por mí para hacer el doctorado bajo tu tutela. Muchas gracias por tus consejos, tus enseñanzas y ánimos ya que debido a todo esto mi formación ha dado un salto qualitativo enorme. Lo único que espero de corazón es haber cumplido tus expectativas y que esto no sea un punto y final.

También quiero dar las gracias al resto de los “capos” del laboratorio (Rafa, Balbina, Elena y Jorge), porque siempre han estado dispuestos a ayudarme en la medida de lo posible con cualquier duda o problema que he tenido.

A Marga, porque has estado ahí cuando te he necesitado, recibiéndome siempre con esa sonrisa que te caracteriza. He aprendido muchas cosas de ti y lo único que espero es poder seguir aprendiendo de tu experiencia y compartiendo todo tipo de historias contigo a pesar de que nuestros caminos profesionales se separen. Ojalá pudiera agradecértelo lo suficiente (¡Y no me llames pelota que te conozco! ¡Que lo digo muy en serio!).

A Cristina, Magda, Bel Brunet, Cati, María, David, Claudia y Toni (¡Busquets!), personas a las que considero mi "familia del laboratorio". Son impagables las risas de todos estos años a vuestro lado. Conseguíais que, por muy mal día que tuviera, siempre fuera con muchas ganas al laboratorio ya que el simple hecho de saber que os iba a encontrar allí me hacía sonreir. No me puedo quejar de la enorme suerte que he tenido de encontraros a todos juntos y haber compartido tantos días con vosotros. Cabe destacar también todo lo que me habéis ayudado siempre que lo he necesitado y sólo espero haberos ayudado a vosotros también en algún sentido. Sabed que, por mucho tiempo que pase, aquel "tío callado del fondo que de repente empezó a hablar y a soltar burradas" os guarda un sitio especial en su corazón.

A mi profesora Margalida Llabrés, a la que considero la mejor docente que he tenido nunca. Gracias por estimular mi interés por el mundo de la biología a través de la pasión e ilusión que demostrabas en cada una de tus clases. Un pedacito de este éxito también es tuyo.

Al resto de integrantes de "los cuatro fantásticos" (Toni, Xavi y Juan), porque estar a vuestro lado durante la carrera ha sido un no parar de reír. Sois los responsables de convertir esos años en la mejor etapa como estudiante de toda mi vida. Siempre nos quedaran frases célebres como "El niño de la pezuña", "El muelle", "Salmón sí, siempre salmón" y esas tardes de prácticas en las que no podíamos estar "quietecitos". En este sentido tampoco puedo olvidarme de Bel Galmès (siempre serás Bel-1 para mi), Bel Brunet (¡Otra vez!), Maria del Mar o Marga Ramis, porque también habéis sido parte importante en mi experiencia universitaria. Gracias de corazón por todo lo vivido (y lo que, espero, nos quede por vivir).

A Francisco Salvà-Serra, por todo lo que me ayudaste en mis "primeros pasos" en el mundo de la bioinformática y por abrirme la puerta profesional a Suecia para vivir una

gran experiencia junto a ti y todo el equipo de investigación de Gotemburgo (¡En especial junto a Hedvig y nuestra "celebrity" Roger!).

A Francisco Aliaga, porque la mejor experiencia de hacer el doctorado ha sido, sin ningún tipo de duda, el haberte conocido. Eres un modelo para mí de compromiso y trabajo duro del que he aprendido grandes cosas tanto profesional como personalmente. Gracias por todas y cada una de las experiencias a tu lado (en especial esas tardes de laboratorio donde se nos hacía media noche, pero las risas seguían siendo el protagonista principal a pesar del cansancio). Ojalá algún día podamos volver trabajar codo con codo para volver a "hacer de las nuestras" en el laboratorio, aunque solo sea una vez. Y si no es posible, por lo menos siempre nos quedaran nuestras comidas/cenas para seguir disfrutando (¡Pero no te olvides del jamón!).

A mi gran amigo Jose Enrique (Kike mola más), porque desde que te conocí hace más de quince años te convertiste en insustituible para mí. Tu apoyo incondicional durante tantos años ha sido fundamental en mi desarrollo personal y tus ánimos en momentos difíciles han sido esenciales para seguir adelante sin importar las dificultades. Uno de los grandes motivos por los que he llegado a este momento eres tú, con tus "no te rindas", "sigue intentándolo" o "Vamos a hacer ruta y así te despejas". Te puedo decir, sin miedo a equivocarme, que me harían falta cien vidas para agradecerte lo suficiente todo lo que has hecho por mi(¡incluso así me quedaría corto!).

A Conchi, Elena, Sai, Pedro (El representante de la ley), Isa, Guillermo y Toni, porque no hay mejor regalo en esta vida que mirar a tu alrededor y verte acompañado de grandes amigos como vosotros. Gracias, porque, aunque no os deis cuenta (o no lo diga lo suficiente), sois parte importante de mi persona y sin vosotros todo esto habría sido mucho más difícil.

A mi novia Cristina, porque, a pesar de haber aparecido en mi vida en las etapas finales de este camino, has estado al pie del cañón día tras día para empujarme siempre adelante. Lo único que siento es que hayas tenido que vivir "la peor etapa", donde el estrés, los retrasos y los bajones de ánimo han sido el fruto de casi cada día. Pero ahí has estado, escuchándome atenta y pacientemente, sin importar que me tirara horas soltando todo lo

que tenía dentro. Gracias por tu apoyo, tus palabras y la alegría que me brindas cada día, porque sin estos tres regalos este último tramo habría sido muy diferente.

Y, por último, a las personas más importantes de mi vida: mis padres y mi hermana. Porque gracias a vosotros no he estado solo en ningún momento de mi vida. Porque si estoy aquí ahora mismo es porque vosotros lo habéis hecho posible. Porque os habéis destrozado las manos y la espalda para que yo tuviera todas las oportunidades que vosotros no tuvisteis y evitar que tuviera que dedicarme a "uno de esos trabajos rompe-espaldas". He sido consciente cada día de todos y cada uno de vuestros esfuerzos y he intentado siempre que ninguno de ellos fuera en vano. Prueba de ello es que estoy aquí, dedicándoos probablemente las palabras que nunca os he dicho pero que siempre he sentido. Gracias por todo, porque os debo a vosotros todo lo que soy a día de hoy.

ABACAS: del inglés *Algorithm Based Automatic Contiguation of Assembled Sequences*

ADN: Ácido desoxiribonucleico

ADNr: Ácido desoxiribonucleico ribosómico

ARN: Ácido ribonucleico

ARNm: Ácido ribonucleico mensajero

ATP: Adenosin trifosfato

BDBH: del inglés *Bi-Directional Best Hits*

BLASR: del inglés *Basic Local Alignment with Successive Refinement*

BLAST: del inglés *Basic Local Alignment Search Tool*

BLASTP: del inglés *Basic Local Alignment Search Tool for Proteins*

CA: del inglés *Celera assembler* (Ensamblador Celera)

CARD: del inglés *Comprehensive Antibiotic Resistance Database*

CDS: del inglés *Coding DNA Sequence*

CMA: Complejo *Mycobacterium avium*

COG: del inglés *Cluster of Orthologous Genes*

COGT: del inglés *Cluster of Orthologous Genes Triangles*

dNTP: Desoxinucletótido trifosfato

ddNTP: Didesoxinucleótido trifosfato

DO: Densidad Óptica

ECF: del inglés *Extra Cromosomal Function*

EDTA: Ácido etilendiaminotetraacético

FRC: del inglés *Feature Reponse Curve*

FS: Factor Sigma

FT: Factor de Transcripción

GC: Guanina-Citosina

HGAP: del inglés *Hierarchical Genome Assembly Process*

HMM: del inglés *Hidden Markov Model*

HP: del inglés *Hypothetical Protein* (proteína hipotética)

HQ: Histidina Quinasa

IPTG: Isopropil β-D-1-tiogalactopiranósido

KEGG: del inglés *Kyoto Encyclopedia of Genes and Genomes*

LAM: LipoArabinoManano

LR: del inglés *Long Read* (lectura larga)

MBL: Metalo-β-Lactamasa

MCL: Micobacterias de Crecimiento Lento

MCR: Micobacterias de crecimiento Rápido

MCS: del inglés Multi Cloning Site

MLE: del inglés *Maximum Llikelihood Estimation*

MLSA: del inglés *Multilocus Sequence Analisis*

MNT: Micobacterias No Tuberculosas

MP: del inglés *Mate-Pair*

NCBI: del inglés *National Center for Biotechnology Information*

OMCL: del inglés *Orthologous Markov CLuster algorithm*

PacBio: del inglés *Pacific Bioscience*

PAGIT: del inglés *Post-Assembly Genome-Improvement Toolkit*

PCR: del inglés *Polymerase Chain Reaction*

PDB: del inglés *Protein Data Bank*

PE: del ingles *Paired-End*.

PGAP: del inglés *Prokaryotic Genome Annotation Pipeline*

PHAST: del inglés *PHAgue Search Tool*

PIM: Fosfatidilinositol Manósido

PQS: PoliQuétido Sintasa

PSK:del inglés *Post-Segregational Killing*

QS: *Quorum Sensing*

QUAST: del inglés *QUality Assessment Tool*

REAPR: del inglés *Recognition of Errors in Assemblies using Paired Reads*

RGI: del inglés *Resistance Gene Identifier*

RR: Regulador de Respuesta

RT: Regulador Transcripcional

SDC: Sistema de Dos Componentes

SIDA: Sindrome de InmunoDeficiencia Adquirida

SMRT: del inglés *Single Molecule Real-Time*

SNG: Secuenciación de Nueva Generación

SSPACE: del inglés *Scaffolding Pre-Assemblies After Contig Extension*

STA: Sistema Toxina-Antitoxina

STRING: del inglés *Search Tool for the Retrieval of Interacting Genes/proteins*

TA: Toxina-Antitoxina

TAC: Toxina-Antitoxina-Chaperona

TADB: del inglés *Toxin-Antitoxin DataBase*

UP: del inglés *Uncharacterized Protein* (Proteína no caracterizada)

UFC: Unidades Formadoras de Colonias

UNAG: UDP-N-AcetilGlucosamina

VIH: Virus de Inmunodeficiencia Humana

Listado de Figuras

Introducción general

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Listado de Tablas

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Listado del Anexo 1

Tabla suplementaria 1. Genomas utilizados para el estudio de genoma esencial y pangenoma del grupo de micobacterias de crecimiento rápido. Se incluyen los números de acceso para cada uno, incluyendo los correspondientes a los plásmidos en los casos en los que están presentes.

Tabla suplementaria 2. Genomas utilizados para el estudio de genoma esencial y pangenoma del grupo de *abscessus-chelonae-immunogenum*. Se incluyen los números de acceso para cada uno, incluyendo los correspondientes a los plásmidos en los casos en los que estan presentes. Los genomas correspondientes a *M. immunogenum* incluidos en esta tabla son los utilizados para el mismo estudio centrado en esta especie.

Tabla suplementaria 3. Genomas utilizados para el estudio de genoma esencial y pangenoma de la especie *M. tuberculosis*. Se incluyen los números de acceso para cada uno.

Tabla suplementaria 4. Listado de los códigos de las categorías funcionales de los COG relacionados con las funciones específicas que representan en cada caso.

Listado del Anexo 2

Tabla suplementaria 1. Listado completo de todos los FT encontrados en las cepas de MCR estudiadas, incluyendo las listadas en la tabla reducida el capítulo 3.

Tabla suplementaria 2. Listado completo de todos los FT encontrados en las cepas cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv., incluyendo las listadas en la tabla reducida el capítulo 3.

Tabla suplementaria 3. Familias de reguladores de respuesta encontrados en los genomas de las cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv. Se indica el número de representantes encontrados en cada caso.

Tabla suplementaria 4. Factores sigma identificados en el análisis del reguloma de los genomas de las cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv. Se indica el número de representantes encontrados en cada caso.

Tabla suplementaria 5. Reguladores negativos (Factores antisigma) identificados en la prospección de los proteomas de las cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv. Se indica la presencia (+) o ausencia (-) de cada factor en los distintos genomas.

Tabla suplementaria 6. Número de elementos relacionados con SDC encontrados en los genomas de las cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv. Se indica el número de histidina quinasas, reguladores de respuesta y el número de SDC completos formados entre ellos. Se indica también el número de elementos para los cuales no se ha hallado el elemento relacionado que completaría el SDC.

Listado del Anexo 3

Tabla suplementaria 1. Cebadores diseñados para la amplificación de los elementos de los STA. En rojo se muestran las dianas de restricción pertinentes en cada caso. Previa a esta diana se insertaron 6 nucleótidos coincidentes con la secuencia original del genoma para incrementar la eficiencia de hibridación en la PCR inicial.

Tabla sumplementaria 2. Porcentajes de identidad entre las secuencias de las proteínas MT0933, la lipasa y MT0934, utilizando la secuencia del genoma de la cepa tipo de *M. chelonae* CCUG 47445^{T} como referencia.

Tabla suplementaria 3. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapB27 de *M. llatzerense* MG13^{T}. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Tabla suplementaria 4. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapC27 de *M. llatzerense* MG13^{T}. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Tabla suplementaria 5. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapB28 de *M. llatzerense* MG13^{T}. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Tabla suplementaria 6. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapC28 de *M. llatzerense* MG13^{T}. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Tabla suplementaria 7. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la proteína hipotética de la cepa *Mycobacterium* MHSD3. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Tabla suplementaria 8. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la toxina zeta de la cepa *Mycobacterium* MHSD3. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Tabla suplementaria 9. Aminoácidos del centro activo de las proteínas con dominio PIN utilizadas. Asp: aspartato, Glut: glutamina, Gly: glicina, Asn: asparagina.

Figura suplementaria 1. Patrón de bandas proteicas obtenidos para los sistemas VapBC27 y VapBC28 (A) y el sistema PH y Toxina zeta (B). Se destacan en rojo los pesos moleculares (en kDa) referentes a la zona donde se enmarcan las proteínas de interés. Se destacan también las bandas potencialmente representativas de las toxinas VapC28 y zeta, así como la proteína hipotética relacionada con esta última**.**

Figura suplementaria 2. Alineamiento obtenido a partir de la superposición de las estructuras terciarias. Entre paréntesis se indican el número de aminoácidos que no se superponen entre los distintos bloques alineados. Se resaltan aquellos aminoácidos conservados del centro activo en amarillo, y en rojo los aminoácidos que corresponden a la cuarta posición del mismo, no conservados en las toxinas VapC28 y VapC5.

Figura suplementaria 3. Predicción estructural realizada por I-TASSER a partir de la secuencia de aminoácidos de la hipotética toxina MT0934 del par MT0933-34 de *M. llatzerense* MG13^{T}.

Resumen/Resum/Summary

Resumen

El objetivo principal de la presente tesis es realizar un estudio detallado desde el punto de vista genómico de todas aquellas características que pueden beneficiar el carácter patogènico de las micobacterias de crecimiento rápido (MCR), en especial de las especies estrechamente relacionadas *Mycobacterium abscessus, Mycobacterium chelonae* y *Mycobacterium immunogenum.* Dichas especies son consideradas como importantes patógenos oportunistas, responsables de infecciones difíciles de tratar por sus características intrínsecas y que complican el estado de salud de los pacientes hospitalizados. Además, con este estudio se pretende compensar el desequilibrio histórico existente en cuanto al estudio del género *Mycobacterium,* donde los esfuerzos se han centrado fundamentalmente en especies patógenas como *Mycobacterium tuberculosis*. Incrementar la información disponible de las MCR mencionadas permitirá futuros estudios centrados en la investigación de los elementos de patogenicidad que poseen, así como proponer nuevas dianas para el desarrollo de tratamientos alternativos de las infecciones que provocan.

Con el fin de abordar el problema del escaso número de genomas presentes en las bases de datos de las especies mencionadas; se desarrolló un protocolo de extracción de ADN eficaz para micobacterias, consideradas como microorganismos difíciles de lisar. Obtenidas las muestras de ADN, se utilizaron tecnologías de secuenciación de nueva generación para la obtención de grandes cantidades de secuencias (denominadas lecturas) a partir de las cuales se obtuvieron los genomas de las respectivas cepas. Con este fin se utilizaron las herramientas bioinformáticas más óptimas de las disponibles para el ensamblaje, mejora y evaluación de los genomas obtenidos. De esa forma se aseguró la obtención de genomas de alta calidad con las herramientas disponibles en el momento.

Los genomas obtenidos, junto con los disponibles en las bases de datos, se utilizaron para el estudio comparativo basado en el cálculo del genoma esencial (que incluye el conjunto de proteínas compartidas por todos los genomas) y pangenoma (conjunto de todos los grupos de proteínas diferentes presentes en un conjunto de genomas). El genoma esencial permitió clarificar las relaciones evolutivas de las diferentes especies consideradas, especialmente en el complejo de subespecies de *M. abscessus*. El estudio del pangenoma

permitió sugerir una alta capacidad adaptativa de las MCR debido a la tendencia abierta detectada en su pangenoma. El mismo resultado se obtuvo en el grupo *abscessus-chelonae-immunogenum*, pero no en la especie *M. immunogenum*, donde la tendencia cerrada del pangenoma refleja la escasa plasticidad genómica de la especie.

Un hito importante conseguido fue la caracterizaron funcional de las proteínas exclusivas de una determinada especie o genoma. En este sentido, se realizó un catálogo de todos aquellos elementos genómicos implicados en la patogenicidad de cada microorganismo. Para ello se utilizaron toda una serie de bases de datos especializadas para la búsqueda de genes de resistencias a antibióticos, factores de virulencia, elementos móviles, proteínas reguladoras y elementos implicados en la percepción del Quórum. De esta forma se determinó un extenso catálogo de elementos genéticos que pueden influir de forma decisiva en la patogenicidad de cada cepa analizada.

Finalmente, se caracterizaron los llamados sistemas toxina-antitoxina (STA), cuya funcionalidad puede influir en la patogenicidad de los microorganismos. En este caso no solo se utilizaron toda una serie de bases de datos especializadas y herramientas bioinformáticas para la descripción de estos elementos, sino que también se estableció un protocolo para realizar el ensayo experimental de su funcionalidad, utilizando a *Escherichia coli* como hospedador y vectores de expresión donde se clonaron los genes de la toxina y la antitoxina por separado. De esta forma se consiguió la descripción completa desde el punto de vista genómico, estructural y funcional de todos los sistemas detectados en las micobacterias objeto de estudio, observando resultados compatibles con un STA en tres ellos.

Resum

L'objectiu principal de la present tesis es realizar un estudi detallat desde el punt de vista genòmic de totes aquelles caracteristiques que poden bneficiar el caràcter patogènic de les micobacteris de creixement ràpid (MCR), en especial de les espècies estretament relacionades *Mycobacterium abscessus, Mycobacterium chelonae* y *Mycobacterium immunogenum*. Les espècies esmentades es consideren importants patògens oportunistes, responsables d'infeccions difícils de tractar degut a les seves característiques intrínseques i que compliquen l'estat de salut dels pacients hospitalitzats. A més a més, amb aquest estudi es pretén compensar el desequilibri històric existent en quant a l'estudi del gènere *Mycobacterium*, on l'esforç s'ha centrat fonamentalment en espècies patògenes com *Mycobacterium tuberculosis*. Incrementar la informació disponible de les MCR esmentades permetrá futurs estudis centrats en la investigació dels elements de patogenicitat que posseeixen, així com proposar noves dianes per al desenvolupament de tractaments alternatius de les infeccions que provoquen.

Amb la finalitat d'abordar el problema de l'escàs nombre de genomes presents en els bases de dades de les espècies destacades, es va desenvolupar un protocol d'extracció d'ADN eficaç per micobacteris, considerades com a microorganismes difícils de lisar. Un cop obtingudes les mostres d'ADN, s'empraren tecnologies de seqüenciació de nova generació per a l'obtenció de grans quantitats de seqüencies (denominades lectures) a partir de les quals es varen obtindre els genomes de les respectives soques. Amb aquest objectiu es varen utilitzar les eines bioinformàtiques més òptimes de les disponibles per a l'ensamblatje, millora y evaluació dels genomes obtinguts. D'aquesta manera es va asegurar l'obtenció de genomes d'alta qualitat amb les eines disponibles en el moment.

Els genomes obtinguts, juntament amb els disponibles en les bases de dades, es varen utilitzar per a l'estudi comparatiu basat en el càlcul del genoma esencial (que inclou el conjunt de proteïnes compartides per tots els genomes) y pangenoma (conjunt de tots els grups de proteïnes diferents presents en un conjunt de genomes). El genoma esencial va permetre aclarir les relacions evolutives de les diferents espècies considerades, especialment en el complexe de subespècies de *M. abscessus*. L'estudi del pangenoma va permetre sugerir una alta capacitat adaptativa de les MCR degut a la tendència oberta

detectada en el pangenoma. El mateix resultat es va obtenir en el grup *abscessus-chelonae-immunogenum*, pero no en la espècie *M. immunogenum*, en la que la tendència tancada del pangenoma reflecteix la baixa plasticitat genòmica de l'espècie.

Un fet important aconseguit va ser la caracterització funcional de les proteïnes exclusives d'una espècie determinada o genoma. En aquest sentit, es va realitzar un catàleg de tots aquells elements genòmics implicats en la patogenicitat de cada microorganisme. Per aconseguir-ho es varen utilitzar tota una sèrie de bases de dades especialitzades per a la recerca de gens de resistència a antibiòtics, factors de virulència, elements mòvils, proteïnes reguladores y elements implicats en la percepció del Quòrum. D'aquesta forma es va determinar un extens catàleg d'elements genètics que poden influir de manera decisiva en la pateogenicitat de cada soca analitzada.

Finalment, es caracteritzaren els coneguts com a sistemes toxina-antitoxina (STA), la funcionalitat dels cuals pot influir en la patogenicitat dels microorganismes. En aquest cas no tan sols es varen utilitzar tota una sèrie de bases de dades especialitzades y eines bioinformàtiques per a la descripció d'aquests elements, sinó que també es va establir un protocol per realitzar l'assaig experimental de la seva funcionalitat, emprant a *Escherichia coli* com hospedador y vectors d'expresió on es varen clonar els gens de la toxina i la antitoxina per separat. D'aquesta manera es va aconseguir la descripció completa des del punt de vista genòmic, estructural i funcional de tots els sistemas detectats en els micobacteris objecte d'estudi, observant resultats compatibles amb un STA en tres dels sistemes detectats.

Summary

The main objective of the present thesis is to perform a detailed study from a genomic point of view of all those characteristics that can benefit the pathogenicity of the rapid growing mycobacteria (RGM), especially of the closely related species *Mycobacterium abscessus*, *Mycobacterium chelonae* and *Mycobacterium immunogenum*. These species are considered important opportunistic pathogens, responsible of infections that are hard to treat due to their intrinsic characteristics. These infections complicate the health status of the hospitalized patients. In addition, this study pretends to compensate the historical imbalance in the study of the genus *Mycobacterium*, in which the efforts have been put in the study of pathogenic species as *Mycobacterium tuberculosis*. Increasing the information available of the RGM mentioned before in the databases will allow future studies to be focused on the research of the pathogenicity elements that they have, as well as propose new targets for the development of alternative treatments of the infections that they provoke.

With the objective to increase the actual low number of genomes present in the databases of the species mentioned before, it was developed a DNA extraction protocol effective for mycobacteria, considered as "hard-to-lyse" microorganism. Having obtained the DNA samples, next-generation sequencing technologies were used to obtain massive amounts of sequences (called reads) from which the genomes of the respective strains were obtained. With this purpose, the most suitable and available bioinformatic tools were used for the assembly, improvement and checking of the genomes obtained. Thus, it was ensured the obtention of high-quality genomes with the tools available at that moment.

The resulting genomes, along with the genomes available in the databases, were used for a comparative study based on the estimation of the core genome (which includes all the proteins shared by all the genomes) and the pangenome (set of all the groups of different proteins present in all the studied genomes). The core genome allowed to clarify the evolutive relationships among the species considered, especially inside the complex of subspecies of *M. abscessus*. The determination of the pangenome allowed to suggest a high adaptive capacity of the RGM group due to the open tendency observed in its pangenome. The same result was obtained in the group *abscessus-chelonae-*

immunogenum, but not in the species *M. immunogenum*, in which the closed tendency reflects the limited genomic plasticity of the species.

An important achievement was the characterization of the proteins exclusive of one specific species or genome. In this sense, it was obtained a detailed catalogue of all those genomic elements related with the pathogenicity of each microorganism using specialized databases to find genes related to antibiotic resistances, virulence factors, mobile genetic elements, regulatory proteins and elements related to the *Quorum sensing*. Consequently, it was determined a large catalogue of genetic elements that can influence decisively in the pathogenicity of each strain.

Finally, the elements called toxin-antitoxin systems (TAS) were characterized. The functionality of these elements can be closely related with the pathogenicity of the microorganisms. In this case, in addition to the use of specialized databases and bioinformatic tools for the description of these elements, it was developed a protocol for the experimental assay of their functionality. For this purpose, it was used *Escherichia coli* as host and expression vectors where the genes of the toxins and antitoxins were cloned separately. With this procedure, the TAS were completely described from the genomic, structural and functional point of view, obtaining results compatible with a TAS in three of them.

Publicaciones

Lista de publicaciones derivadas de la presente tesis

Del trabajo realizado en la presente tesis se han generado los siguientes artículos científicos:

- Daniel Jaén-Luchoro, Francisco Salvà-Serra, Francisco Aliaga-Lozano, Carolina Seguí, Antonio Busquets, Antonio Ramírez, Mikel Ruíz, Margarita Gomila, Jorge Lalucat, Antoni Bennasar-Figueras. 2016. **Complete genome sequence of *Mycobacterium chelonae* type strain CCUG 47445, a rapidly growing species of nontuberculous mycobacteria.** Genome Announc 4(3):e00550-16. doi:10.1128/genomeA.00550-16

- Daniel Jaén-Luchoro, Carolina Seguí, Francisco Aliaga-Lozano, Francisco Salvà-Serra, Antonio Busquets, Margarita Gomila, Antonio Ramírez, Mikel Ruiz, Edward Moore, Jorge Lalucat, Antoni Bennasar-Figueras. 2016. **Complete genome sequence of the *Mycobacterium immunogenum* type strain CCUG 47286.** Genome Announc 4(3):e00401-16. doi:10.1128/genomeA.00401-16

- Daniel Jaén-Luchoro, Francisco Aliaga-Lozano, Rosa Maria Gomila, Margarita Gomila, Francisco Salvà-Serra, Jorge Lalucat, Antoni Bennasar-Figueras. **First insights into a type II toxin-antitoxin system from the clinical isolate *Mycobacterium* sp. MHSD3, similar to epsilon/zeta systems.** PLoS ONE 12(12): e0189459. https://doi.org/10.1371/journal.pone.0189459

1. Introducción general

1.1. El Género *Mycobacterium*

1.1.1. Características generales de las micobacterias

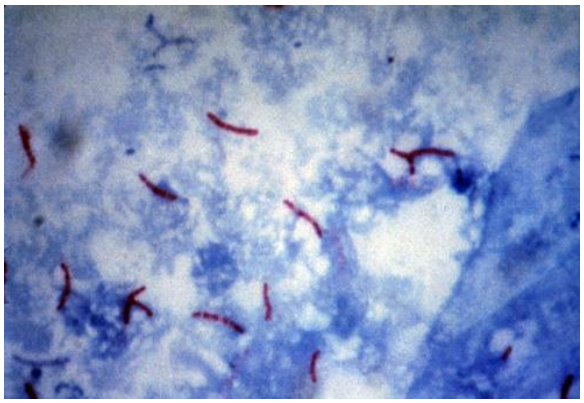


Figura 1.1. Bacilos de *Mycobacterium tuberculosis* tras la tinción Ziehl-Neelsen. Imagen de dominio público

El género *Mycobacterium* pertenece a la clase *Actinobacteria*, uno de los grandes grupos de bacterias Gram positivas, caracterizadas por ser bacterias con forma de bacilo o filamentosas, en su mayoría no infecciosas, a pesar de grandes excepciones. El género *Mycobacterium* consta en la actualidad de 178 especies y 13 subespecies [1] ampliamente distribuidas en multitud de nichos ecológicos diferentes. Se trata de microorganismos aerobios, no móviles y no formadores de esporas que, a pesar de ser considerados estructuralmente Gram positivos, no se tiñen a través de esta técnica, sino que se utiliza una técnica específica conocida como la tinción de bacilos ácido-alcohol resistentes (BAAR) o Ziehl-Neelsen (Figura 1.1). Sus requerimientos nutricionales son relativamente simples, pudiendo crecer en su mayoría en medios minerales con amonio y glicerol como únicas fuentes de nitrógeno y carbono respectivamente. Son bacterias con un elevado contenido GC, que en cultivo crecen formando colonias densas, compactas y a menudo rugosas [2].

En líneas generales se clasifican en dos grandes grupos: micobacterias de crecimiento lento, donde se concentran las micobacterias patógenas (MCL, crecimiento en más de 7 días), y las micobacterias de crecimiento rápido (MCR, crecimiento en menos de 7 días). En función de su capacidad de producir pigmentación es posible clasificarlas en cuatro grandes grupos: 1) fotocromogénicas (grupo de Runyon I), cuya pigmentación aparece al hacerlas crecer expuestas a la luz; 2) escotocromogénicas (grupo de Runyon II), las cuales crecen formando colonias intensamente amarillas o naranjas estando expuestas o no a la luz; 3) No cromogénicas, las cuales no producen pigmentación o producen colonias pálidamente amarillas, cuya intensidad de color no incrementa si se expone o no a la luz (grupos de Runyon III y IV) [2].

Una de las particularidades más destacadas de este género es la peculiar pared bacteriana que presentan sus integrantes (Figura 1.2). Esta se caracteriza por ser muy gruesa y rica en lípidos, siendo los ácidos micólicos un tipo de lípido muy característico de su estructura [3]. Estos lípidos se encuentran en la superficie bacteriana unidos covalentemente al peptidoglucano de la pared celular. Esta configuración de pared bacteriana hace que sea una cubierta cerosa, altamente consistente e hidrofóbica; lo que la hace una de las principales responsable de la gran resistencia a factores externos tales como desinfectantes, metales pesados, antibióticos, etc. [4]. De hecho, esta hidrofobicidad es la que hace que estos bacilos no puedan teñirse a través de la tinción de Gram.

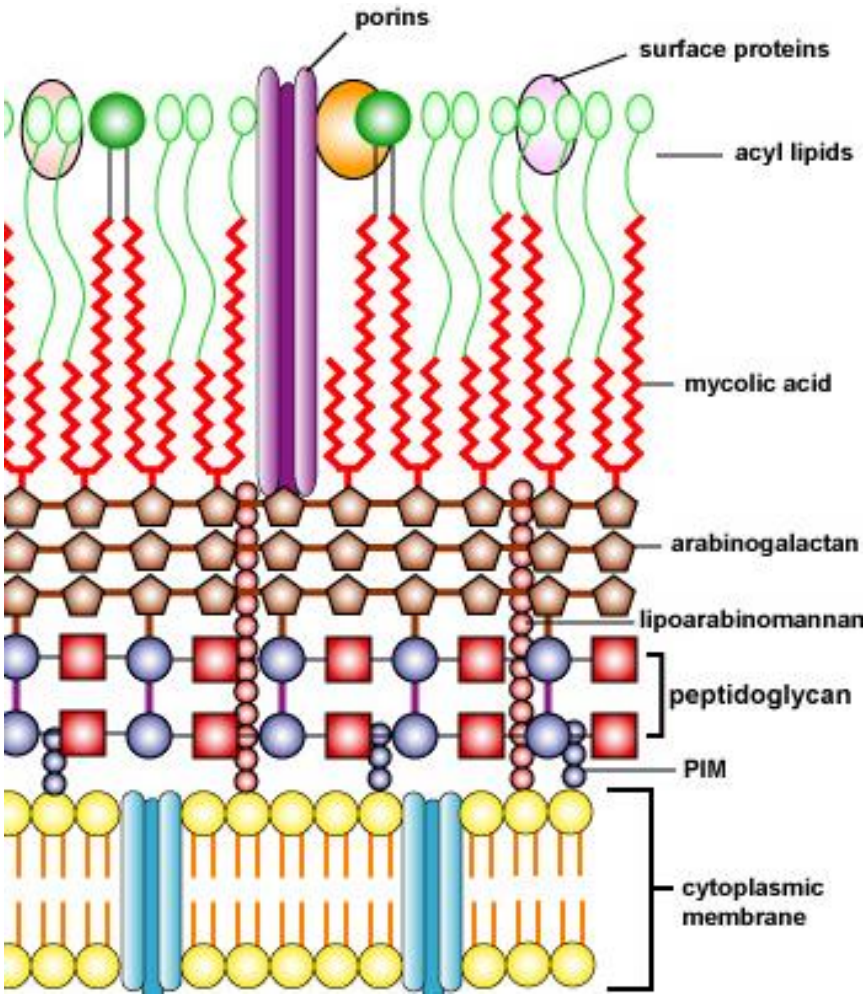


Figura 1.2. Esquema general de la pared micobacteriana. De arriba abajo, se destaca a) las proteínas de membrana (incluidas las porinas), b) acil-lípidos, c) ácidos micólicos, del arabinogalactano, e) lipoarabinomanano (LAM), e) peptidoglicano, g) fosfatidilinositol manosidos (PIM); y h) la membrana citoplasmática. Fuente de la imagen: *Community College of Baltimore County* (faculty.ccbcmd.edu).

1.1.2. Contextualización histórica

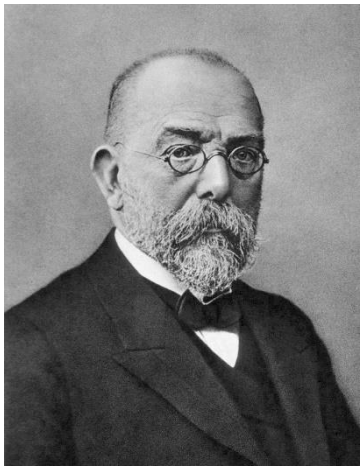


Figura 1.3 Born Heinrich Hermann Robert Koch (1843-1910). Fuente de la imagen https://alchetron.com/Robert-Koch

Históricamente, el gran peso del estudio de este género bacteriano ha recaído sobre las cepas oficialmente reconocidas como patógenas, como *Mycobacterium tuberculosis.* Una buena prueba de ello se encuentra en las bases de datos, donde la gran mayoría de la información sobre este género está basada en estudios realizados sobre dicha especie. El motivo hay que buscarlo en las graves enfermedades que provoca en el ser humano y que desde la antigüedad se sabe que han venido causando graves problemas de salud [5,6]. La tuberculosis es una enfermedad fundamentalmente pulmonar, caracterizada por la expectoración de esputo sanguinolento, tos, fiebre, sudores y pérdida de peso, aunque también puede extenderse a otras regiones del cuerpo; como son los sistemas nervioso, circulatorio, linfático o huesos, causando un cuadro sintomático diverso en cada caso [7]. La tuberculosis se calcula que ha acompañado al ser humano durante miles de años. Ya en tiempos de Hipócrates, en la antigua Grecia, se conocían perfectamente los efectos de esta enfermedad, entonces conocida como *phthisis*, la cual se consideraba altamente contagiosa. Son numerosas las epidemias que ha provocado a lo largo de la historia de la humanidad y que han sesgado un número incalculable de vidas hasta el momento [8], motivos que históricamente han dado lugar a la imperiosa necesidad de conocer cuál era la causa y las posibles soluciones. El agente causal de la tuberculosis fue descrito finalmente por el médico y microbiólogo alemán Heinrich Hermann Robert Koch (1843-1910) (Figura 1.3) el 24 de marzo de 1.882 en Berlín. Durante sus estudios consiguió aislar un agente microbiano que era capaz de reproducir la enfermedad en animales [2]. Gracias a su descubrimiento, numerosos investigadores comenzaron el camino para intentar entender los entresijos del proceso de infección provocado por este microorganismo, carrera que, a día de hoy, todavía no ha concluido. En la actualidad se calcula que aproximadamente un tercio de la población mundial está infectada por *M. tuberculosis,* y que este patógeno se sigue cobrando

alrededor de un millón de vidas anuales [9], convirtiéndose en el que probablemente sea el microorganismo que más muertes ha provocado con respecto al resto.

Con estos antecedentes históricos, es entendible la urgencia y el esfuerzo de estudio puesto sobre dicho microorganismo, pero esto ha provocado que se dejaran un poco de lado a las micobacterias llamadas "ambientales", también conocidas como micobacterias no tuberculosas (MNT) o micobacterias atípicas. Este último grupo de micobacterias se está convirtiendo en un grupo de gran interés clínico en los últimos años ya que algunos de sus representantes pueden desencadenar infecciones oportunistas de considerable gravedad. Dichas infecciones provocadas por micobacterias ambientales se conocen desde principios del siglo XX [5] y con el tiempo han ido cobrando más importancia debido a los problemas clínicos que generan y que complican la recuperación de los pacientes [10]. Además, así como en el caso de *M. tuberculosis* se realiza un control y reporte exhaustivo del número de infecciones detectadas, no ocurre lo mismo con los aislamientos de MNT en pacientes, por lo que la información epidemiológica en este último caso es difusa. Sin embargo, algunos centros han comenzado a contabilizar los casos de infecciones donde el agente causal es una MNT, observándose en algunos casos que el número de aislamientos de MNT esta empezando a sobrepasar al número de aislamientos del agente causal de la tuberculosis [11], por lo que estas micobacterias deben considerarse como una emergente e importante fuente de agentes potencialmente infecciosos.

1.1.3. Relevancia clínica de las micobacterias no tuberculosas

Las MNT con capacidad de desarrolar infecciones suelen aprovechar situaciones en las que el paciente ya tiene algún tipo de afección para sortear las defensas del organismo y desencadenar una infección [5], hecho por el que son consideradas patógenos oportunistas, es decir, microorganismos que ante una situacion normal no nos afectan, pero que ante determinadas circunstancias en las que existe alguna alteración de los sistemas de defensa del organismo, son capaces de desencadenar un proceso infectivo. Este particular hecho se ha agravado por el incesante incremento de pacientes inmunodeprimidos como, por ejemplo, los pacientes que han desarrollado el síndrome de

inmunodeficiencia adquirida (SIDA) a causa de la infección por el virus de inmunodeficiencia humana (VIH) [5].

Las infecciones provocadas por MNT, debido a las características de resistencia intrínseca a antibióticos de uso común, así como las propiedades inherentes a la propia pared celular de estos microorganismos, hacen que no sean sencillas de tratar y que generalmente sea de gran importancia conocer qué especie en concreto está causando el problema para aplicar un tratamiento específico y dirigido, combinando los fármacos más adecuados en cada caso [12,13]. Así, por ejemplo, MNT como *M. abscessus* son mas susceptibles a antibióticos como amikacina, cefoxitina, imipenem, claritromicina o azitromicina, mientras que *M. chelonae* responde mejor a macrólidos, linelozida y tobramicina. Este tipo de antibióticos suelen ser bastante agresivos, por lo que provocan efectos secundarios de considerable importancia en el paciente, tales como vómitos, perdida de audición, anorexia, insomnio, reacciones cutáneas e incluso problemas neuromusculares, entre muchos otros efectos adversos [11]. Si ante estos efectos secundarios se añade el hecho de que los tratamientos suelen ser prolongados en el tiempo, comprendiendo periodos que pueden extenderse hasta los 6 y 12 meses [11], la situación del paciente se complica en gran medida. En ocasiones se detectan infecciones recurrentes debido, según se apunta en algunos casos, a la aparición de células persistentes a través de la activación de diversos elementos genéticos. Estas células persitentes son capaces de quedar en estado latente hasta que el agente agresor externo (en este caso un antibiótico) desaparece, para volver después a activar toda su maquinaria metabólica y reproducir la infección [14], reiniciando el proceso de tratamiento.

Entre todas las especies incluídas en el grupo de las MNT, existen tres especies de MCR estrechamente relacionadas entre sí que se han posicionado como unas de las más problemáticas desde el punto de vista clínico: *M. chelonae*, el complejo *M. abscessus* (en el que se incluyen *M. abscessus*, *M. abscessus* subps. *bolletii* y *M. abscessus* subsp. *massiliense*) además de *M. immunogenum* [15]. Dichas especies suelen representar las micobacterias mas comunmente aisladas en infecciones nosocomiales, presentando una relativa alta resistencia a desinfectantes de uso común así como una alta resistencia intrínseca a una gran variedad de antibióticos [15–18].

En cuanto a sus manifestaciones clínicas, *M. chelonae* suele relacionarse con infecciones cutáneas, aparición de abscesos, celulitis localizadas, osteomelitis, infecciones producidas por catéteres contaminados e incluso infecciones diseminadas. Las cepas de *M. chelonae* han sido reconocidas como responsables de diversos brotes de infección relacionados con la industria del tatuaje y como un importante agente infeccioso relacionado con los transplantes, implantacion de prótesis e incluso procesos de hemodiálisis [19–24]. Aunque no es muy frecuente su presencia en infecciones pulmonares como agente causal, sí se han reportado casos en pacientes con graves afecciones en las vías respiratorias, tales como la fibrosis quística [25].

Por su parte, el complejo *M. abscesus* es un grupo taxonomicamente complejo que ha sufrido varias reclasificaciones en el transcurso de los años [26–28], tal como se puede observar en la Figura 1.4. En cualquier caso, son varias las manifestaciones clínicas que se pueden asociar a este complejo. Una de ellas son las infecciones de las vías respiratorias, especialmente peligrosas en pacientes con una enfermedad pulmonar de base, como la fibrosis quística, donde *M. abscessus* suele ser de entre las micobacterias la más frecuente. Dichas infecciones pueden comenzar de forma completamente asintomática e ir evolucionando lentamente en el paciente, comprometiendo cada vez más su calidad de vida; aunque también se han descrito avances repentinos en la infección con resultados fulminantes para el paciente, implicando, por ejemplo, fallos pulmonares fatales [29–31].

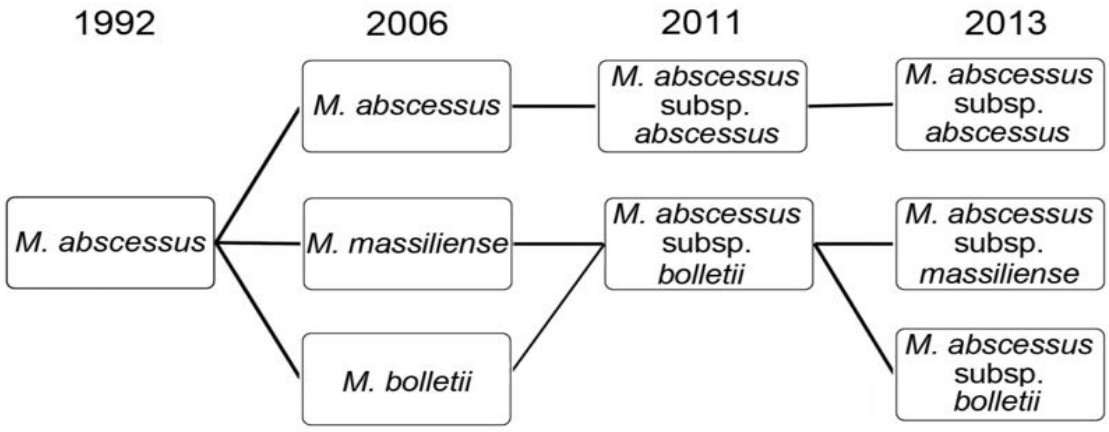


Figura 1.4. Esquema de las reclasificaciones taxonómicas sufridas por el complejo *M. abscessus* desde el año 1992 al año 2013. Fuente de la imagen: "*Mycobacterium abscessus Complex Infections in Humans*" (Emerging Infectious Diseases • www.cdc.gov/eid • Vol. 21, No. 9, September 2015).

Por su parte, *M. immunogenum* suele relacionarse con infecciones cutáneas, infeciones durante tratamientos de mesoterapia o pneumonitis hipersensible. Además, se ha relacionado con brotes o pseudobrotes debidos a lavados broncoalveolares, brotes de queratitis e incluso hay documentados casos de choque séptico debido a la diseminación de la propia infección [36–40].

Teniendo estas consideraciones en cuenta, es evidente que desde el punto de vista clínico es importante tener presentes estas especies, no sólo por las infeciones que pueden causar sino por la dificultad de su tratamiento, largo y con importantes efectos secundarios derivados. Esto convierte a estas MCR en patógenos oportunistas significativos que forman parte del grupo de las MNT.

1.1.4. Ecología y adaptabilidad

Como se ha indicado previamente, las MCR son ubicuas y pueden aislarse a partir de suelos, rocas e incluso agua, pudiendo incluso formar parte de bioaerosoles. Su presencia puede detectarse en ambientes altamente exigentes, como por ejemplo nichos ecológicos en los que se dan bajos valores de pH o altas temperaturas, entre otros [11]. Además, al ser poco exigentes nutricionalmente, pueden desarrollarse en ambientes oligotróficos, como son las aguas potables, y es precisamente este ambiente el que mejor define su relación con el ser humano. Efectivamente, los sistemas de aguas potables funcionan como un vehículo idóneo para la dispersión de estos microorganismos, permitiéndoles penetrar y distribuirse por todo tipo de dependencias utilizadas por el hombre. Evidentemente, a este tipo de instalaciones se les suelen aplicar sistemas de control, fundamentalmente basados en procesos de desinfección como mecanismo para el control del crecimiento y la circulación de microorganismos. Sin embargo, tal y como se ha determinado previamente, las MCR como *M. chelonae*, *M. abscessus* y *M. immunogenum* presentan una especial resistencia a los desinfectantes de uso común utilizados para este proceso, tales como el cloro, compuestos organomercuriales o glutaraldehído [41]. Ésto las convierte en bacterias difíciles de erradicar y con una facilidad relativa para superar las barreras defensivas y disponer de un acceso considerablemente fácil al ser humano.

Al ser patógenos oportunistas, no es un problema especialmente grave cuando infectan a personas con buenas condiciones de salud. No obstante, cuando este fenómeno ocurre en

hospitales, donde dichas especies son capaces de contaminar todo tipo de material, es cuando el contacto entre bacteria y ser humano se vuelve peligroso, ya que los pacientes hospitalizados son un grupo especialmente vulnerable. En numerosas ocasiones se han conseguido aislar este tipo de bacterias a partir de aguas hospitalarias, así como también de las redes que suministran aguas de uso doméstico [42,43]. Un caso extremo es el agua utilizada en los circuitos de hemodiálisis, purificada pero no estéril. Esta puede presentar una comunidad microbiana que debe ser controlada para evitar efectos adversos al preparar y suministrar el agua de hemodiálisis a los pacientes, y donde se ha visto que pueden proliferar micobacterias a pesar de todos los controles que se aplican sobre estas aguas [43]. Esto incrementa el riesgo para la salud de los pacientes que requieren este tipo de tratamientos, especialmente en pacientes inmunodeprimidos [44]. En el caso de un estudio realizado sobre este tipo de aguas en el Hospital Universitario de Son Dureta (Palma de Mallorca, Islas Baleares) se puso de manifiesto no sólo este hecho, si no que además la diversidad microbiana era relativamente elevada (Figura 1.5) [43].

Un último pero no menos importante aspecto a destacar hace referencia a la participación de estos microorganismos en la formación de biopelículas, estructuras organizadas de microorganismos embebidos en una matriz extracelular [2]. Este tipo de estructuras incrementa todavía más la resistencia de sus integrantes a factores externos, y la presencia de micobacterias en biopelículas tempranas las convierte en importantes integrantes pioneros en su formación [45–47]. La formación de estas estructuras en las tuberías de suministro de aguas o en otro tipo de estructuras en contacto con el agua (grifos, duchas, etc.) hace que sean incluso más difíciles de eliminar del ambiente hospitalario, convirtiéndose en una constante fuente de suministro de estos agentes potencialmente infecciosos. En base a este conjunto de motivos, no es extraño que los representantes del grupo de MCR aparezcan constantemente como responsables de numerosas infecciones nosocomiales.

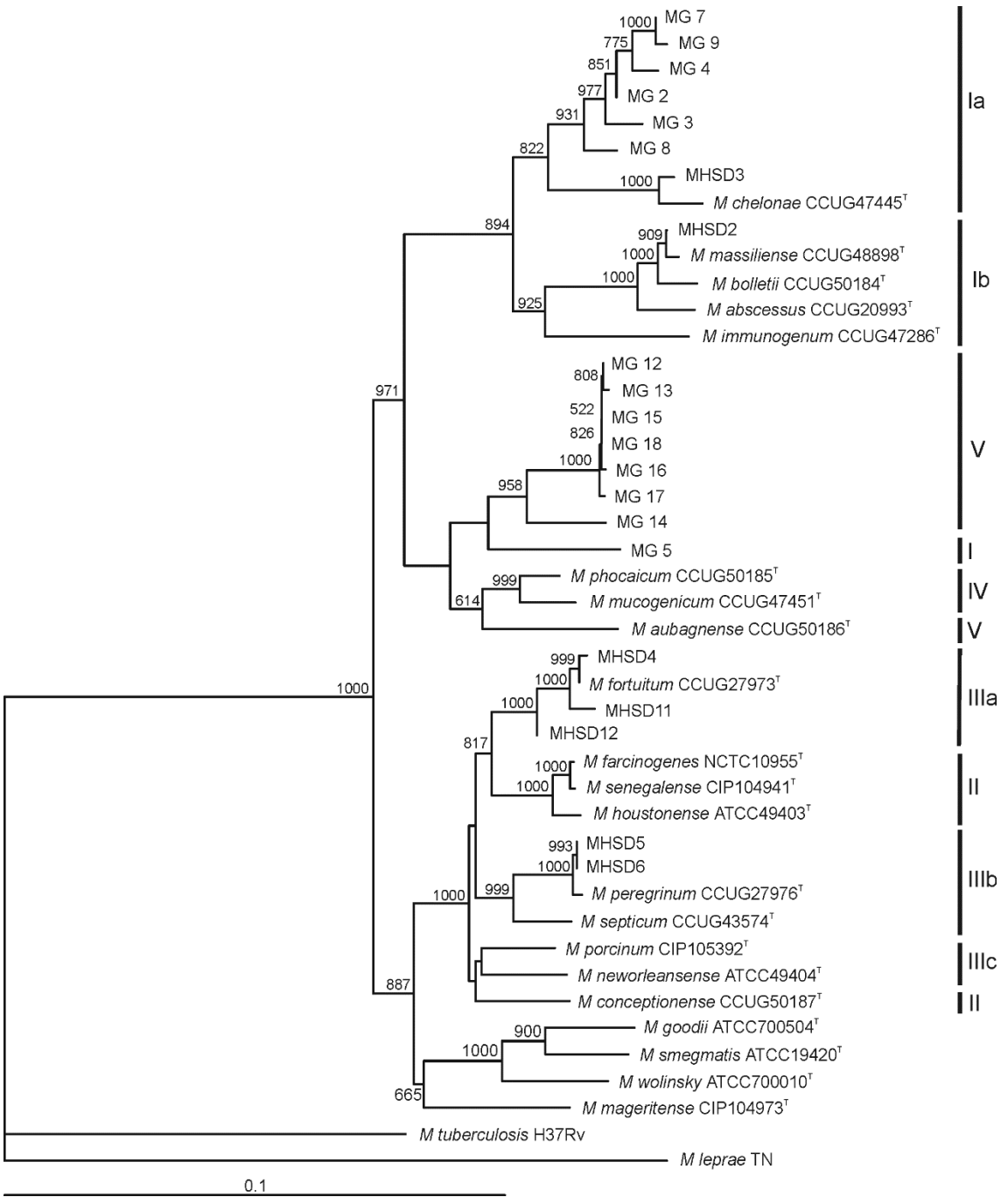


Figura 1.5. Distribución de las cepas de micobacterias aisladas a partir de muestras de agua de hemodiálisis. El árbol refleja las relaciones evolutivas basadas en el análisis de secuencia multilocus basada en los genes gyrB, hsp65, recA, rpoB, sodA, y ADNr 16S. Fuente de la Figura: "*Diversity of environmental Mycobacterium isolates from hemodialysis water as shown by a multigene sequencing approach*", Gomila y colaboradores, Applied Environmental Microbiology, 2007).

1.2. Origen y desarrollo de la secuenciación del ADN

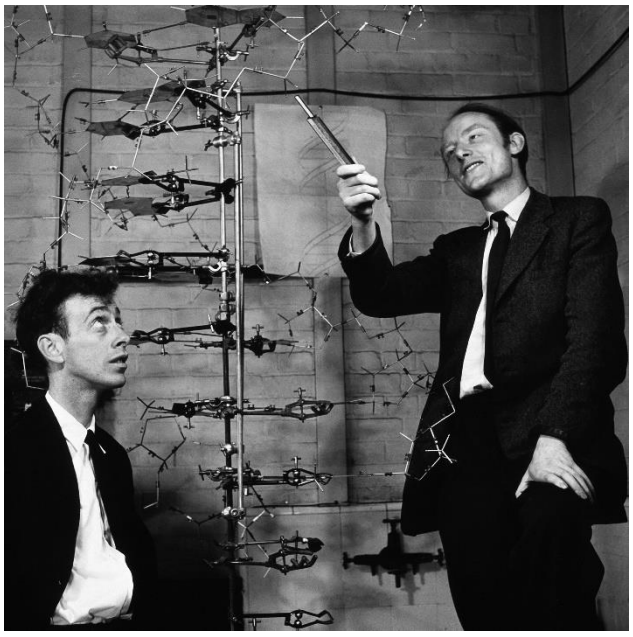


Figura 1.6. James Dewey Watson y Francis Crick frente a un modelo tridimensional de la estructura del ADN. Fuente de la imagen: https://abcienciade.wordpress.com.

Desde el descubrimiento del ácido desoxiribonucleico (ADN) por James Dewey Watson y Francis Crick en 1953, así como la posterior definición de dicha molécula como la responsable del almacenamiento de la información genética, ha existido un gran interés en el estudio de las características y funcionamiento de la misma. Al ser la molécula que contiene la información genética de un organismo, puede considerarse como el manual de instrucciones del mismo, el estudio del cual permitiría descifrar su funcionamiento y entender mejor su relación con el ambiente. Uno de los grandes pasos en este sentido fue el desarrollo de métodos que permitieran determinar su secuencia nucleotídica. Uno de los métodos más importantes de secuenciación fue diseñado por el bioquímico británico Frederick Sanger en 1.975 [49], el método de terminación de cadena. Dicho método se basa en la utilización de nucleótidos modificados químicamente, llamados didesoxinucleótidos (abreviados como ddNTP) y que son nucleótidos que carecen de un grupo 3'-hidroxilo (-OH) en la desoxirribosa que entraña la imposibilidad de formar un enlace fosfodiéster con los nuevos nucleótidos, y por tanto la adición de un nuevo dNTP a la cadena de nueva síntesis; durante la replicación de una molécula de ADN que tiene lugar durante su secuenciación por el método de Sanger, esta característica provoca que el proceso se detenga de forma específica ante la necesidad de incorporación de una adenina (A), guanina (G), citosina (C) o timina (T) (Figura 1.7A).

Tal fue el éxito de esta metodología que incluso a día de hoy se dispone de kits comerciales y sistemas automatizados de secuenciación de ADN basados en este método, habiéndose sustituido la radioactividad utilizada inicialmente como marcaje por ddNTPs marcados con fluorocromos (los cuales emiten fluorescencia de un color determinado

para cada nucleótido), permitiendo así realizar las cuatro reacciones de forma simultánea en un mismo tubo y, gracias a los avances metodológicos actuales, resolver los fragmentos por capilaridad [50]. Otro gran método de secuenciación fue desarrollado por Allan Maxam y Walter Gilbert en 1.977 [48] y está basado en la modificación química del ADN para permitir la ruptura del mismo en bases específicas (Figura 1.7B). Sin embargo, el método desarrollado por Friederick Sanger desbancó esta metodología por completo, extendiéndose en el tiempo hasta la actualidad tal y como se ha indicado previamente.

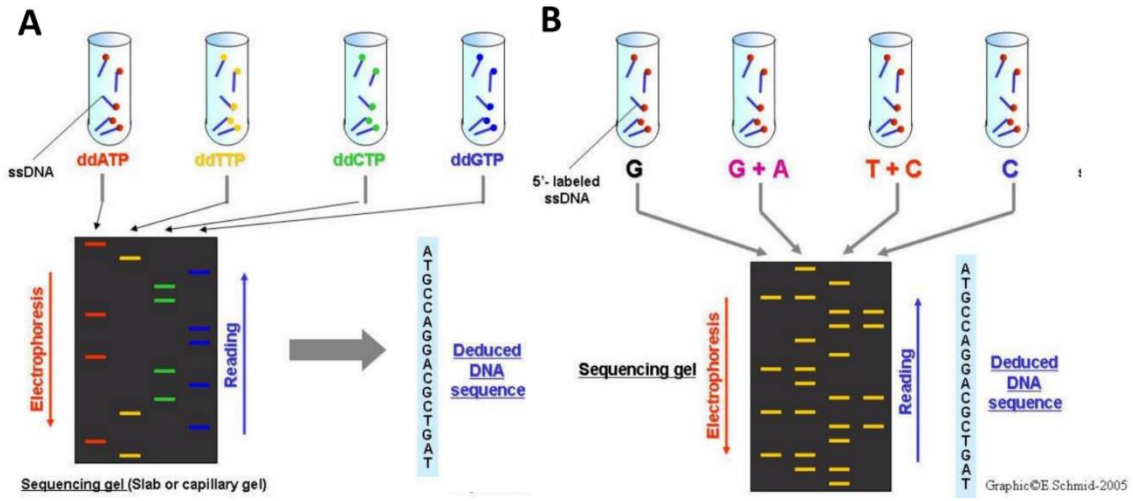


Figura 1.7. Representación del funcionamiento del A) método de secuenciación por terminación de cadena de Frederick Sanger; y B) método de secuenciación desarrollado por Maxam y Gilbert. Fuente de la imagen: http://classroom.sdmesa.edu/eschmid.

Desde su aparición, las mejoras del método de secuenciación basada en el método Sanger con el paso de los años han permitido su aplicación hasta en la secuenciación de genomas completos, entendiendo como genoma el conjunto de toda la información genética contenida en un organismo. Sin embargo, los primeros procesos de secuenciación de genomas mediante la tecnología Sanger eran altamente tediosos, costosos y requerían mucho tiempo para la obtención de secuencias completas, además de obtener un bajo rendimiento y dificultando su aplicación extendida al campo de la investigación, especialmente en laboratorios con recursos limitados [50] . Desde la secuenciación del primer genoma bacteriano en 1.995, hasta aproximadamente el año 2.005, se sucedieron toda una serie de avances tecnológicos utilizando diferentes enfoques con un mismo objetivo; generar una gran cantidad de secuencias con el menor coste posible. Así surgieron lo que se conocen como metodologías se secuenciación masiva, las cuales

dieron paso a la era de las tecnologías de secuenciación de nueva generación (SNG). Estas tecnologías utilizan un enfoque conocido como el principio de secuenciación por síntesis, en el que los nucleótidos se van incorporando a la cadena de forma controlada. Esta incorporación, dependiendo de la tecnología, se detecta a través de la emisión de algún tipo de señal, como fluorescencia (tecnologías Illumina o PacBio), o mediante la variación del pH producida por la liberación de protones durante el proceso (tecnología Ion-Torrent) [50]. En cualquier caso, el resultado es la obtención de un gran número de fragmentos de secuencia de nucleótidos conocidos como lecturas. Uno de los objetivos principales desde el nacimiento de las tecnonologías de secuenciación masiva ha sido la obtención de lecturas cada vez de mayor longitud, y en este sentido el gran avance se ha conseguido con la plataforma PacBio (Pacific Bioscience®) basada en el enfoque conocido como secuenciación en tiempo real de una única molécula de ADN o SMRT (del inglés *Single Molecule Real Time*) [51]. Esta tecnología en la actualidad es capaz de generar lecturas de hasta 40 Kb de longitud, frente a las 100 - 400 pb que se consiguen con las plataformas Illumina y Ion-Torrent. Estas tecnologías paralelizan los procesos de secuenciación de forma masiva, permitiendo generar de cientos a miles de megabases por día, obteniéndose así grandes cantidades de información en un único proceso, en relativamente poco tiempo y a un coste que ha ido descendiendo de forma significativa año tras año (Figura 1.8A). Fue precisamente el descenso en el coste de la secuenciación por genoma producido lo que contribuyó decisivamente a abrir las puertas a pequeños laboratorios de todo el mundo para incorporar estos procedimientos en sus investigaciones, al convertirlas en una opción asequible. En consecuencia, todo ello ha redundado en un incremento masivo año tras año del número de genomas disponibles en las bases de datos (Figura 1.8B).

De esta forma, se puede afirmar que existen dos grandes periodos de secuenciación: la era basada en la secuenciación por la metodología Sanger (1.995-2.005) y la era de SNG, iniciada alrededor del año 2.005 y todavía en expansión en la actualidad. Los grandes avances ocurridos en estas dos grandes etapas han ido de la mano del incremento en el conocimiento y entendimiento de los microorganismos, aportando grandes avances en campos como la taxonomía, epidemiología o la búsqueda de nuevos genes. Además, el estudio de los genomas no sólo permite descifrar las capacidades genéticas de un organismo en concreto, sino que abre la posibilidad de la comparación de esas

capacidades entre representantes de una misma especie o, incluso, de especies diferentes, en lo que se conoce como genómica comparada. En definitiva, todos estos avances permitieron el desarrollo de una de las áreas de la biología más vanguardistas y en expansión en la actualidad: la genómica, una rama interdisciplinar dedicada al estudio de los genomas desde el punto de vista funcional, evolutivo y de su origen.

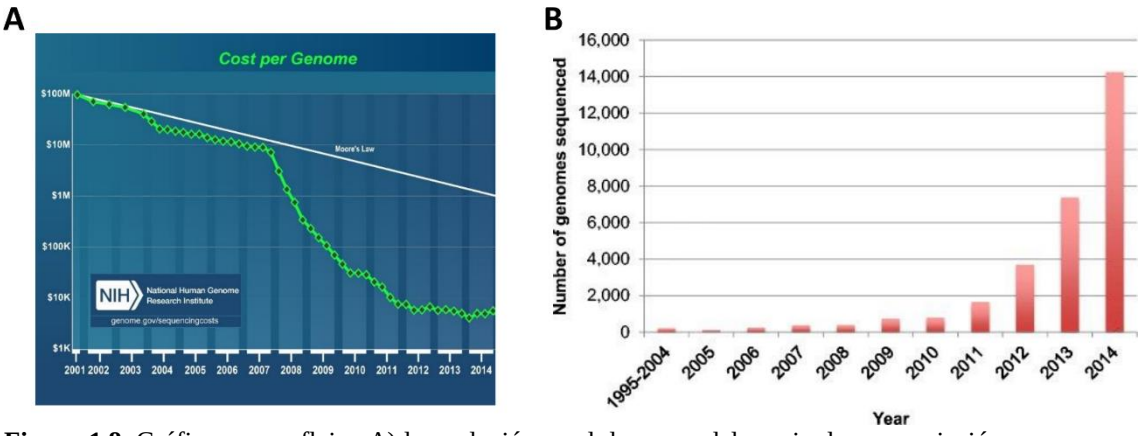


Figura 1.8. Gráficas que reflejan A) la evolución en el descenso del precio de secuenciación por genoma durante el periodo 2001-2014 (Fuente de la figura: *National Human Genome Research Institute*, https://www.genome.gov/) y B) Número de genomas depositados en las bases de datos durante el periodo 1995-2014 en GenBank (Fuente de la imagen: "*Insights from 20 years of bacterial genome sequencing*", Miriam Land y colaboradores. Functional integrative Genomics, 2015).

1.3. Secuenciación y estudios genómicos en el género *Mycobacterium*

1.3.1. Secuenciación de nueva generación en micobacterias

El estudio del género *Mycobacterium* en particular, y la microbiología en general, se ha beneficiado del uso de las tecnologías de secuenciación masiva como mecanismo para desarrollar unos conocimientos más profundos de sus entresijos genómicos. Sin embargo, en sus inicios una vez más los esfuerzos se focalizaron en especies patógenas, tales como *M. tuberculosis*, en claro detrimento de las especies de MNT. Los motivos por otra parte son evidentes: disponer del genoma de una cepa del agente causal de la tuberculosis permitía profundizar en el estudio de los mecanismos implicados en el proceso de infección al completo y en su contexto genómico original; así como la secuenciación de aislamientos realizados en pacientes alrededor del mundo, comparar sus genomas e intentar arrojar luz a la epidemiología de la especie a nivel global, entre otros muchos estudios. En definitiva, el objetivo es utilizar la genómica para entender mejor el

funcionamiento, la virulencia y la distribución a nivel mundial de estos microorganismos. De esa manera se puede generar información de gran importancia a la hora de intentar controlar su diseminación y hacerles frente en el ámbito de sus implicaciones clínicas [52].

Un buen ejemplo para reflejar estos hechos expuestos es el número de genomas de *M. tuberculosis* depositados año tras año en bases de datos públicas como GenBank, frente al de especies como *M. abscessus*, *M. abscessus* subsp. *bolletii*, *M. chelonae* y *M. immunogenum* (Figura 1.9); por su parte cada vez más relevantes desde el punto de vista clínico.

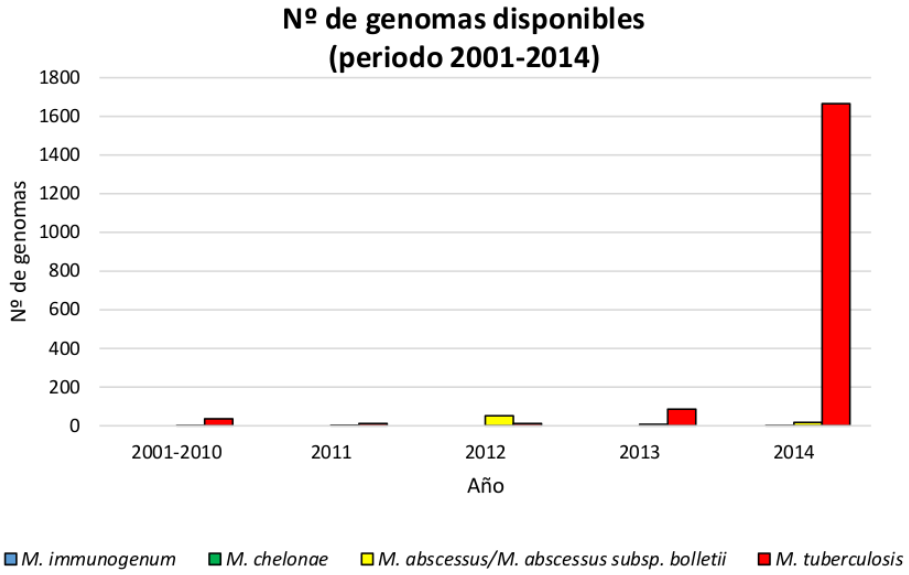


Figura 1.9. Número de genomas publicados por año en GenBank durante el periodo 2001-2014 de las especies *M. immunogenum*, *M. chelonae* y *M. abascessus/M. abscessus subsp. bolletii*. Datos obtenidos de *The National Center for Biotechnology Information* (https://www.ncbi.nlm.nih.gov/).

Estos datos, además de hacer patentes las diferencias entre los casos expuestos, también refleja el progresivo incremento en el interés por las especies de MCR, especialmente *M. abscessus* y *M. abscessus* subsp. *bolletii*. Así, probablemente fruto de las mejoras en las metodologías de identificación de los aislamientos clínicos que ha puesto de manifiesto su notable capacidad de desarrollar infecciones nosocomiales oportunistas, se puede observar un escalonado incremento de los genomas secuenciados de estas especies de MCR. A pesar de todo, *M. chelonae* y *M. immunogenum* siguen siendo dos especies pobremente representadas y sobre las que se necesitaría más información genómica con el fin de realizar de forma significativa los estudios de genómica comparada.

1.3.2. Potencial patogènico desde el punto de vista genómico

Paralelamente al desarrollo de la genómica y la SNG, se han ido desarrollando y perfeccionando toda una serie de bases de datos especializadas en aspectos concretos relativos a la información derivada de los propios genomas depositados en las bases de datos. De esta forma, podemos encontrar bases de datos especializadas en factores de virulencia, resistencias a antibióticos, proteínas reguladoras, factores sigma, conjuntos de proteínas homólogas, dominios proteicos, y un largo etcétera. Dichas bases de datos resultan de gran utilidad a la hora de realizar estudios preliminares que permitan identificar potenciales genes o proteínas implicadas en un determinado proceso y concentrar los esfuerzos posteriores en caracterizar *in vitro* ese subconjunto de proteínas identificado *in silico*. De esa manera se agiliza el proceso de comprensión de las capacidades de un determinado organismo a partir del punto de vista genómico.

En este sentido, *M. tuberculosis* y otras micobacterias patógenas suelen estar ampliamente representadas en dichas bases de datos. Los amplios y prolíferos estudios que se han realizado sobre sus genomas han permitido generar una gran cantidad de información de los elementos génicos implicados en sus mecanismos infectivos, de resistencia y demás aspectos relacionados con su patogenicidad. De esa forma, al ser depositadas en las bases de datos se dispone de un buen número de secuencias de referencia que pueden ser utilizadas para hallar homólogos de estos elementos en otros genomas. Sin embargo, dada la relevancia clínica de especies como *M. chelonae*, *M. immunogenum* y el complejo *M. abscessus*, es importante el profundizar en su estudio abordando todos los aspectos posibles, desde su faceta ambiental a su faceta de patógenos oportunistas. De esta forma se puede dibujar la ruta que siguen desde el ambiente hasta el huésped y, una vez en él, de los mecanismos de los que se valen para el desarrollo de la infección. El principal inconveniente en este último grupo es la falta de información genómica, el cual debe resolverse para poder profundizar en este tipo de estudios.

1.3.3. Impacto social de la aplicación de la genómica al estudio de las MCR

La genómica es una rama de la biología relativamente joven y en plena expansión que, como se ha destacado, permite una comprensión global de las características genéticas de un microorganismo desde un punto de vista teórico. Estas características pueden correlacionarse con los aspectos ecológicos y clínicos de una determinada especie o conjunto de especies.

Las especies *M. abscessus* (junto con *M. abscessus* subsp. *bolletii*), *M. chelonae* y *M. immunogenum* son un buen ejemplo de microorganismos que requieren de un intenso estudio de las características de sus genomas, para intentar esclarecer muchos de los aspectos implicados en su faceta de patógenos oportunistas y que tantos problemas clínicos están causando en la actualidad, especialmente en aquellos pacientes inmunodeprimidos. Incrementar este conocimiento es importante para abrir nuevas vías de investigación sobre cómo combatir de forma más eficaz este tipo de infecciones, tan difíciles de tratar por otra parte. En este sentido, los conocimientos que puede aportar la genómica pueden contribuir de forma decisiva a largo plazo en mejorar los tratamientos aplicados (facilitando la eliminación de la propia infección), en el desarrollo de sistemas de control más eficaces que dificulten la penetración de este tipo de bacterias en dependencias hospitalarias y la identificación precisa del agente causal; tan importante en grupos taxonómicamente difíciles como es el complejo de *M. abscesus*. Un gran ejemplo de las ventajas que pueden proporcionar estos estudios está relacionado con un tema de gran relevancia en la actualidad: la propagación de las resistencias a antibióticos entre los microorganismos. Se estima que en pocos años la actual generación de antibióticos podría no ser útil para combatir infecciones microbianas, y por lo tanto urge el desarrollo de una nueva generación de fármacos con este fin o de estrategias alternativas para combatir los microorganismos patógenos. El estudio comparado de los genomas, en combinación con los diferentes tipos de bases de datos y programas bioinformáticos disponibles, puede ayudar a identificar de forma relativamente rápida el conjunto de proteínas codificadas en ellos y que pueden suponer potenciales dianas para el desarrollo de nuevos tratamientos [53], o incluso ayudar a identificar los motivos genéticos subyacentes por los cuales un determinado tratamiento no está funcionando. La genómica comparada

puede a su vez ser de gran utilidad en el análisis de la diseminación de resistencias, al emprender el análisis de los diferentes proteomas (conjunto de las secuencias de todas las proteínas codificadas por un genoma) de diferentes cepas [54]. La genómica pone las bases de futuros estudios de transcriptómica y proteómica en los que se pondrán de manifiesto las circunstancias en las que se expresan los genes y los factores implicados en su regulación.

En definitiva, la secuenciación de genomas, las bases de datos resultantes, la genómica comparada y las numerosas herramientas bioinformáticas disponibles a día de hoy pueden ser una buena estrategia para arrojar luz a todas estas cuestiones abiertas y realizar una detallada y exhaustiva descripción de todos aquellos elementos génicos codificados en un genoma que definen las implicaciones ecológicas y clínicas de una determinada cepa o especie. La respuesta a estas preguntas mediante la aplicación de un enfoque basado en la genómica aplicada específicamente dentro de un grupo de creciente importancia clínica, como es el formado por las MCR *M. chelonae*, *M. immunogenum* y el complejo de *M. abscessus,* es el eje central del trabajo que se propone en la presente tesis, cuyos objetivos se expondrán a continuación.

2. Objetivos

1. Secuenciación y anotación de los genomas pertenecientes al grupo de MCR, especialmente dentro del grupo que incluye las especies *M. chelonae*, *M. immunogenum* y el complejo *M. abscessus*.
2. Análisis comparativo de genomas mediante los cálculos del genoma esencial y pangenoma.
3. Profundizar en el análisis de los mecanismos que influyen en la patogenicidad y adaptabilidad de las MCR: resistoma, mobiloma, reguloma, factores de virulencia y mecanismo implicados en el *Quorum Sensing*.
4. Realizar un catálogo de los sistemas toxina-antitoxina presentes en micobacterias ambientales, así como la demostración experimental de su funcionalidad.

3. Capítulo 1: Secuenciación y ensamblaje de genomas

3.1. Introducción

El objetivo del presente capítulo es la obtención de genomas de cepas representativas de micobacterias ambientales. El proceso de obtención de un genoma se conoce como ensamblaje. El ensamblaje parte de las lecturas generadas a través de las plataformas de secuenciación. Dichas lecturas son procesadas por programas especializados, conocidos como ensambladores, cuya finalidad es combinar la información contenida en las mismas para obtener secuencias de mayor longitud. Estas secuencias más largas, conocidas en inglés como *contigs*, pueden ser combinadas para conseguir una representación todavía más continua del genoma en un proceso de ordenado conocido como *scaffolding*. Este sistema de ordenación puede dar lugar a situaciones en la que se sabe que dos o más *contigs* van juntos, pero se desconoce la secuencia que los une. En este tipo de casos los programas rellenan el hueco entre *contigs* con indeterminaciones o Ns. El resultado del proceso de *scaffolding* son los llamados *scaffolds.* La diferencia entre un *contig* y un *scaffold* es que los *contigs* son secuencias continuas de nucleótidos sin ningna indeterminación a lo largo de su secuencia, mientras que los *scaffolds* sí las contienen. En cualquier caso, esas indeterminaciones constituyen huecos o *gaps* que pueden ser rellenados con lecturas con el objetivo de obtener una secuencia continua.

En este estudio, el ensamblaje de genomas se ha centrado en el del grupo formado por las especies *M. chelonae*, *M. immunogenum* y el complejo *M. abscessus*, miembros del grupo de MNT que más problemas causa desde el punto de vista clínico después de los patógenos representativos del género, como puede ser *M. tuberculosis.* La finalidad es ampliar la información existente del grupo incrementando el número de genomas disponibles del mismo y poder abordar así su análisis comparativo.

En base a los análisis de secuencia multi-locus o MLSA (del inglés *Multi-Locus Sequence Analysis*) realizados previamente [43] (Figura 3.1), se seleccionaron diversos aislamientos y cepas tipo incluidas en dicho grupo, así como especies cercanas bien definidas como *M. llatzerense* y *M. immunogenum*. Al inicio del estudio todas las cepas seleccionadas estaban disponibles en la colección del laboratorio de Microbiología de la Universidad de las Islas Baleares.

Para la obtención de los distintos ensamblajes se han aplicado diferentes estrategias en las que se ha evaluado el uso de distintos métodos y programas, para así determinar la opción más idónea en cada caso. El objetivo final ha sido la obtención de genomas completamente cerrados en el caso de las cepas tipo, y genomas a nivel de *draft* (genomas no cerrados) de alta calidad para el resto.

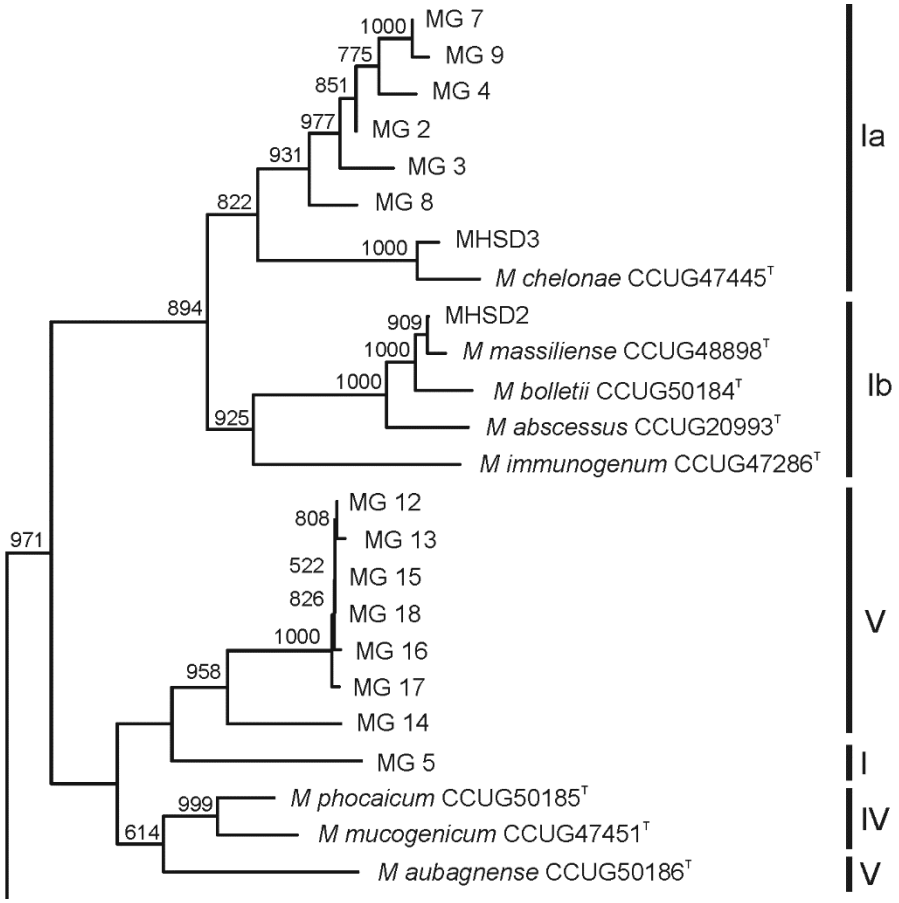


Figura 3.1. Sección superior del árbol obtenido a partir del concatenado de las secuencias del ADNr 16S y de los genes *gyr*B, *hsp*65, *rec*A, *rpo*B, *sod*A realizado por Gomila y colaboradores (2007) [43].

3.2. Material y métodos

3.2.1. Cepas seleccionadas

Las cepas seleccionadas para secuenciar sus genomas fueron *M. chelonae* CCUG 47445^T, *M. llatzerense* MG13^T, *M. immunogenum* CCUG 47286^T, *M. abscessus* subsp. *bolletii* CCUG 50184^T como cepas tipo y los aislamientos MG2, MG8, MHSD2, MHSD3, previamente descritos [43]. Otros dos aislados procedentes de sendos pacientes de la

Clínica Rotger (Palma de Mallorca, Islas Baleares, España), *M. abscessus* subsp. *bolletii* CR-UIB1 y *M. tuberculosis* CR-UIB2, fueron incluidos en el estudio (Tabla 3.1).

Tabla 3.1. Procedencia de las cepas seleccionadas para su secuenciación.

Cepa	Tipo de muestra	Origen	Año
***M. chelonae* CCUG 47445^T**	Tubérculo de tortuga	Desconocido	1923
***M. immunogenum* CCUG 47286^T**	Lavado bronco-alveolar humano	Missouri (USA)	1990
***M. llatzerense* MG13^T**	Agua para hemodiálisis	Mallorca (España)	2003
***M. abscessus* subsp. *bolletii* CCUG 50184^T**	Lavado bronquial humano	Marsella (Francia)	///
***Mycobacterium* sp. MG2**	Agua para hemodiálisis	Mallorca (España)	2002
***Mycobacterium* sp MG8**	Agua para hemodiálisis	Mallorca (España)	2003
***Mycobacterium* sp MHSD2**	Esputo humano	Mallorca (España)	2004
***Mycobacterium* sp MHSD3**	Biopsia de tejido humano	Mallorca (España)	2004
***M. abscessus* subsp. *bolletii* CR-UIB1**	Frotis de un tendón humano	Mallorca (España)	2015
***M. tuberculosis* CR-UIB2**	Tuberculosis extrapulmonar	Mallorca (España)	2015

3.2.2. Extracción de ADN

El protocolo aplicado consta de cuatro etapas bien diferenciadas: cultivo de las cepas, extracción de ADN total, análisis cuantitativo, cualitativo y confirmación taxonómica del mismo ADN (Figura 3.2).

3.2.2.1. Condiciones de cultivo de las cepas

Las cepas se cultivaron en placas de agar R2A (Sharlau, Barcelona, España) y se incubaron a 30 ºC durante 4 ó 6 días, en función de la velocidad de crecimiento de cada cepa. Posteriormente se recogió con la ayuda de un hisopo la biomasa a partir de las placas en tubos Eppendorf de 1,5 ml que contenían 1 ml de Ringer (Merck Milllipore, Billerica, Massachusetts, EEUU). La suspensión celular se homogenizó con la ayuda de un vórtex y las células se recogieron por centrifugación a 6.000 g durante 4 minutos. El sobrenadante se descartó y las células se guardaron a -20 ºC hasta el momento de su utilización.

3.2.2.2. Pretratamiento y protocolo de extracción

Las células, una vez descongeladas a temperatura ambiente, se resuspendieron en 720 µl de tampón ATL (Qiagen, Izasa, Madrid, España) y 120 µl de un stock de proteinasa K a 10 mg/ml (GE Healthcare, Chicago, Illinois, EEUU) siguiendo la relación de volúmenes recomendada por el fabricante del tampón (20 µl de proteinasa K a 10 mg/ml por cada 120 µl de tampón ATL). Se homogenizó con la ayuda de un vórtex y se incubó a 56 ºC durante 1 hora, tras lo cual se homogenizó nuevamente.

En este punto se aplicó un paso de rotura mecánica utilizando un sistema *bead-beater Disruptor Genie*® (Scientific industries, Bohemias, Nueva York, EEUU). Para ello se añadió la suspensión celular a tubos de rosca de 2 ml que contenían 0,5 g de microesferas de vidrio de 0,1 µm de diámetro, preparados y esterilizados previamente. Los tubos se dispusieron en el *Disruptor Genie*® para ser sometidos a un movimiento orbital multidireccional durante 5 minutos, permitiendo una eficiencia de choque entre las microesferas de vidrio y las células del medio suficiente para favorecer a través de esos impactos el daño o la rotura de la superficie celular.

Los sobrenadantes se limpiaron mediante un breve pulso de centrífuga y se pasaron a un nuevo tubo Eppendorf de 1,5 ml. En este punto se continuó en el paso 6 del protocolo de extracción de ADN para bacterias Gram positivas y Gram negativas del kit comercial *Wizard Genomic DNA purification* (Promega, Madison, Wisconsin, EEUU). Finalmente, para la rehidratación del ADN, y con el fin de asegurar la resuspensión de la máxima cantidad de ADN posible, se incubó 1 h a 65 ºC y se dejó reposar a 4 ºC toda la noche. El ADN obtenido se purificó con el kit *DNA Clean&Concentrator^{TM}-100* (Zymoresearch; Irvine, California, EEUU).

3.2.2.3. Comprobación del ADN

El ADN obtenido se analizó cualitativa y cuantitativamente. Para el análisis cualitativo se utilizaron 5 µl de la muestra obtenida para realizar una dilución 1/10, a partir de la cual se realizó una comprobación mediante electroforesis en geles de agarosa de 0,8 % (p/v), a 100 V durante 35 minutos. La digitalización de los geles se realizó con la ayuda del capturador de imagen *Universal Hood II* (Bio-Rad) y utilizando el programa informático

Image-Lab v2.0.1 build 18 del propio sistema. El ADN se cuantificó con el sistema *Nanodrop 2000c spectrophotometer* (Thermo Fisher Scientific, Waltham, Massachusetts, EEUU). El valor cuantitativo de pureza se determinó a partir de las relaciones de los valores de absorbancias 260/280 y 260/230.

3.2.2.4. Confirmación de la procedencia del ADN

A partir de 10-40 ng de ADN se realizó la amplificación por PCR del ADNr 16S y de los genes *gyr*B, *rpo*B y *hsp*65, los cuales forman parte de los llamados *housekeeping genes*. El tamaño estimado para los amplicones se confirmó por electroforesis en geles de agarosa de 1,5 % (p/v) (100 V, durante 40 minutos), tras lo cual se purificaron con el kit *Illustra GFX PCR DNA and Gel Band Purification* (GE Healthcare, Chicago, Illinois, EEUU). Los productos purificados se volvieron a comprobar por electroforesis en las condiciones ya indicadas.

Los productos purificados se utilizaron para llevar a cabo las reacciones de secuenciación por el método de Sanger con el kit *BigDye® Terminator v3.1 Cycle Sequencing* (Applied Biosystems™, Foster City, California, EEUU) siguiendo las indicaciones del fabricante y utilizando 5-10 ng de ADN. El lavado de los productos obtenidos se realizó mediante dos pasos sucesivos de precipitación por centrifugación. Brevemente, los productos obtenidos en la reacción de secuenciación se diluyeron con 90 µl de agua MiliQ y se pasaron a un Ependorff que contenía 250 µl de etanol absoluto y 10 µl de acetato sódico 3 M pH 5,2. Tras mezclar con vórtex, se centrifugó durante 30 minutos a 16.000 g. El sobrenadante se descartó y se añadieron 300 µl de etanol al 70 % (v/v) para volver a centrifugar durante 20 minutos a 16.000 g. El etanol se aspiró cuidadosamente y se dejó secar durante 15 minutos a 30 ºC para evaporarlo completamente.

Las muestras se mantuvieron congeladas a -20 ºC hasta el momento de su carga en un *Genetic Analyzer 3130* (Applied Byosistems™, Foster City, California, EEUU), paso realizado por los servicios cientificoténicos de la Universidad de las Islas Baleares. Para ello se resuspendieron en 20 µl de agua MiliQ. Las secuencias obtenidas se compararon con las secuencias de ADNr 16S y los genes *hsp*65, *rpo*B y *gyr*B previamente obtenidas por Gomila y colaboradores [43] (Tabla 3.2).

Tabla 3.2. Números de acceso de las secuencias de genes "*housekeeping*" originales de las cepas secuenciadas.

	CCUG 47445^T	**CCUG 47286^T**	**MG2**	**MG8**	**MHSD2**	**MHSD3**
ADNr 16S	AY457072	AY457080	AJ746059	AJ746065	AM421246	AM421247
***hsp*65**	AF547818	AY458081	AM421333	AM421338	AM421349	AM421350
***gyr*B**	AM421324	AM421326	AM421295	AM421301	AM421246	AM421315
***rpo*B**	AY147163	AY262739	AM421380	AM421386	AM421398	AM421399

3.2.3. Secuenciación de genomas con plataformas de SNG

La secuenciación de las muestras de ADN genómico se llevó a cabo en dos empresas externas: Lifesequencing S.L. (Paterna, Valencia) y BaseClear B.V. (Leiden, Holanda). Las plataformas de secuenciación seleccionadas fueron Illumina HiSeq 2500 para la generación de lecturas *Paired-End* aplicando 2x100 ciclos *Paired-End* (PE), con un tamaño de inserto de la librería de 250-350 pb; la plataforma *Pacific Bioscience* para la generación de lecturas largas o LR (del inglés, *Long-reads* PacBio RSII, librerías de 10 kb) mediante la tecnología SMRT (del inglés *Single Molecule Real Time*); y la plataforma 454 GS FLX-Titatium (Roche Diagnostics, Barcelona, España) para la generación de lecturas MP (del inglés *Mate-Pair*) con un tamaño de inserto de librería de 8-16 kb. En cada caso se siguieron los protocolos establecidos para cada plataforma y optimizaciones recomendadas por las respectivas empresas.

3.2.4. Obtención de los genomas

Para la obtención de ensamblajes de alta calidad se aplicaron varias estrategias, especialmente diferentes en función de si el objetivo era obtener un genoma completamente cerrado (como en el caso de las cepas tipo), o un genoma a nivel de *draft* (en el caso del resto de cepas), tal y como se ha indicado anteriormente. Las diferentes etapas del proceso fueron: 1) preparación de las lecturas, 2) ensamblaje, 3) mejora del ensamblaje mediante el rellenado de huecos, 4) validación y 5) anotación. Los programas utilizados, así como las condiciones determinadas para cada caso se explican a continuación.

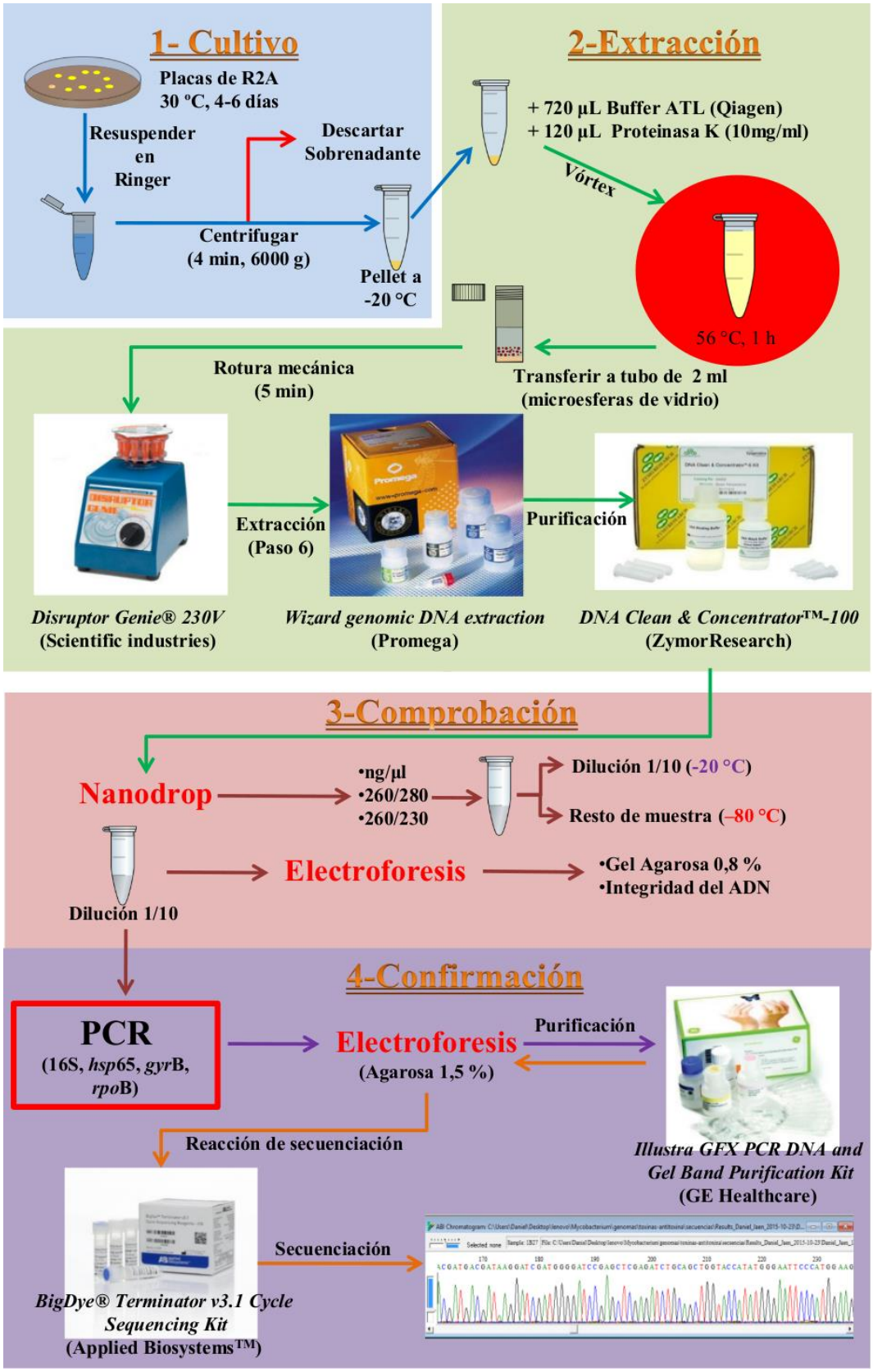


Figura 3.2. Esquema del protocolo para la preparación de ADN destinado a la secuenciación por plataformas SNG. Se indican los pasos incluidos en las cuatro grandes etapas del mismo: 1) cultivo, 2) extracción, 3) comprobación y 4) confirmación.

3.2.4.1. Preparación de las lecturas

Las lecturas obtenidas fueron filtradas para la obtención de lecturas de alta calidad. Este paso se aplicó únicamente a las lecturas Illumina. Se consideraron lecturas de alta calidad aquellas con índice de calidad Phred (Q) superior a 30. Este índice de calidad se otorga de forma individual a cada nucleótido, reflejando la probabilidad de que ese determinado nucleótido sea erróneo (P). Un índice de calidad Phred 30 determina que esa probabilidad de error sea de 1 entre 1000, asegurando una precisión del 99,9 %. El cálculo de Q refleja una relación logarítmica entre él mismo y P [54,55].

$$\mathbf{Q = -10 \log_{10} P}$$

Para el filtrado de las lecturas se utilizaron tres programas. El primero de ellos fue Sickle v1.2 [56], utilizando la modalidad específica para PE (opción -pe), indicando el descarte de todas aquellas lecturas con indeterminaciones en su secuencia, marcadas como N (opción -n) y extrayendo todas aquellas lecuras sin pareja (*Single reads*, SR) en un archivo separado (opción -s). La segunda estrategia fue a través de la herramienta de filtrado de secuencias (*Trim sequences*) de programa CLC Genomics Workbench v6.5.1 (CLC bio, Aarhus, Dinamarca). El tercer y último programa utilizado fue BBMap v35.34 (https://sourceforge.net/projects/bbmap). La calidad de las lecturas, basadas en el índice de calidad Phred, se comparó antes y después del proceso de filtrado con el programa FastQC v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) para comprobar la mejora en la media del índice de calidad de las mismas.

A partir de los archivos de lecturas Illumina de alta calidad (.fastq) se generaron archivos de cobertura con el programa fastqSample del paquete CABOG (*Celera Assembler with Best Overlap Graph*) [57], utilizando los tamaños estimados de cada genoma (opción -g), una longitud mínima de 100 nucleótidos por lecturas (opción -l 100) y fijando una cobertura de 50 veces el genoma (opción -c 50). Esta cobertura se seleccionó en base a los resultados obtenidos en el Trabajo Fin de Master (TFM) de Francisco Salvà-Serra, en el que se determinó como la cobertura idónea para la obtención de ensamblajes de alta calidad optimizando los recursos de cálculo disponibles.

3.2.4.2. Ensamblaje

Para la obtención de los genomas se utilizaron diferentes programas con el fin de abarcar diferentes estrategias de ensamblaje en cada caso, partiendo en todos ellos de lecturas de alta calidad. En aquellos casos donde fue necesaria la selección de un tamaño de K-mer óptimo se realizó de forma automática utilizando los programas KmerGenie [58] y la opción *stimate best K-mer size* de Vague [59], siendo K-mer todas las posibles subsecuencias de una determinada longitud (k) que pueden obtenerse a partir de una lectura. Los programas, así como las condiciones específicas en cada caso, se enumeran a continuación:

- **CLC Genomics Workbench v6.5.1 (CLC bio, Aarhus, Dinamarca):** partiendo de lecturas Illumina de alta calidad se realizó un ensamblaje *de novo*. Se estableció la selección automática del los parámetros *Word size*, *Bubble size* y la detección automática de la distancia entre las lecturas PE. Se seleccionó el tamaño mínimo de *contig* por defecto (1.000 pb) y se anuló la opción de realizar *scaffolding* a partir de los *contigs* obtenidos.

- **ABySS v1.5.1** [60]**:** los ensamblajes con este programa se llevaron a cabo con librerías PE de Illumina, lecturas sueltas de un extremo procedentes del descarte de las librerías PE durante el filtrado de calidad (sin pareja o eliminada al ser de baja calidad).

- **HGAP V2.0.2:** el protocolo HGAP (del inglés *Hierarchical Genome Assembly Process*) [61] (Figura 3.3) se basa en la utilización de LR. Brevemente, las LR fueron alineadas entre sí en un proceso de mapeo sobre las LR más largas, pre-seleccionadas previamente. Este proceso se realizó con el programa BLASR (del inglés *Basic Local Alignment with Successive Refinement*) [62]. El algoritmo pbdagcon se utilizó para obtener secuencias consenso de mayor longitud a partir de estos alineamientos, donde los nucleótidos erróneos se autocorrigieron por superposición de lecturas. Esta primera etapa se conoce como Preesamblaje, y los datos obtenidos como lecturas preensambladas. Estas lecturas se utilizaron para la realización del ensamblaje propiamente dicho a través del programa CABOG

[57]. Este programa ensambla las lecturas por solapamiento (CA/*overlap*), buscando la mejor superposición al final de cada lectura (CA/*unitigger*), para obtener finalmente una secuencia consenso que corresponde a los *contigs* finales (CA/*utgcns*). Los *contigs* obtenidos fueron analizados con Quiver (https://github.com/PacificBiosciences), con el fin de reducir el número de errores que se hubieran podido producir durante el proceso de ensamblaje.

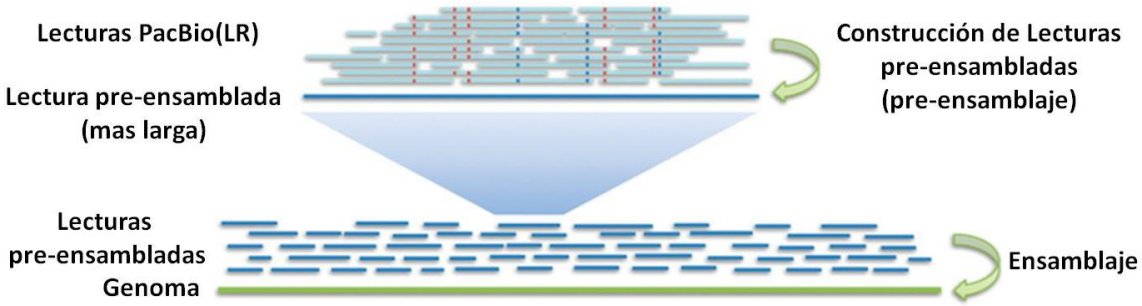


Figura 3.3. Principios del protocolo de ensamblaje HGAP.*Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data*, Chen-Shan Chin y col. (2013)[61]

- **AllPaths:** este programa altamente especializado de ensamblaje se utilizó con lecturas PE de Illumina, lecturas PE simuladas (*wgsim*) con un tamaño de inserto de 180 pb y una longitud de 100 pb a partir de lecturas o ensamblajes de alta calidad PacBio a una cobertura 50x, así como dos *Long Jumping Distance* (LJD) Illumina simuladas (*fastqSimulate*) con separaciones de 3 y 10kb.

- **Velvet v1.1.04** [63]**:** utilizado para la obtención de ensamblajes a partir de lecturas PE, este ensamblador basa su procedimiento en diagramas de *de Brujin* en lugar de solapar las lecturas. La ejecución del programa consta de dos partes bien diferenciadas. En primer lugar, se utilizó el programa Velveth para preparar los datos a utilizar en la siguiente etapa. En este paso se creó un directorio donde se depositarían los archivos generados (*output_directory*), se seleccionó una longitud de K-mer (en este caso de forma automática tal y como se explicó previamente), se indicó el formato del archivo de las lecturas (-fastq) y el tipo de las mismas (*-shortPaired).* Realizado este paso se ejecutó el subprograma Velvetg para realizar el ensamblaje propiamente dicho. Para este paso, además de la ubicación del directorio creado por Velveth, se indicó la cobertura esperada (-

exp_cov), el mínimo de cobertura requerida (-*cov_cutoff*), el tamaño mínimo de *contig* (-*min_contig_lgth*) y especificando que no debía realizarse *scaffolding* con los *contigs* obtenidos (-*scaffolding* no).

- **SPAdes v3.0** [64]**:** otro programa también basado en diagramas de *de Brujin*, puede utilizar lecturas PE, MP, lecturas desapareadas, LR e incluso datos generados por la plataforma SNG Ion-Torrent. En este caso se utilizó para la obtención de ensamblajes a partir de lecturas PE. Aunque por defecto el programa realiza el filtrado de lecturas, se desactivó esta opción indicando que solo debía proceder con el ensamblaje (--*only-assembly*) ya que se partía de archivos de lecturas de alta calidad previamente filtradas. Se utilizaron los valores por defecto del resto de parámetros, indicando únicamente los archivos de lecturas forward (-1) y reverse (-2) y tratando de reducir el número de errores e inserciones-delecciones cortas con la opción -careful.

- **GS-de Novo Assembler (Roche. Basilea, Suiza):** este ensamblador está especialmente diseñado para el ensamblaje de lecturas MP obtenidas con la plataforma 454 GS FLX-Titatium de Roche Diagnostics. Pero también permite la opción de utilizarlo con lecturas PE y basa su funcionamiento en el solapamiento de lecturas. Partiendo de lecturas PE, se indicó en la interfaz gráfica del programa el tipo de lecturas (opción *FASTQ reads* y *Paired-End*), un tamaño mínimo de *contigs* de 1.000 pb, dejando el resto de opciones por defecto.

3.2.4.3. Mejora del ensamblaje

Con el fin de cerrar los genomas de las cepas tipo u obtener ensamblajes de mayor calidad y continuidad en el caso del resto de cepas, se utilizaron varias herramientas y se optimizaron los protocolos para la consecución de este objetivo:

- **SSPACE-LongRead** [65]**:** partiendo de *contigs* generados con alguno de los ensambladores previamente descritos y LR, se utilizó este programa para ordenar y orientar dichos *contigs* aprovechando la información espacial proporcionada por las lecturas obtenidas con la tecnología SMRTbell, en un proceso conocido como

scaffolding híbrido. Se alinearon las LR contra los *contigs* preensamblados a través de BLASR (considerados en este caso "las lecturas más largas"), utilizando el valor por defecto de 200 pb de solapamiento mínimo para dar por válido un alineamiento, reduciendo así la posibilidad de obtener falsos positivos (opción -g). Basado en el resultado de los alineamientos sobre las LR se determinó la posición y orientación entre los mismos, calculándose al mismo tiempo la posible distancia que les separa. Los huecos situados entre los *contigs* ordenados se rellenaron con indeterminaciones (N), obteniendo *scaffolds* representativos del orden de los *contigs* en el genoma en cuestión.

- **ABACAS (del inglés *Algorithm-Based Automatic Contiguation of Assembled Sequences*)** [66]**:** incluido en el paquete de programas PAGIT (del inglés *Post-Assembly Genome-Improvement Toolkit*, del Instituto Sanger) [67], este programa se basa en la utilización de un genoma de referencia para la ordenación y orientación de los *contigs* obtenidos en un ensamblaje. La elección del genoma de referencia es el punto crítico del proceso ya que debe asegurarse que es un genoma próximo al de la cepa cuyo genoma queremos mejorar. En este caso, para asegurar ese punto se realizaron cálculos de ANIs (del inglés *Average Nucleotide Identities*) con el programa JSpecies v1.2.1 [68], y se consideraron como genomas de referencia válidos aquellos cuyo valor de ANI era igual o superior al 95%. Con el genoma de referencia seleccionado y el archivo con los *contigs* a ordenar (ambos en formato FASTA), se indicó al programa el nombre del archivo de referencia y el nombre del archivo a procesar, utilizando el resto de parámetros por defecto. Los huecos presentes entre los *contigs* una vez ordenados se rellenaron nuevamente con indeterminaciones, obteniéndose así *scaffolds* representativos del orden del genoma (pseudogenoma).

- **GapFiller v1.10** [69]**:** Utilizado para el rellenado de aquellos huecos o *gaps* presentes en los *scaffolds* obtenidos, haciendo uso para ello librerías de lecturas PE de alta calidad, así como también lecturas MP en el caso de que estuviesen disponibles. Para ello se configuró un archivo de texto basado en 7 columnas con la información necesaria para el funcionamiento del programa: nombre de la

librería (columna 1), algoritmo de alineamiento a utilizar (en este caso bwa, columna 2), nombre de los archivos de lecturas PE (columnas 3 y 4), tamaño de inserto de las librerías (entre 250 y 450 pb para lecturas PE de Illumina y entre 3.000 y 8.000 pb para lecturas MP, columna 5), el porcentaje de desviación permitida en el tamaño de inserto (0,25 (25 %), columna 6) y si los archivos de lecturas corresponden a lecturas en sentido directo o *forward* (F) o lecturas en sentido inverso o *reverse* (R) respectivamente (columnas 7). Configurado este archivo se lanzó el cálculo con el programa indicando el nombre del archivo FASTA del ensamblaje (-s) y el archivo de texto con la configuración a utilizar (-l), indicando además que se realizaran 10 iteraciones del proceso (-i 10). El resto de parámetros del programa se utilizaron con los valores optimizados por defecto.

Los genomas fueron procesados con scripts de Perl propios con el fin de eliminar todas las indeterminaciones presentes en los ensamblajes para trabajar con genomas formados únicamente por *contigs*.

3.2.4.4. Validación de ensamblajes

La validación de los ensamblajes se basó en la aplicación de una serie de programas para la detección de errores e incongruencias de ensamblaje, así como herramientas para comparar diferentes ensamblajes de un mismo genoma obtenidos con diferentes programas y estrategias, para hacer una relación ordenada y seleccionar el más óptimo en términos de continuidad y calidad.

- **GS-Assembler Mapper v2.7 (Roche, Basilea, Suiza):** este programa se utilizó para el mapeo de los genomas con los archivos de cobertura 50x de lecturas de alta calidad. Al programa se le proporcionó el genoma a mapear (opción *Reference*) y los dos archivos de lecturas (en este caso en la opción "*FASTA and FASTQ reads*" al tratarse de lecturas PE), dejando el resto de parámetros por defecto. El reclutamiento de lecturas se visualizó con el programa Tablet v1.14.10.21 [70] a partir del archivo BAM obtenido por defecto. Este programa permite ver la cobertura en lecturas alcanzada a lo largo del genoma, favoreciendo la localización de regiones con coberturas mayores o inferiores a lo esperado;

consideradas como zonas conflictivas y/o repetitivas que requerían ser analizadas con más detalle.

- **REAPR (del inglés *Recognition of Errors in Assembly using Paired Reads*)** [71]**:** este protocolo nos permitió evaluar la exactitud o precisión de los ensamblajes utilizando la información que recibe de lecturas mapeadas (PE y/o MP) sobre el ensamblaje del genoma, sin necesidad de utilizar un genoma de referencia. Aquellas regiones que el programa interpreta como mal ensambladas, basándose en los errores encontrados, son rotas y separadas como *contigs*. El protocolo se utilizó para realizar un análisis nucleótido a nucleótido de la secuencia; identificando sustituciones de bases, inserciones o deleciones, así como errores estructurales derivados del cambio en la distribución de la cobertura esperada o lecturas fuera del rango de tamaño de inserto teórico utilizado en el proceso de secuenciación. De esta forma, al final del proceso se obtuvo un nuevo archivo FASTA con el ensamblaje sin puntos conflictivos y la nueva combinación de *contigs*. Los *contigs* o *scaffolds* obtenidos en ese archivo final se consideraron como regiones ensambladas de gran precisión acorde con la información contenida en las lecturas. El protocolo completo REAPR se aplicó íntegramente, incluyendo las opciones iniciales de comprobación del formato y nomenclatura de los *contigs* y/o *scaffolds* del genoma ensamblado *de novo* (opción *facheck*), el mapeado con las opciones *perfectmap* (*reapr perfectmap*, donde es importante indicar el tamaño de inserto de la librearía) y *smaltmap* (*reapr smaltmap*). Finalmente, con los archivos de mapeo (BAM) correspondientes y el ensamblaje a evaluar se inició la fase de análisis propiamente dicha (*reapr pipeline*). Los ensamblajes resultantes al final de este proceso se consideraron como óptimos en función de la información disponible y fueron utilizados en las posteriores aplicaciones.

- **Curvas FRC (del inglés *Feature Reponse Curves*)** [72]**:** en los casos en los que se realizaron varios ensamblajes del mismo genoma, se utilizaron las curvas FRC para analizar el contenido de errores a lo largo de los mismos con el fin de escoger el mejor ensamblaje, aquél con menor contenido de errores acumulados. En este

punto se utilizó un programa de procesamiento por lotes diseñado por el Dr. Antoni Bennasar Figueras para la automatización de los protocolos, en el que introducimos además del nombre de cepa, el prefijo de las parejas de archivos de lecturas PE, tamaños mímino y máximo de inserto de las librerías (misma información para las lecturas MP en el caso de estar disponibles), tamaño esperado del genoma y el nombre del archivo que contiene el ensamblaje. Los valores numéricos se utilizaron para la representación gráfica de las respectivas curvas.

- **QUAST v2.3 (del inglés *Quality ASsessment Tool*)** [73]**:** este programa se utilizó para la comparación de los ensamblajes de un mismo genoma. Concretamente, de toda la información proporcionada por el programa se centró la atención en el número de *contigs*, longitud total del genoma, % de GC, el estadístico N50 (que se define como la longitud de los *contigs* que ordenados de igual o mayor tamaño se obtiene la mitad de los nucleótidos del genoma; y se calcula ordenando todos los *contigs* obtenidos de mayor a menor y determinando el conjunto mínimo de *contigs* cuyo tamaño total sea la mitad de la del genoma, tomando como valor N50 el valor del tamaño del *contig* con el que se consigue llegar al 50 % del tamaño del genoma), N75 (tamaño del *contig* con el que se consigue llegar al 75 % del tamaño del genoma), L50 (número de *contigs* necesarios para alcanzar el 50 % de la longitud del genoma), L75 (número de *contigs* necesarios para alcanzar el 75 % de la longitud del genoma) y el número de indeterminaciones o Ns por cada 100 kb.

3.2.4.5. Anotación de ensamblajes

La anotación de los genomas de las cepas tipo y del resto de cepas, para su publicación en la base de datos NCBI, se realizó a través del protocolo PGAP (del inglés *Prokaryotic Genome Annotation Pipeline*) [74]. Para los trabajos de comparación realizados, tanto los genomas propios como los obtenidos del NCBI fueron reanotados con el programa Prokka v1.10 [75] para normalizar la anotación de todos los genomas bajo los mismos criterios.

3.2.4.6. Ampliación y mantenimiento de la base de datos de MCR

Los genomas obtenidos fueron publicados en la base de datos GenBank [76] de NCBI. A su vez fueron incorporados en la base datos interna de micobacterias de crecimiento rápido del *Grup de Recerca de Microbiologia* creada en el ámbito del proyecto CGL2012-39604, en el cual se enmarca la presente tesis (http://microbiologia.uib.es/bioinformatica/mcr).

3.3. Resultados

3.3.1. Extracción y secuenciación

Los datos referentes a los ADNs obtenidos se recogen en la Tabla 3.3.

Tabla 3.3. Características de las muestras de ADN obtenidas aplicando el protocolo optimizado. La cantidad de ADN (en µg) se calcula sobre un volumen de 145 µl, excepto en la muestra MHSD2-D (88 µl).

Muestra de ADN	Concentración (ng/µl)	Cantidad total (µg)	260/280	260/230
MG8-A	63,3	9,05	1,85	2,06
MG8-B	265,4	37,9	2,15	1,91
MG2-A	122	17,4	1,91	2,18
MG2-B	86,8	12,4	1,91	2,21
MHSD2-C	112,6	16,1	1,96	2,19
MHSD2-D	121,2	10,6	1,9	2,34
MHSD3-A	144,1	20,6	1,92	2,15
MHSD3-B	157,7	22,5	1,9	2,18
CR-UIB1	477,5	69,2	1,9	2
CR-UIB2	116	16,8	1,97	2,37
M.immunogenum-A	259,8	37,6	1,92	2,13
M.immunogenum-B	206,4	29,9	1,92	2,08
M.llatzerense-A	185,6	26,9	1,92	2,14
M.llatzerense-B	120,9	17,5	1,92	2,13
M. chelonae-A	143,8	11,5	1,89	1,67
M. abscessu subsp. bolletii-A	227,3	32,9	1,89	1,80

Las relaciones de valores de absorbancia 260/230 y 260/280 muestran la excelente calidad y pureza del ADN obtenido según los estándares aceptados para su utilización en los

protocolos de secuenciación de SNG, según los cuales un ADN de buena calidad y pureza debe tener una relacion de absorbancias 260/280 comprendido entre 1,7-2 y una relación 260/230 superior a 1,5 *(Protocols and Applications Guide*, Promega).

Las muestras MG2-A, MG8-B, MHSD2-D, MHSD3-A, CR-UIB1, CR-UIB2 *M. llatzerense*-A y *M.immunogenum*-A, *M. chelonae*-A, *M. abscessus* subsp. *bolletii*-A fueron enviadas a secuenciar después de ser confirmadas por las secuencias de los genes ADNr 16S, *hsp65*, *rpoB*, *gyrB*. Los resultados para la plataforma Illumina y PacBio se recogen en las Tablas 3.4 y 3.5.

Tabla 3.4. Rendimientos de secuenciación obtenidos con la plataforma Illumina HiSeq-2500.

Cepa	Nº de lecturas	Longitud media (pb)	Total (Mb)	Indice Phred (Media)
M. chelonae CCUG 47445^T	7.526.798 (x2)	50,95	766	37
M. immunogenum CCUG 47286^T	2.863.242 (x2)	126	720	33
M. abscessus subps. *bolletii* CCUG 50184^T	9.252.842 (x2)	99	1.832	35
M. llatzerense MG13^T	7.903.724 (x2)	101	1.596	32
MG2	2.864.666 (x2)	126	720	34
MG8	3.489.681 (x2)	126	878	34
MHSD2	3.600.793 (x2)	126	906	34
MHSD3	3.362.577 (x2)	126	846	34
CR-UIB1	696.708 (x2)	240	334,4	32
CR-UIB2	2.464.806 (x2)	126	612	31

Tabla 3.5. Rendimientos de secuenciación obtenidos con la plataforma PacBio RSII.

Cepa	Nº de lecturas	Longitud media (kb)	Total (Mb)	GC (%)
M. chelonae CCUG 47445^T	136.465	2.266	425	62,34
M. immunogenum CCUG 47286^T	61.476	2.677	164	63,2
M. llatzerense MG13^T	192.883	6.340	1.223	64

Los resultados obtenidos mediante la plataforma 454 GS FLX-Titatium de Roche Diagnostics, que sólo fue utilizada con la cepa tipo de *M. llatzerense*, proporcionaron

371.061 lecturas MP, de las cuales fueron útiles 252.057 con un tamaño medio de 410 nt por lectura y un rendimiento de 103 Mb aproximadamente.

3.3.2. Resultados de ensamblaje

3.3.2.1. *Mycobacterium chelonae* CCUG 47445^T, *Mycobacterium immunogenum* CCUG 47286^T, *Mycobacterium llatzerense* MG13^T y *Mycobacterium abscessus* subps. *bolletii* CCUG 50184^T.

El objetivo de este apartado fue el cierre completo de los genomas de las cepas tipo, a excepción del genoma de la cepa tipo de *M. abscessus* subps *bolletii* 50184^T. Se descartó el cierre de este último genoma debido a que al inicio del proyecto aparecieron en las bases de datos un gran número de genomas cerrados de esta misma especie incluyendo la cepa tipo BD^T, por lo que se decidió obtener un genoma a nivel de *draft* de alta calidad y centrar esfuerzos en el cierre del genoma de las otras cepas tipo. En la Tabla 3.6 se muestran los resultados de los ensamblajes obtenidos.

Tabla 3.6. Resultados de los protocolos de ensamblaje aplicados para cada una de las cepas tipo.

Cepa	*Contigs*	Tamaño (pb)	GC (%)
M. chelonae CCUG 47445^T	1	5.029.817	63,92
M. immunogenum CCUG 47286^T	1	5.573.781	64,27
M. llatzerense MG13^T	13	6.274.239	66,36
M. abscessus subps. *bolletii* CCUG 50184^T	26	5.053.525	64,06

El ensamblaje del genoma de *M. chelonae* se inició con lecturas PE que fueron filtradas y ensambladas con CLC Genomics Workbench v6.5.1. Dicho ensamblaje fue sometido a un proceso de *scaffolding* híbrido con el programa SSPACE-Long Read scaffolder v1.0, utilizando para ello todas las LR disponibles. Los huecos del *scaffold* obtenido fueron cerrados con lecturas PE utilizando el programa GapFiller. Por su parte, para el ensamblaje del genoma de *M. immunogenum*, el filtrado de lecturas se realizó con BBMap v35.34, y todas las lecturas de alta calidad obtenidas fueron ensambladas *de novo*, con ABySS v1.5.1. En este punto se aplicó el mismo protocolo de *scaffolding* híbrido y rellenado de huecos seguido para *M. chelonae*. Sin embargo, en este caso los huecos solo fueron rellenados parcialmente, por lo que se procedió a generar una base de datos a partir de *contigs* previamente ensamblados con Velvet v1.1.04 y lecturas Illumina de alta

calidad con una cobertura 50x. Esta base de datos se utilizó para la confirmación y el cierre de los huecos restantes, mientras que la resolución y confirmación de puntos conflictivos del ensamblaje se realizó manualmente y con ayuda de BLAST. El resultado fueron dos genomas completamente cerrados de 5.029.817 pb para *M. chelonae* CCUG 47445^{T} y de 5.573.781 pb para *M. immunogenum* CCUG 47286^{T}.

Por su parte, el genoma de la cepa tipo *M. llatzerense* se ensambló utilizando múltiples programas y estrategias para conseguir completarlo, además de combinar lecturas Illumina, PacBio y 454 GS FLX-Titatium. Pese a los diferentes enfoques abordados, no se ha conseguido cerrar el genoma. El mejor ensamblaje obtenido constó de 13 *contigs* y un tamaño total de 6.274.239 pb y se escogió en base a los resultados obtenidos de las curvas FRC. El proceso aplicado para la obtención de este genoma se inició ensamblando todas las LR a través del protocolo HGAP. A partir de tres ensamblajes obtenidos HGAP se realizaron simulaciones de librerías con lecturas de longitud y tamaños de inserto deseadas (longitud de lecturas de 100 pb e inserto de 180 pb, longitud de lecturas de 100 pb e inserto de 3 Kb, longitud de lecturas de 100 pb e inserto de 10 Kb) a través del programa wgsim. A partir de lecturas Illumina-PE filtradas con el programa Sickle, se preparó un archivo de cobertura 50x que, en combinación con las tres librerías previamente prediseñadas, fueron utilizadas para su ensamblaje con AllPaths-LG v4.6. Posteriormente se fueron incorporando al proceso de forma iterativa cada uno de los tres archivos de lecturas PacBio disponibles, mejorando el ensamblaje de forma progresiva. Finalmente, se rellenaron parcialmente los huecos con GapFiller y se eliminaron el resto de indeterminaciones para acabar obteniendo el ensamblaje reflejado en la Tabla 3.6.

Por último, para el genoma de *M. abscessus* subsp. *bolletii* CCUG 50184^{T} se realizó un ensamblaje utilizando el programa Velvet v1.1.04, utilizando un archivo de cobetura 50x de lecturas PE de alta calidad. Los *contigs* resultantes de este ensamblaje fueron ordenados utilizando el programa ABACAS incluido en PAGIT y utilizando el genoma de *M. abscessus* subps. *bolletii* 50594 (número de acceso CP004374.1) como referencia. Los huecos de los *scaffolds* resultantes fueron rellenados una vez más con el programa GapFiller, utilizando todas las lecturas Illumina de alta calidad disponibles, eliminando posteriormente las indeterminaciones restantes, para acabar obteniendo un ensamblaje de 26 *contigs* y 5.053.525 pb de longitud.

3.3.2.2. Cepas MG2, MG8, MHSD2, MHSD3, CR-UIB-1 y CR-UIB2

A partir de archivos de lecturas PE de alta calidad y una cobertura 50x se realizaron ensamblajes paralelos con los programas Velvet v1.1.04, SPAdes v3.0 y GS - de novo Assembler. En el caso del aislamiento clínico CR-UIB-1 se utilizó una cobertura 25x al no disponer lecturas de alta calidad suficientes para trabajar con una cobertura mayor. Una vez revisados los errores acumulados mediante curvas FRC, se optó por aquel que presentase un menor contenido de errores. Según este criterio, en todos los casos el mejor ensamblaje fue el obtenido mediante Velvet (Figuras 3.4 y 3.5A).

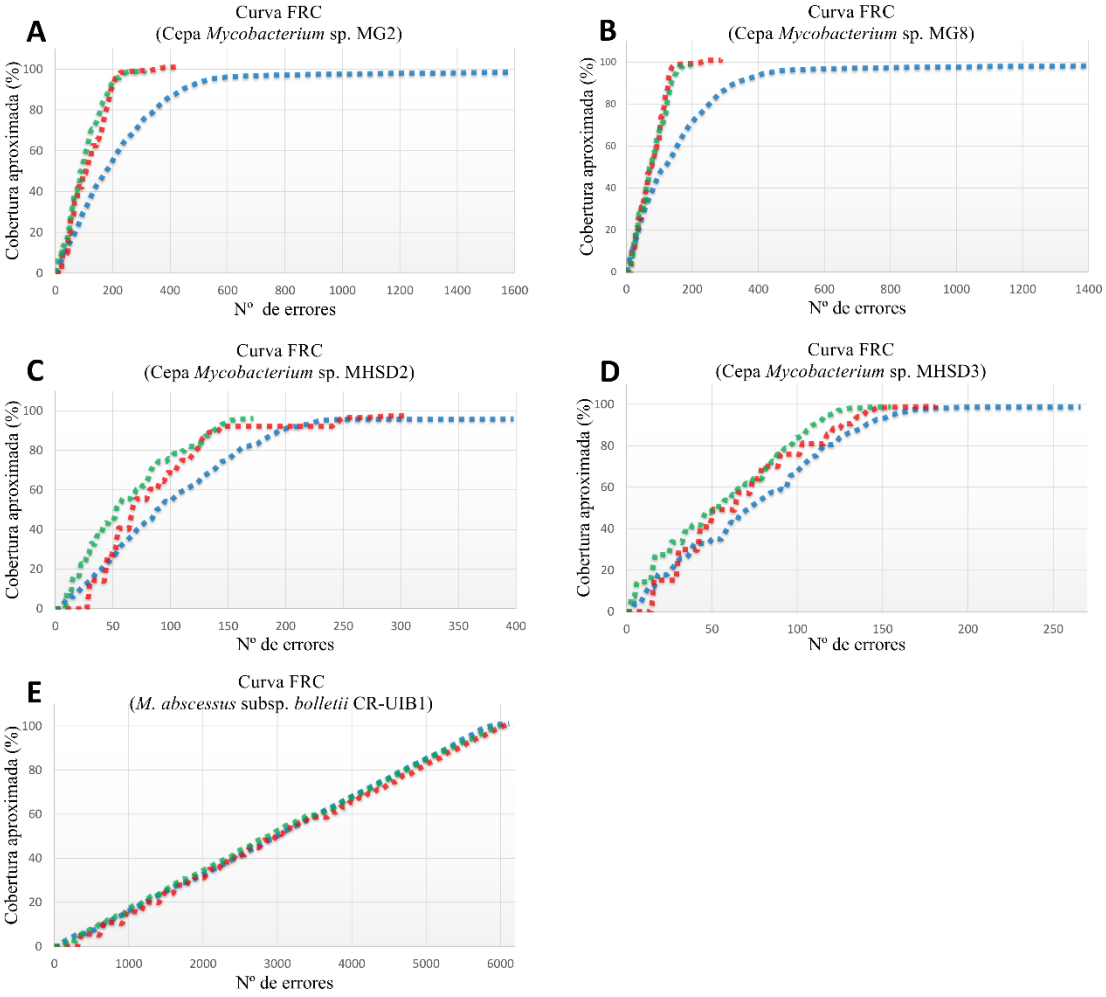


Figura 3.4. Curvas FRC de los asilamientos A) MG2, B) MG8, C) MHSD2, D) MHSD3 y E) CR-UIB1. Se representan los resultados para los ensamblajes obtenidos con Velvet (Verde), SPAdes (Rojo) y Newbler (Azul).

Después del análisis detallado de los datos y ensamblajes obtenidos con Velvet para cada uno de los genomas de las cepas de MCR (Tabla 3.7) se observó que los genomas en los

que se logró una mayor continuidad de secuencia corresponden a las cepas MHSD2 y MHSD3, destacado por los valores de N50 y N75 (tamaño del último *contig* con el que se consigue cubrir el 50 % ó 75 % respectivamente del tamaño estimado del genoma) y por los valores de L50 y L75 (nº de *contigs* necesarios para cubrir el 50 % y 75 % del tamaño estimado del genoma). Por su parte, y según estos mismos parámetros, el genoma en el que se obtuvo la menor continuidad fue en el aislamiento CR-UIB1, posiblemente debido a la cobertura aplicada, claramente inferior (25x frente a 50x con la que se ensamblaron el resto de genomas).

Tabla 3.7. Características de los mejores ensamblajes obtenidos para cada una de las cepas.

	MG2	**MG8**	**MHSD2**	**MHSD3**	**CR-UIB1**
Cobertura	Illumina 50x	Illumina 50x	Illumina 50x	Illumina 50x	Illumina 25x
Nº de *contigs*	152	123	65	70	225
***Contig* más largo**	178.136	202.360	426.852	361.417	139.377
Tamaño total	4.936.161	4.936.033	4.800.506	4.939.932	5.008.010
GC (%)	64,02	64,02	64,18	64,12	64,26
N50	72.383	74.693	129.773	141.658	40.937
N75	39.094	43.273	81.380	87.013	21.222
L50	21	23	10	10	39
L75	43	45	22	20	81
Ns por 100 kb	0	0	0	0	0

Como se puede observar en la Tabla 3.7, todos los genomas presentaron un contenido GC que osciló entre 64 y 65%, reflejándose el alto contenido GC característico, aunque no exclusivo, de las micobacterias.

En lo que respecta a la cepa *M. tuberculosis* CR-UIB2, tal y como se destacó anteriormente, el mejor ensamblaje según las comparativas realizadas mediante curvas FRC también fue el obtenido con Velvet (Figura 3.5A). En un intento de mejorar este ensamblaje, los 332 *contigs* resultantes del ensamblaje con Velvet se ordenaron con ayuda del programa ABACAS y utilizando el genoma completo de la cepa *M. tuberculosis* H37Rv (número de acceso NC_000962) como referencia. Después de rellenar los huecos resultantes y eliminar de las correspondientes indeterminaciones (Ns) se obtuvo un

ensamblaje con una considerable disminución en el contenido de errores (Figura 3.5B). La optimización de la continuidad del ensamblaje final se manifestó claramente en términos de un menor número de *contigs*, unos valores de N50 y N75 mayores y unos valores de L50 y L75 menores (Tabla 3.8).

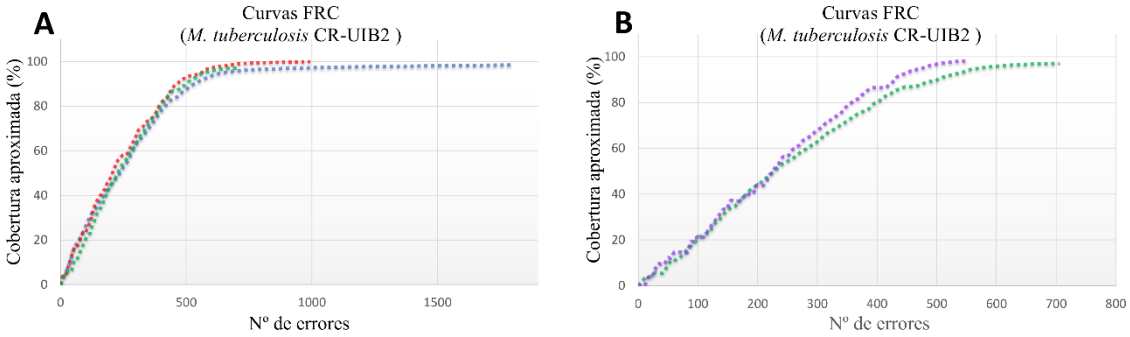


Figura 3.5. Curvas FRC de (A) los tres ensamblajes iniciales obtenidos con Velvet (Verde), SPAdes (Rojo) y Newbler (Azul) y (B) la comparación del contenido de errores entre el ensamblaje inicial de Velvet (Verde) y el mismo ensamblaje después de ser procesado con PAGIT y GapFiller (Morado).

Tabla 3.8. Resultados comparativos de cada una de las etapas del ensamblaje del aislamiento CR-UIB2.

	Velvet	**PAGIT**	**Sin Ns**
Nº de *contigs*	332	1	149
***Contig* más largo**	87.870	4.432.668	158.412
Tamaño total (pb)	4.269.887	4.432.668	4.317.131
GC (%)	65,31	65,35	65,35
N50	25.406	4.432.668	62.472
N75	15.123	4.432.668	32.048
L50	58	1	24
L75	113	1	49
N's por 100 kb	0	2.587	0

3.3.3. Anotación de los genomas con Prokka.

Las anotaciones obtenidas con Prokka v1.10 (Tabla 3.9) de los genomas que pertenecen al grupo MCR mostraron un rango que osciló entre los 4.691 (cepa MHSD2) a 6.453 (cepa tipo de *M. llatzerense*) CDS. En el caso de los ARNt, la cepa MHSD2 con 40 fue el genoma que menor número presentó, mientras que el genoma donde se encontró el mayor número fue *M. immunogenum* CCUG 47286^{T}, donde se detectaron hasta un total

de 61 ARNt. Por lo que respecta a los operones ribosomales, todos los genomas a excepción de *M. immunogenum* (en el que se detectaron 2), presentaron un único operón completo (ADNr 16S, 23S y 5S). Por su parte, en el genoma de la cepa *M. tuberculosis* CR-UIB2 se cuantificaron un total de 4.202 CDS, 50 ARNt y un operón ribosomal completo.

Tabla 3.9. Resultados de la anotación con Prokka v1.10 para los genomas secuenciados y ensamblados.

Cepa	Genes	CDS	ARNt	Operon ribosomal	tmRNA
M. chelonae CCUG 47445^T	4.940	4.888	48	1	1
M. immunogenum CCUG 47286^T	5.574	5.484	61	2	1
M. llatzerense MG13^T	6.510	6.453	53	1	2
M. abscessus subsp. *bolletii* CCUG 50184^T	4.956	4.932	51	1	1
MG2	4.887	4.838	45	1	1
MG8	4.903	4.855	44	1	1
MHSD2	4.735	4.691	40	1	1
MHSD3	4.909	4.862	43	1	1
CR-UIB1	5,033	4.983	46	1	1
CR-UIB2	4.297	4.202	50	1	1

3.4. Discusión

3.4.1. Extracción de ADN y secuenciación

Los problemas observados en términos de rendimiento y calidad de las lecturas de partida en los procesos iniciales de secuenciación llevados a cabo en la presente tesis son un fiel reflejo de la importancia de la obtención de ADN de alta calidad, pureza y en cantidad suficiente para la secuenciación de genomas con las nuevas tecnologías y plataformas de altas prestaciones. Este aspecto ha resultado clave para la obtención de buenos resultados en los procesos de secuenciación de nueva generación abordados en el proyecto. La presencia de impurezas, baja cantidad o calidad insuficiente (por ejemplo, ADN con un alto grado de degradación) son claramente responsables de un funcionamiento subóptimo o de la inhibición de los procesos en los que se basan estas plataformas de secuenciación. Consecuencia directa de ello, es la obtención de lecturas de mala calidad, bajos rendimientos o resultados poco consistentes.

A lo previamente comentado hay que añadir el hecho de que las micobacterias se catalogan como "bacterias difíciles de lisar" y que al romper sus paredes celulares generan muchos residuos que hay que eliminar. Por lo tanto, el proceso de extracción de ADN se complica en gran medida, ya que se requieren protocolos más agresivos a la vez que purificaciones más efectivas con el fin de obtener un ADN en cantidad suficiente y de alto peso molecular, dos premisas a veces difíciles de consensuar. Estos aspectos comprometen en gran medida los tres puntos clave destacados anteriormente. De esta forma, todos aquellos protocolos preestablecidos así como los kits comerciales utilizados de forma rutinaria en el laboratorio, altamente eficientes, por ejemplo, con microorganismos pertenecientes al género *Pseudomonas*, resultaron completamente ineficaces al ser utilizados con las cepas de micobacterias seleccionadas en este estudio.

Las búsquedas bibliográficas combinadas con las sucesivas pruebas de optimización llevadas a cabo en diferentes condiciones experimentales, permitieron esbozar y ajustar un protocolo eficaz para todas las cepas de micobacterias sobre las que se ha aplicado en el proyecto. Los puntos más críticos del protocolo diseñado consistieron en incluir un paso de rotura mecánica posterior a un pretratamiento enzimático con proteinasa K, y el paso de purificación y concentración final, en el que además se combinaron varias muestras de ADN obtenidas del mismo microorganismo, para conseguir alcanzar los mínimos de concentración y cantidad exigidas por los diferentes sistemas de generación de genotecas de las actuales plataformas de secuenciación masiva, especialmente exigente en este sentido PacBio.

La incorporación del tampón de lisis ATL (ver materiales y métodos), especialmente diseñado para lisis de tejidos, resultó de gran utilidad para conseguir una completa disgregación de los grumos celulares formados, en especial por algunas cepas como MHSD2 y CR-UIB1. Efectivamente, al realizarse la recolección de las células a partir de las colonias formadas en placa se creaban agregados de textura extremadamente seca y similares al aspecto de “migas de pan”, por otra parte, característico de *M. tuberculosis*, las cuales, al añadir la solución isotónica de Ringer para sedimentar las células, resultaban practicamente imposibles de disgregar. Como resultado se obtenían sedimentos celulares poco compactos, que al ser resuspendidos seguían teniendo el mismo aspecto anterior y constituían muestras difíciles de manejar y poco óptimas para proceder a la extracción de

sus ácidos nucleicos, tal como lo demostró el hecho de que a partir de estas muestras se obtenía muy poco ADN. La incorporación al protocolo del tampón de lisis ATL combinado con la actividad de la proteinasa K a 56 ºC, fueron la solución idónea para este punto crítico, ya que tras sendas modificaciones del protocolo los grumos se disgregaban casi sin esfuerzo, simplemente al ser sometidos a unos segundos de vórtex. La suspensión celular resultante presentaba un aspecto definitivamente homogéneo, bastante más adecuado para los sucesivos pasos de extracción de ADN total de las micobacterias.

La rotura mecánica con ayuda de microesferas de vidrio fue también clave para debilitar y conseguir la rotura final de la gruesa y resistente pared celular de las micobacterias, facilitando así mismo la penetración de los reactivos utilizados posteriormente, que ejercieron su acción de forma más eficaz. Sin embargo, durante este proceso es normal que el daño en las paredes celulares o especialmente la rotura de células provoque la liberación temprana de moléculas de ADN desprotegido al medio, sobre el cual también se provocan daños enzimáticos y mecánicos. La rotura del ADN se debe controlar al máximo posible para obtener un ADN genómico de alto peso molecular y poco degradado. En consecuencia, después de repetidas optimizaciones se determinó un tiempo óptimo de rotura mecánica de 5 minutos, necesario para inducir el suficiente daño en las envolturas celulares y mantener la máxima integridad posible del ADN obtenido.

3.4.2. Ensamblaje de genomas

3.4.2.1. Ensamblaje de cepas tipo

Como refleja el apartado de resultados, los genomas de las cepas tipo de *M. chelonae* y *M. immunogenum* se cerraron completamente, obteniendo cromosomas bacterianos de 5,0 y 5,5 Mb respectivamente, sin indeterminaciones y que fueron confirmados mediante mapeos, curvas FRC y el test con el programa REAPR. El cierre de estos genomas se realizó de manera casi inmediata con la simple combinación de lecturas Illumina y PacBio de alta calidad, lo cual parecería suficiente para el cerrado de genomas. Sin embargo, la misma estrategia aplicada en el genoma de la cepa tipo *M. llatzerense* resultó insuficiente para conseguir el mismo objetivo. Por ello, se consideró conveniente la inclusión de información complementaria que ayudase tanto en la ordenación como en la orientación

de los *contigs* resultantes. Con esta finalidad se generaron lecturas MP adicionales sobre la plataforma 454 GS FLX-Titatium, para incorporar información adicional que ayudase a construir nuevas relaciones de continuidad. La mala calidad de las primeras tandas de lecturas PacBio y de las lecturas MP fueron un punto clave negativo, ya que se seguía careciendo de información estructural fiable necesaria para abordar la resolución definitiva de un considerable número de puntos conflictivos del genoma. La inclusión de una nueva tanda de lecturas PacBio de mayor calidad mejoró en gran medida el ensamblaje, pero no permitió su cierre completo. Este ensamblaje seguramente necesita de un considerable esfuerzo de intervención manual en la resolución del rompecabezas que plantea; aunque también es posible la existencia de regiones del genoma mal cubiertas en términos de lecturas por las diferentes plataformas. Una segunda causa podría tener su origen en que la información concerniente a estas regiones en términos de lecturas es inexistente, consecuencia de que se encuentran pobremente representadas en el ADN total obtenido; hecho que podría explicar el que ninguna de las plataformas haya aportado esta información. La cepa tipo de *M. llatzerense* presenta además un genoma con gran cantidad de elementos repetitivos, como transposasas, integrasas o elementos de micobaceriofagos (datos no mostrados). Estos elementos lejos de facilitar la tarea de ensamblaje, contribuyen a la dificultad del proceso. Teniendo todo esto en cuenta, se aplicaron varias alternativas de ensamblaje y con diferentes programas. El mejor resultado obtenido, explicado previamente en el apartado de resultados, corresponde al de un tamaño de genoma de 6,2 Mb y que incluye hasta 6.453 CDS en su secuencia. En cualquier caso, se trata de un número claramente superior al de las regiones codificantes obtenidas en el resto de genomas secuenciados en el presente estudio.

En lo que se refiere a la cepa *M. abscessus* subps. *bolletii* CCUG 50184^{T}, se ha obtenido un ensamblaje de su genoma a nivel de *draft* de alta calidad. El ensamblaje consta de 26 *contigs* que alcanzan un tamaño total de genoma de 5.053.525 pb que codifican para 4.932 CDS. Estos resultados son concordantes con los que se pueden extraer de los genomas completos disponibles para esta especie en GenBank. No obstante, las estrategias de ensamblaje basadas en un genoma de referencia sólo son aplicables en casos donde se esté muy seguro de la proximidad de la especie del genoma en cuestión y siempre que existan genomas completos de esa misma especie en las bases de datos con los que se espera una elevada correlación en términos de sintenia. Aun así, la cautela debe ser

máxima ya que la existencia de reordenaciones cromosómicas en genomas de cepas de la misma especie puede llevar a ordenaciones erróneas de los *contigs*, generando por consiguiente ensamblajes que no se corresponden con la realidad. Por ello, es imprescindible llevar a cabo las oportunas confirmaciones de calidad del ensamblaje final independientes de un genoma de referencia; por ejemplo, partiendo de las lecturas PE y del ensamblaje final obtenido hacer la comprobación como se hizo con REAPR para saber si el ensamblaje obtenido respondía a la información que se podía obtener partiendo de las lecturas Illumina y calculando la acumulación de errores basada en curvas FRC.

3.4.2.2. Ensamblaje de los aislamientos del grupo MCR

Las diferencias que se observaron entre los ensamblajes obtenidos con cada uno de los tres programas utilizados rutinariamente fueron notables, esencialmente debido a los fundamentos en los que se basa cada programa. Por ejemplo, Newbler utiliza un enfoque basado en el solapamiento de lecturas y fue diseñado para trabajar con lecturas de pirosecuenciación obtenidas en la plataforma 454 GS FLX-Titatium. aunque puede utilizar datos de otras plataformas como lecturas PE, lecturas sin pareja o incluso lecturas obtenidas por la tecnología Sanger. Por su parte, Velvet es un programa especializado en lecturas Illumina, aunque también admite lecturas Sanger o 454 GS FLX-Titatium. Sin embargo, para la realización de un ensamblaje se basa, al igual que SPAdes, en los diagramas de *de Brujin*. Esta estrategia le permite tener una mayor capacidad de discriminación y resolución de zonas repetidas. Por su parte, SPAdes puede incorporar además lecturas de otras plataformas como Ion-torrent, 454 GS FLX-Titatium o incluso PacBio.

Los diferentes enfoques aplicados para realizar un ensamblaje y la habilidad para de utilizar un determinado tipo de lecturas, crean las diferencias entre los resultados obtenidos con Newbler, Velvet y SPAdes. La diferencia más evidente fue el contenido de errores en los ensamblajes, presentando un número muy superior los ensamblajes obtenidos con Newbler. Por otro lado, Velvet y SPAdes no mostraron una diferencia tan grande en cuanto a acumulación de errores con respecto a Newbler. En todos los casos, Velvet consiguió realizar un mejor trabajo en las mismas condiciones de cobertura con lecturas Illumina (50x). Por este motivo, los ensamblajes obtenidos con este programa

fueron los seleccionados para su utilización en los sucesivos pasos, por ejemplo, en el ordenado de *contigs* utilizando genomas de referencia.

Mención aparte merece el genoma del aislamiento CR-UIB1. Este genoma se secuenció utilizando una construcción de librerías diferente actualizada que permite la obtención de lecturas Illumina PE considerablemente más largas. La media de tamaño en las lecturas en este caso fue de 240 nucleótidos, habiendo lecturas de incluso más de 300. Sin embargo, se obtuvo un rendimiento en Mb mucho menor, no comparables con los resultados obtenidos para lecturas del mismo tipo en otras cepas. De hecho, el número de lecturas fue tan bajo que resultó imposible la obtención de un archivo de cobertura superior a 25x a pesar de la mayor longitud de las lecturas. Posiblemente, el bajo rendimiento se explicaría por una baja eficiencia en los pasos de alargamiento ejecutados por la polimerasa. Efectivamente, debido a la naturaleza de la química y el proceso de amplificación implicados, un ADN que no satisface los mayores requerimientos de pureza y tamaño (ADN menos degradado) para generar lecturas más largas, puede afectar a la capacidad de secuenciación por síntesis basada en la acción enzimática de la polimerasa en la plataforma Illumina. Así pues, esta menor cobertura está en el origen del hecho de que el ensamblaje de la cepa CR-UIB1 no pudiera mejorarse más allá de los 225 *contigs* finalmente obtenidos.

3.4.2.3. Ensamblaje de *Mycobacterium tuberculosis* CR-UIB2.

El proceso bioinformático de ensamblaje del genoma del aislamiento de *M. tuberculosis* siguió el mismo esquema que en el caso de los aislamientos del grupo MCR mencionados anteriormente. Nuevamente, Velvet, en condiciones de cobertura con lecturas Illumina (50x) iguales, fue el que dio lugar al mejor ensamblaje según los criterios de calidad aplicados habitualmente. Esta cepa se identificó fenotípicamente tras su aislamiento en la Clínica Rotger de Palma de Mallorca, como *M. tuberculosis.* En base a ello y tras elegir el mejor ensamblaje, se decidió intentar mejorar el mismo utilizando como genoma de referencia el correspondiente a la cepa tipo de *M. tuberculosis* H37Rv. El resultado preliminar obtenido con el programa ABACAS consiguió ordenar todos los *contigs* en un único *scaffold*, además de una buena correlación en el alineamiento entre la cepa CR-UIB2 y el genoma correspondiente a la cepa tipo. Un considerable número de

indeterminaciones se corrigieron mediante el rellenado de huecos utilizando el conjunto de todas las lecturas Illumina de alta calidad obtenidas. La eliminación de las indeterminaciones restantes después de todo el proceso resultó en un ensamblaje considerablemente mejorado no solo en continuidad, presentando una longitud del *contig* más largo de 158.412 pb respecto a las 87.870 pb iniciales, y un L50 = 24, con respecto a los 58 *contigs* iniciales, tal como se puede comprobar en los valores que hacen referencia a este parámetro (Tabla 3.8), sino también en un menor contenido de errores acumulados (Figura 3.5).

4. Capítulo 2: Determinación y análisis del genoma esencial y pangenoma del grupo de micobacterias de crecimiento rápido

4.1. Introducción

Se define como el genoma esencial de una especie a todos aquellos genes o proteínas compartidas por todos los genomas considerados en un estudio, mientras que el pangenoma representa el conjunto de todas las familias génicas o proteicas diferentes presentes en los mismos [77]. Así pues, el pangenoma, además del genoma esencial, incluiría todos aquellos elementos accesorios que pueden ser exclusivos de un genoma o compartido por algunos de ellos y que le dotarían de diferencias metabólicas clave para la ocupación de un determinado nicho ecológico o para la realización de una función concreta. A este conjunto de genes o proteínas se le conoce como genoma accesorio [2]. Un pangenoma puede ser abierto o cerrado en función de la evolución en su cálculo a medida que se incorporan más genomas al estudio. Por definición, un pangenoma "abierto" es aquel en el que la incorporación de nuevos genomas de interés produce la incorporación de nuevos genes no presentes en los ya incluidos en el estudio, mientras que un pangenoma "cerrado" es aquel en el que la incorporación de nuevos genomas en el estudio apenas varía el tamaño del mismo [78]. Por ejemplo, en el caso de *Streptococcus agalactiae* el contenido medio de genes por genoma es de aproximadamente 2.000, mientras que un estudio de pangenoma realizado sobre esta especie determinó un tamaño del mismo en torno a 6.000 familias génicas [77], tres veces superior, representando un claro ejemplo de pangenoma abierto. Presentar un pangenoma abierto o cerrado proporciona una idea aproximada de la capacidad adaptativa de una especie, o la diversidad funcional de un conjunto de especies.

Cuando se realizan estudios de pangenoma, la comparación de genomas o proteomas permite la identificación de lo que se conocen como ortólogos, genes o proteínas derivadas de un ancestro común y que están presentes en distintas cepas, mientras que los genes parálogos son aquellos que proceden de fenómenos de duplicación génica, en los cuales, la nueva copia puede evolucionar de forma independiente hasta el punto de poder aportar una nueva función a la célula, siendo este fenómeno uno de los principales motores impulsores de la evolución microbiana [2]. Así pues, el genoma esencial estaría compuesto por genes ortólogos, mientras que el pangenoma, por definición, incluiría

tanto genes ortólogos como parálogos, cubriendo todo el espectro de funciones distintas presentes en el conjunto de genomas estudiados.

Partiendo de los estudios de pangenómica es posible determinar grupos de proteínas exclusivas de un genoma o un conjunto de ellos una vez los proteomas implicados han sido comparados entre ellos. A grandes rasgos, es interesante determinar qué funciones pueden llevar a cabo ese subconjunto de elementos para, de alguna manera, relacionar qué necesidades funcionales ha desarrollado, por ejemplo, una determinada especie para ocupar un determinado nicho ecológico. Esto es especialmente importante cuando se trata de proteínas hipotéticas. En este sentido, las bases de datos de agrupaciones de genes ortólogos o COG (del inglés *Clusters of Orthologous Genes*) son de gran utilidad. La base de todo esto es que todas aquellas proteínas incluidas en una de estas agrupaciones presumiblemente desempeñan la misma función [79], permitiendo realizar una clasificación funcional de un conjunto de proteínas o genes concretos al ser contrastados con estas bases de datos. De esta forma se pueden relacionar, de alguna manera, las implicaciones ecológicas que se derivan de la presencia de ese conjunto de elementos exclusivos en concreto, es decir, cómo se han visto obligadas a reaccionar las bacterias al ser sometidas a determinadas presiones ambientales.

En el presente trabajo se propone la realización de estudios de genoma esencial y pangenoma desde un nivel más general, considerando diversas especies, a un nivel más concreto, considerando especies estrechamente relacionadas o una única especie. Por lo que se trata de un estudio comparativo a tres niveles bien diferenciados:

- Genoma esencial y pangenoma de especies diferentes (grupo MCR)
- Genoma esencial y pangenoma de especies muy cercanas entre sí (grupo *abscessus-chelonae-immunogenum*)
- Genoma esencial y pangenoma de diferentes cepas de la misma especie (*M. immunogenum* y *M. tuberculosis*)

4.2. Materiales y métodos

4.2.1. Genomas obtenidos de la base de de datos GenBank

Los genomas de especies del grupo MCR disponibles en GenBank [76] fueron obtenidos en formato FASTA utilizando el protocolo descrito en NCBI para la descarga de múltiples genomas a partir del servidor FTP que almacena dicha base de datos (http://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq). En caso de genomas con plásmidos incluidos, su secuencia fue descargada conjuntamente a la del correspondiente genoma de forma automática a través del mismo procedimiento e incluida en el análisis.

4.2.2. Contextualización evolutiva de las especies consideradas

Partiendo de las secuencias nucleotídicas de todos los genomas (Tablas suplementarias 1, 2 y 3 del Anexo 1) se extrajo la secuencia del ADNr 16S por procedimientos descritos previamente [80]. Estas secuencias fueron alineadas con ClustalW [81] y la comparación filogenética se realizó construyendo un árbol con el programa PhyML [82], utilizando una estimación por máxima verosimilitud o MLE (del inglés *Maximum Likelihood Estimation*) para la construcción del mismo y aplicando un bootstrap 100 como soporte estadístico para las ramas. En este caso se incluyeron, como guía, las secuencias de ADNr 16S de las cepas tipo de todas las especies incluidas en este estudio. Estas últimas se obtuvieron a través de la base de datos *StrainInfo* (http://www.straininfo.net).

4.2.3. Determinación del genoma esencial y pangenoma

Con la finalidad de normalizar la anotación de todos los genomas, incluyendo aquellos ya anotados obtenidos a partir de GenBank, se realizó con el programa Prokka v1.10, tal y como se explicó previamente para los genomas propios en el capítulo anterior. Los archivos en formato FASTA resultantes del proceso de anotación, y que contenían el proteoma (secuencias proteicas de todos los ORF predichos) para cada genoma fueron utilizados para su comparación. En términos de homología, dos proteínas pueden ser consideradas como probables homólogas si presentan, como mínimo, un 30% de identidad en el 100 % de la longitud de las secuencias comparadas, aunque pueden definirse proteínas homólogas con porcentajes de identidad menores, basándose en términos de E-value y bit-score [83]. En base a esta información, en el presente caso la

comparativa se realizó mediante BLASTP (del inglés *Basic Local Alignment Search Tool for Proteins*), del programa BLAST (del inglés *Basic Local Alignment Search Tool*) [84], de todos los proteomas contra todos siguiendo la regla 50/50 para especies diferentes definida en trabajos previos. Según esa regla, para que dos proteínas fueran agrupadas conjuntamente debían mostrar, como mínimo, un 50 % de identidad en, por lo menos, un 50 % de la longitud de la secuencia más larga entre las que se realiza esta comparación [80] (el equivalente a un 25 % de identidad en el 100 % de la secuencia). En el caso de las comparativas realizadas entre cepas de la misma especie se aplicó un criterio más restrictivo, considerando en este caso un 70 % de identidad en por lo menos un 50 % de la secuencia más larga. Los resultados obtenidos se utilizaron para la generación de matrices de BLAST basadas en las identidades obtenidas, empleadas para la estimación del número de proteínas compartidas entre genomas.

Para identificar y colectar secuencias de proteínas ortólogas compartidas entre los genomas anotados se aplicaron tres algoritmos de agrupación distintos: BBDH (del inglés *Bi-Directional Best Hits*) [85] COGtriangle (COGT, del inglés *Cluster of Orthologous Genes*) [86] y OrthoMCL (OMCL, del inglés *Orthologous Markov Cluster*) [87], tal y como se describe en el programa *Get_homologues* [88]. Con estos tres algoritmos se pudo estimar el número de familias proteicas compartidas por todos los genomas (genoma esencial) así como el número de familias diferentes acumuladas entre todos los genomas de interés (pangenoma). Las representaciones teóricas de las curvas de genoma esencial y pangenoma fueron calculadas y ajustadas a modelos exponenciales de acuerdo con Tetellin [77] y Willenbrock [89] para la estimación del tamaño y proyección de los mismos.

El número total de familias diferentes detectadas que conforman el pangenoma se obtuvo de la intersección (consenso) de los algoritmos COGT y OMCL. Las diferentes familias de proteínas del pangenoma fueron clasificadas en función del grado de conservación y distribución en los genomas estudiados, dando lugar a cuatro grandes agrupaciones: *genoma esencial* (presentes estrictamente en todos los genomas), *Soft*-genoma esencial o genoma esencial laxo (presentes en un número igual o superior al 95 % de los genomas considerados), *Shell* (presentes en varios genomas, pero por debajo del 95 %) y *Cloud* (presente en solo unos pocos genomas) [90]. Por definición, el número de familias que se

reflejan en el genoma esencial laxo incluye las agrupadas en el genoma esencial estricto. La determinación del número de genomas en los que debe estar presente una familia de proteínas para ser clasificado en cada una de las tres últimas categorías, depende del número de genomas introducidos para el cálculo y es determinado de forma automática [88]. Por último, para la obtención del árbol representativo del pangenoma se partió de una matriz generada a partir de la presencia/ausencia de proteínas en un determinado genoma.

Por su parte, la determinación del genoma esencial formado por genes que se encuentran en copia única en todos los genomas se obtuvo de la intersección o consenso de los algoritmos BDBH, OMCL y COG.

4.2.4. Análisis del genoma esencial "estricto" monocopia

El conjunto de proteínas codificadas por genes que se hallan en copia única en un determinado genoma fueron concatenados, obteniendo una macro secuencia aminoacídica para cada uno de los microorganismos. Los concatenados se alinearon con Clustal Omega (ClustalO) [91]. El alineamiento múltiple obtenido fue procesado con GBLOCKS [92] para la extracción y concatenación de los bloques de posiciones homólogas. El árbol evolutivo a partir del alineamiento resultante se obtuvo con PhyML y aplicando el algoritmo aLRT (del inglés *Approximate Likelihood-Ratio Test*) [93] para el soporte estadístico de las ramas.

4.2.5. Caracterización de las proteínas específicas aportadas por cada cepa o grupo de cepas al pangenoma

Utilizando el programa *Get_Homologues*, se determinó el número de genes o proteínas exclusivas de una cepa o conjunto de cepas con respecto al total de genomas utilizados en el estudio, determinando así su aportación específica al pangenoma. A partir de un archivo con dichas secuencias proteicas en formato FASTA se realizaron análisis de BLASTP contra una base de datos COGs (del inglés *Clusters of Orthologous Genes*) de proteínas recientemente actualizada [94]. Esta asignación a un COG determinado se realizó con la ayuda del programa RPS-BLAST v2.2.26+ (del inglés *Reverse Position Specific BLAST*) (http://nebc.nerc.ac.uk/bioinformatics/docs/rpsblast.html), asignando por este procedimiento la posible función atendiendo a la categoría funcional en la que el

COG reportado se incluye en cada caso. Los resultados de los BLASTP se consideraron significativamente informativos a partir de un E-value $\geq 10^{-5}$. En el caso de que una misma proteína fuese asignada a diferentes COGs, los resultados con mejores valores estadísticos (E-value menor), fueron elegidos, eliminando los demás resultados y evitando de esta manera duplicidades en el estudio. Los códigos de cada categoría funcional, así como la función específica que representan, se encuentran en la Tabla suplementaria 4 del Anexo 1.

4.3. Resultados

4.3.1. Genoma esencial y pangenoma del grupo MCR

Un Total de 52 genomas representativos de 17 especies de MCR, incluidos los genomas secuenciados *de novo* y los obtenidos de la base de datos GenBank, fueron utilizados para realizar el estudio comparativo. Estos genomas incluyen todos los genomas de MCR disponibles a día 1 de septiembre de 2015. Debido al gran número de genomas disponibles de las especies *M. abscessus* y *M. abscessus* subsp. *bolletii*, con el fin de no sobrerrepresentar dichas especies, sólo se utilizaron genomas representativos para el estudio del genoma esencial y el pangenoma del grupo de MCR, priorizando para genomas de cepas tipo y genomas completamente cerrados.

El genoma esencial obtenido estabilizó su descenso por debajo de 2.000 familias proteicas. A partir de la incorporación de 10 genomas los cambios no fueron significativos, mostrando un descenso moderado que esboza perfectamente el efecto meseta. Por su parte, el pangenoma mostró una tendencia ascendente con una pendiente muy pronunciada, ajustándose prácticamente a un crecimiento lineal, aunque el ajuste a la curva mostró una ligera tendencia a la saturación (Figura 4.1). El tamaño del pangenoma fue de 21.056 grupos de proteínas (Figura 4.2A) según el consenso obtenido entre los algoritmos COG y OMCL. La clasificación de los grupos en las distintas categorías congregó a 1.253 familias protecas en el genoma esencial, 1.645 en el denominado genoma esencial laxo (incluyendo el genoma esencial estricto), 5.646 en el *Shell* y 13.765 grupos en el *Cloud* (Figura 4.2B). Estas tres últimas categorías corresponderían al llamado genoma accesorio (presente en diversos genomas, pero no en todos) o prescindible. Esto implica que el genoma esencial estricto representaría solo el 5,95 % del pangenoma, frente a un 65,37 % de las proteínas detectadas que están

presentes en tan solo 1 ó 2 genomas (*Cloud*), mientras que el 28,67 % restante (parte del genoma esencial laxo y el *Shell*) representarían a aquellas proteínas que están presentes en varios genomas (entre 2 y 52). Teniendo en cuenta que el número de CDS en los genomas estudiados es de 5.544 ± 719, el genoma esencial representaría entre un 17,5 % y un 27,4 % de las proteínas codificadas por un genoma concreto.

4.3.1.1. Establecimiento de relaciones evolutivas basadas en el genoma esencial monocopia y pangenoma del grupo MCR

Con el propósito de presentar la información en el contexto evolutivo adecuado se construyó un árbol filogenético basado en las secuencias de ADNr 16S extraídas de los 52 genomas de MCR contemplados en el estudio, incluyendo además las secuencias de ADNr 16S para las cepas tipo de las 17 especies contempladas, obtenidas a través de la base de datos *StrainInfo* como control interno (Figura 4.3). De esta forma se obtuvo el marco evolutivo adecuado para la presentación definitiva de los datos obtenidos de la comparativa de los proteomas de los genomas utilizados

Observando las agrupaciones obtenidas, se puede percibir cómo prácticamente todos los ADNr 16S extraídos de los distintos genomas se agruparon en torno al ADNr 16S de la cepa tipo correspondiente (utilizados como marcadores internos). Sin embargo, las cepas *M. smegmatis* JS623, *M. rhodesiae* NBB3, *M. fortuitum* Z58 y *M. chelonae* 1518 se posicionaron de forma incoherente, alejadas de las agrupaciones donde deberían haberse incluido de acuerdo con las secuencias control de dichas especies.

A partir de las 1.005 familias de proteínas constituyentes del genoma esencial de genes en copia única del grupo de MCR estudiado (Figura 4.2C), una vez concatenadas para cada genoma y alineadas entre ellos, se obtuvieron un total de 343.312 posiciones. Después de eliminar los bloques divergentes, ambiguos o pobremente alineados; se obtuvieron 297.091 posiciones (86,53 % del alineamiento original) correspondientes a bloques conservados y, por lo tanto, posiciones homólogas. Sobre este alineamiento se construyó el árbol evolutivo para los 52 genomas que se representa en la Figura 4.4A. Por su parte, a partir de la matriz de presencia/ausencia, donde se contempla la distribución de las 21.056 familias de proteínas en los distintos genomas, se construyó el árbol donde se refleja la información del pangenoma resultante (Figura 4.4B).

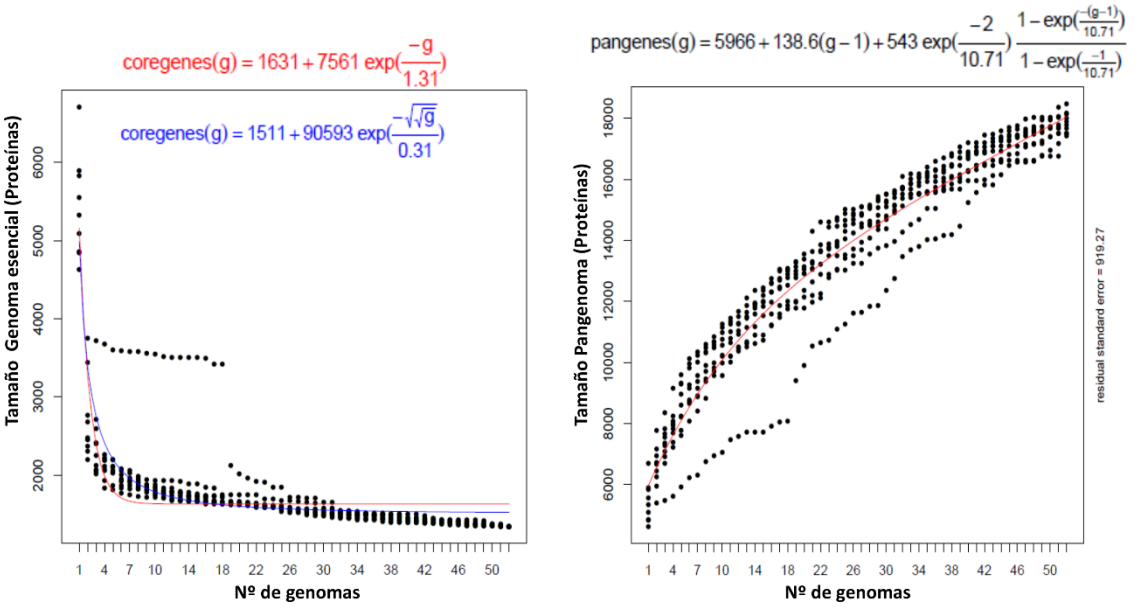


Figura 4.1. Curvas representativas de la proyección del genoma esencial y pangenoma del grupo MCR.

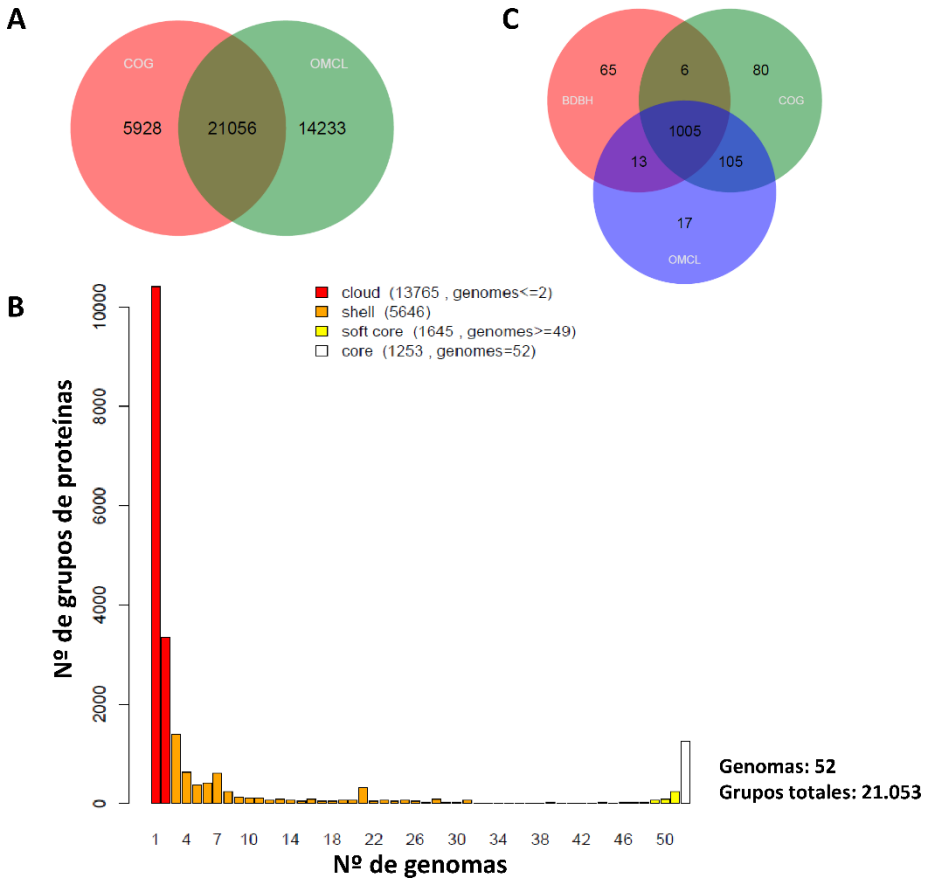


Figura 4.2. Número de grupos que conforman el A) pangenoma, B) clasificación de las distintas familias proteicas del pangenoma en genoma esencial, genoma esencial laxo, *Shell* y *Cloud* y C) genoma esencial monocopia. El número de familias proteicas del genoma esencial laxo representado en la leyenda resulta de la suma de grupos presentes en el 95 % de los genomas y el genoma esencial estricto.

.

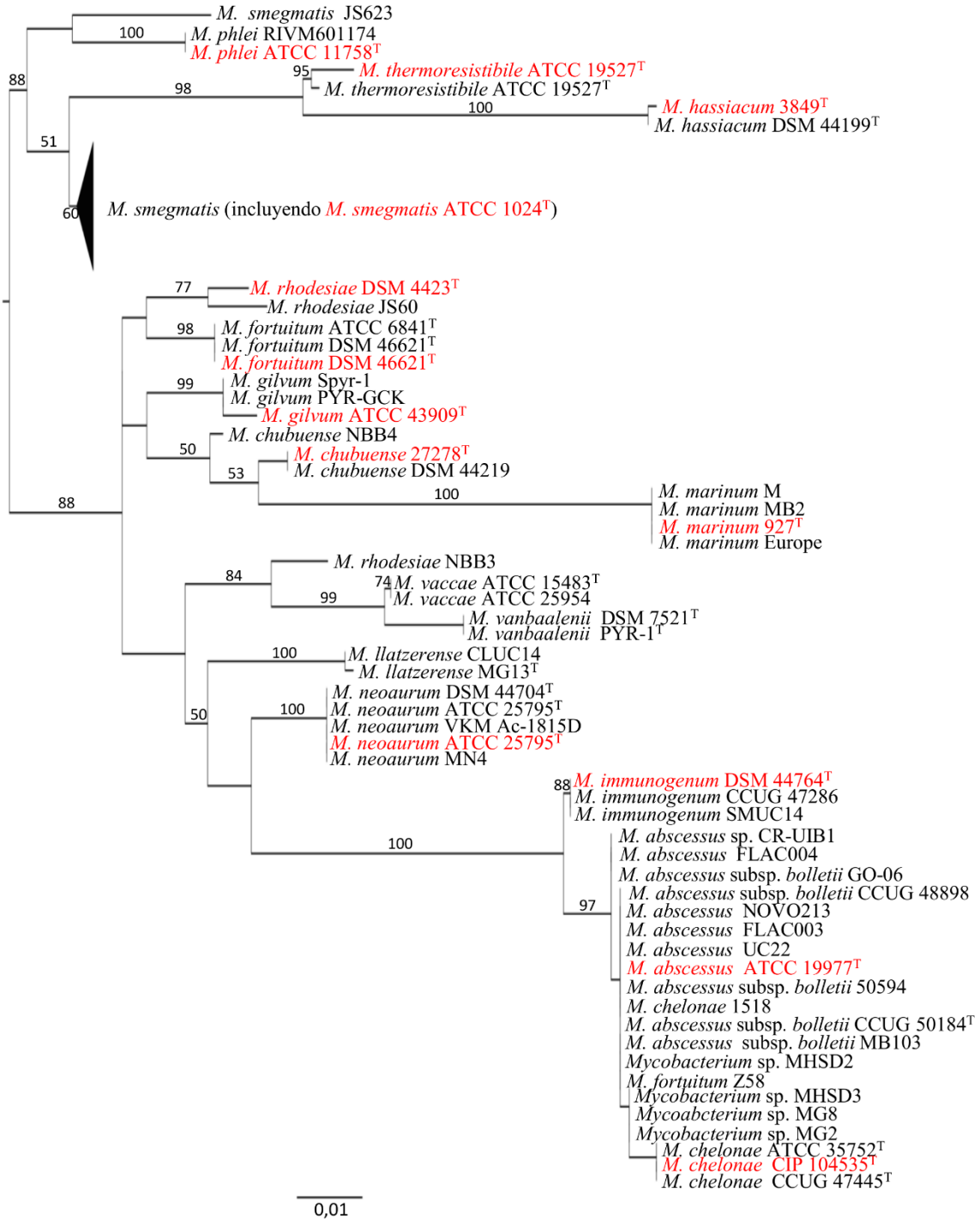


Figura 4.3. Árbol filogenético del grupo MCR basado en la secuencia del ADNr 16S. Las secuencias destacadas en rojo corresponden a las obtenidas en GenBank después de la identificación de la cepa tipo de las distintas especies a través de Strainfo.

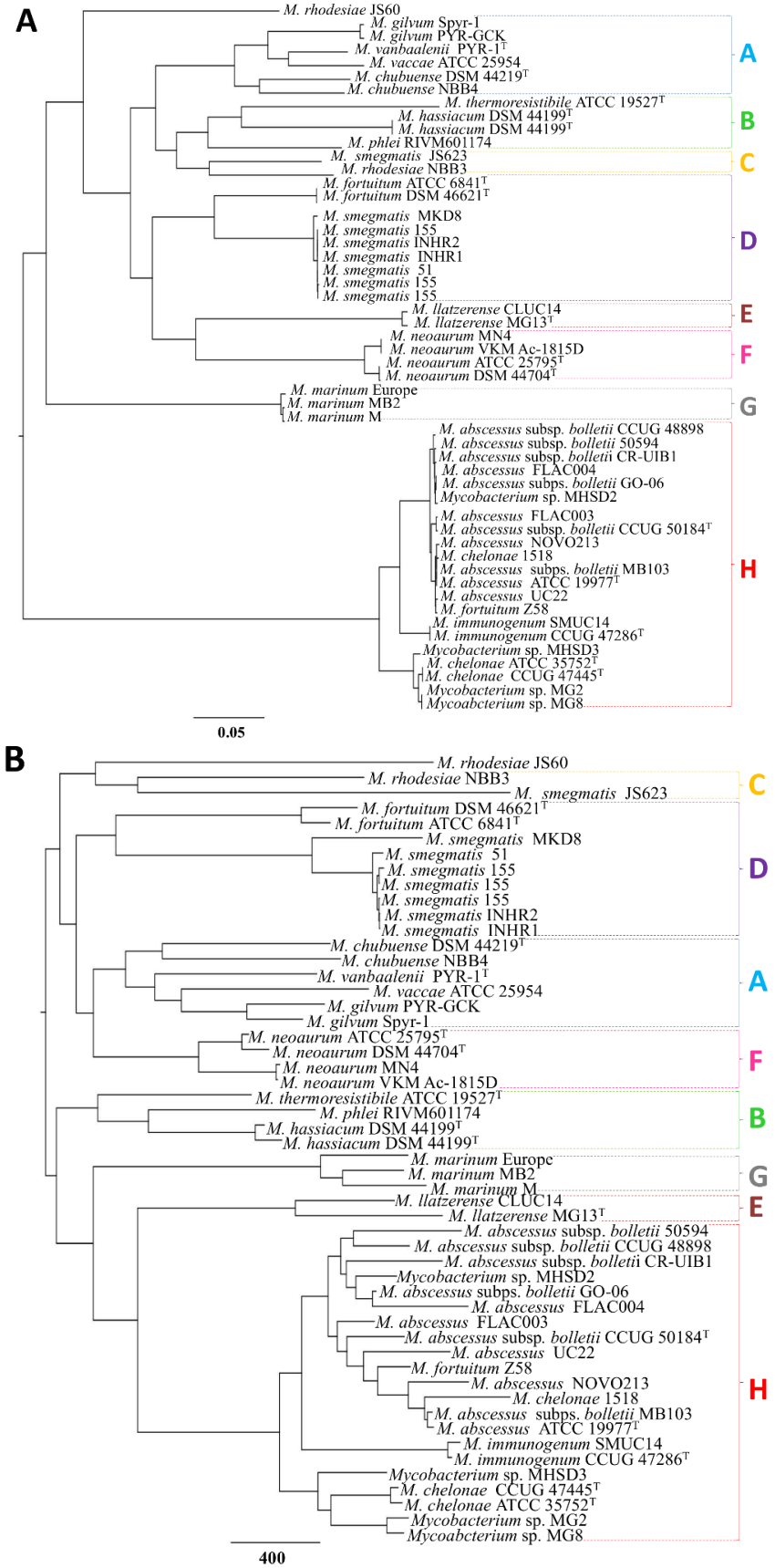


Figura 4.4. Representación del árbol basado en el genoma esencial monocopia (A) y del dendograma basado en la matriz de presencia/ausencia del pangenoma (B).

En ambas representaciones se pueden observar hasta 8 grupos completamente estables (A, B, C, D, E, F, G y H), en los que coinciden las mismas cepas. Dichos grupos varían su posición en el dendograma del pangenoma con respecto al árbol del genoma esencial debido a que definen su posición en torno a aquellas especies más similares funcionalmente según las proteínas presentes o ausentes determinadas en la matriz del pangenoma. Los genomas *M. rhodesiae* NBB3, *M. smegmatis* JS623, *M chelonae* 1518 y *M. fortuitum* Z58 volvieron a aparecer separados del resto de cepas de la especie correspondiente en ambas representaciones, tal y como ocurría con el análisis preliminar con el ADNr 16S. En el grupo *abscessus-chelonae-immunogenum*, las dos últimas especies quedaron perfectamente diferenciadas. Sin embargo, la separación entre las especies *M. abscessus* y *M. abscessus* subsp. *bolletii* no quedó bien delimitada. Efectivamente, tal como se observa en la Figura 4.4, se formaron dos grandes ramas a partir de su separación de *M. immunogenum*, dentro de las cuales se englobaron indistintamente genomas de ambas especies.

La determinación de las proteínas diferenciales que definen la separación en una rama independiente de las distintas especies en el dendograma del pangenoma, se realizó teniendo en cuenta todas aquellas proteínas comunes entre las distintas cepas de una misma especie y que no están presentes en el resto de genomas (Tabla 4.1). Para ello no se tuvieron en cuenta aquellas cepas supuestamente anotadas como una especie determinada pero que, según el estudio de ADNr 16S y el genoma esencial monocopia, quedaron desplazadas de su ubicación esperada.

En la Tabla 4.1 también se recoge el número de COGs diferentes detectados, con un E-value igual o inferior a 10^{-5}, en los grupos de familias proteicas exclusivas de cada especie, reflejando la variedad funcional para la separación de cada especie en la representación del dendograma del pangenoma. Cabe destacar que esta determinación solo atañe a los elementos comunes a todos los genomas de las cepas utilizadas de una especie determinada y no de aquellos elementos específicos de cada cepa. Además, la dificultad para definir la separación entre las cepas pertenecientes a las especies *M. abscessus* y *M. abscessus* subsp. *bolletii* llevó a considerarlas en este caso como un único grupo compuesto, el cual fue posteriormente estudiado más detalladamente

Tabla 4.1. Número de familias proteicas exclusivas aportadas por especie y número de familias clasificadas funcionalmente en base a los COGs.

Especie	Nº cepas	Nº grupos	Nº COGs	% COGs
M. abscessus-bolletii	14	18	7	38,9
M. chelonae	5	64	17	26,6
M. chubuense	2	35	13	37,1
M. fortuitum	2	490	97	19,8
M. gilvum	2	94	14	14,9
M. hassiacum	2	297	48	16,2
M. immunogenum	2	333	47	14,1
M. llatzerense	2	360	66	18,3
M. marinum	3	641	146	22,8
M. neoaurum	4	160	52	32,5
M. phlei	1	410	60	14,6
M. rhodesiae	1	1.071	145	13,5
M. smegmatis	7	441	156	35,4
M. thermoresistibile	1	350	58	16,6
M. vaccae	1	449	113	25,2
M. vanbaleenii	1	419	79	18,9

El grupo *M. abscessus-M. abscessus* subsp. *bolletii* fue el que menor número de proteínas diferenciales presentó (18 proteínas) con respecto a las otras especies, además de ser el que presentó una menor variedad en cuanto a COGs identificados (7 COGs diferentes entre las proteínas clasificadas). Por su parte, la especie *M. rhodesiae*, representada en este caso únicamente por la cepa NBB3, presentó 1.071 proteínas exclusivas, el mayor número con diferencia. Sin embargo, a pesar de este elevado número, no fue la especie donde se observó la mayor variedad de COGs, ya que en *M. smegmatis* se obtuvieron 135 COGs a partir de 441 proteínas exclusivas del conjunto de 7 cepas que conformaron esta especie en el estudio.

La asignación de cada uno de estos COGs a su categoría funcional correspondiente mostraría el perfil funcional, definido por las proteínas específicas, de cada especie (Figura 4.5). En este sentido, se observó que la variedad funcional asociada a la

especiación del grupo MCR fue elevada. El grupo *M. abscesus-M. abscessus* subsp. *bolletii* fue la que menos proteínas diferenciales presentaba y por lo tanto una de las que presentó menos categorías funcionales. En *M. smegmatis*, por ejemplo, se observó cómo una de las categorías funcionales predominantes en cuanto a proteínas exclusivas es la de "metabolismo y transporte de carbohidratos" (G) con 28 proteínas implicadas en estos procesos, seguido de *M. marinum* (12 proteínas) y *M. vaccae* (11 proteínas). Otra categoría, la correspondiente a "transcripción" (K), destacó en las especies *M. fortuitum*, *M llatzerense* y *M. smegmatis* con 11 proteínas representantes en la primera especie y 14 en cada una de las dos últimas especies. La categoría funcional relativa a "biogénesis de la pared celular/membrana/envoltura" (M) destacó en *M. rhodesiae* (11 proteínas) y *M. vanbaleenii* (7 proteínas). Por su parte, en *M. rhodesiae*, *M. marinum* y *M. smegmatis* se catalogaron 12, 13 y 13 proteínas exclusivas respectivamente, implicadas en la "producción y conversión de energía" (C). En estas mismas especies, junto con *M. vaccae* y *M. vanbaleenii* destacaron también el número de proteínas asociadas al "transporte y metabolismo de lípidos" (I) y "predicción de función general" (R). En el caso de *M. marinum* destacaron 14 proteínas asociadas a COGs incluidos en la categoría funcional "función desconocida". En líneas generales, todas las especies presentan un buen número de proteínas exclusivas de especie asociadas a diferentes categorías funcionales.

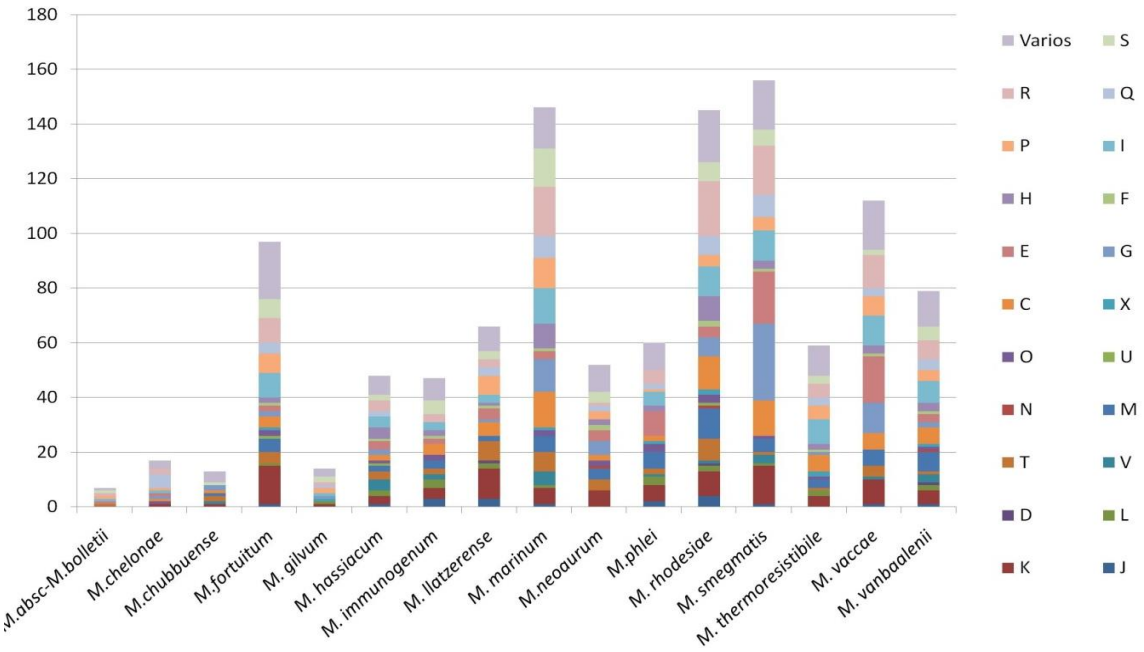


Figura 4.5. Clasificación por categorías funcionales [94] de las proteínas exclusivas del conjunto de genomas de especie analizados en el presente estudio. Se indica el número hallado para cada categoría.

4.3.2. Genoma esencial y pangenoma del grupo *abscessus-chelonae-immunogenum*

De los 263 genomas disponibles, se utilizaron 24 genomas en principio pertenecientes a la especie *M. abscessus* y 24 a la especie *M. abscessus* subsp. *bolletii*, así como los genomas disponibles de *M. chelonae, M. immunogenum* y además de los correspondientes a los de las cepas propias: MHSD2, MHSD3, MG2, MG8 y CR-UIB1 Entre los genomas utilizados se incluyen los correspondientes a *M. abscessus* y *M. abscessus* subsp. *bolletii* ya utilizados en el estudio de genomas esencial y pangenoma del grupo MCR. (Tabla suplementaria 2 del Anexo 1)

Así, para este estudio se utilizaron hasta un total de 66 genomas, sin tener en cuenta el genoma de *M. chelonae* 1518 de acuerdo con los resultados para este genoma obtenidos y descritos en el apartado anterior. La curva del genoma esencial obtenida tal como se observa en la Figura 4.6, estabilizó su descenso por debajo de los 3.000 grupos de proteínas, mostrando una tendencia a no disminuir de forma significativa mediante la incorporación de más genomas. Por su parte, la estimación del pangenoma mostró un tamaño aproximado de 10.000 familias de proteínas diferentes, ajustándose a una tendencia de crecimiento lineal (Figura 4.6).

El tamaño total del pangenoma, determinado a partir de la intersección de los algoritmos COGT y OMCL y considerando las condiciones C50/S50, se estableció en 11.561 proteínas (Figura 4.7A), con una media de 175,2 proteínas nuevas por genoma incorporado. De estas, 2.978 se engloban en el genoma esencial estricto, 3.463 en el genoma esencial laxo (incluyendo el genoma esencial), 3.428 en el *Shell* y 4.670 en el *Cloud* (Figura 4.7B). Considerando estos números, se obtuvo un genoma esencial estricto y un *Cloud* que representan el 25,8 % y 40,4 % del total respectivamente. El 33,8 % restante estarían representados por las 485 familias restantes del genoma esencial laxo (altamente conservadas entre los distintos genomas, pero no en todos) y el *Shell*, distribuidas en un número variable de genomas, pero no tan conservados como los presentes en el genoma esencial laxo ni tan exclusivos como los presentes en el *Cloud* (en este caso >2 y <62). Por su parte, el genoma esencial monocopia presentó 2.650 proteínas diferentes codificadas por genes en copia única en el conjunto de los 66 genomas utilizados (Figura 4.7C).

4.3.2.1. Relaciones basadas en el genoma esencial y pangenoma del grupo *abscessus-chelonae-immunogenum*

El alineamiento del concatenado de las 2.650 proteínas del genoma esencial monocopia dio como resultado 835.554 posiciones, de las cuales GBLOCKS extrajo 810.212 posiciones conservadas (96 % del total). En base a este alineamiento se calculó el árbol del genoma esencial para las 66 cepas utilizadas (Figura 4.8). Como resultado se obtuvieron tres grandes ramas o grupos: (A) *M. abscessus* y *M. abscessus* subsp. *bolletii*, (B) *M. immunogenum* y (C) *M. chelonae*. Dentro del grupo A apareció una primera rama que incluye a las dos cepas tipo de *M. abscessus* subsp. *bolletii* utilizadas, BD y CCUG 50184 (A_1) y una segunda rama a su vez subdividida en dos grandes grupos. Un primer subgrupo formado principalmente por genomas de *M. abscessus*, entre ellos el de la cepa tipo ATCC 19977 (A_2), y un segundo subgrupo formado principalmente por genomas de *M. abscessus* subsp. *bolletii* (A_3).

Por su parte, en el dendograma del pangenoma basado en la presencia/ausencia de hasta 11.561 proteínas, se reprodujeron estas mismas agrupaciones de genomas; aunque con una organización diferente, especialmente en la rama que separa las especies *M. abscessus* y *M abscessus* subsp. *bolletii* del resto. Sólo las cepas *M abscessus* UC22 y *M. abscessus* subsp. *bolletii* 1513 cambiaron completamente de ubicación, pasando de los respectivos grupos A_2 y A_3, al grupo A_1 (Figura 4.9).

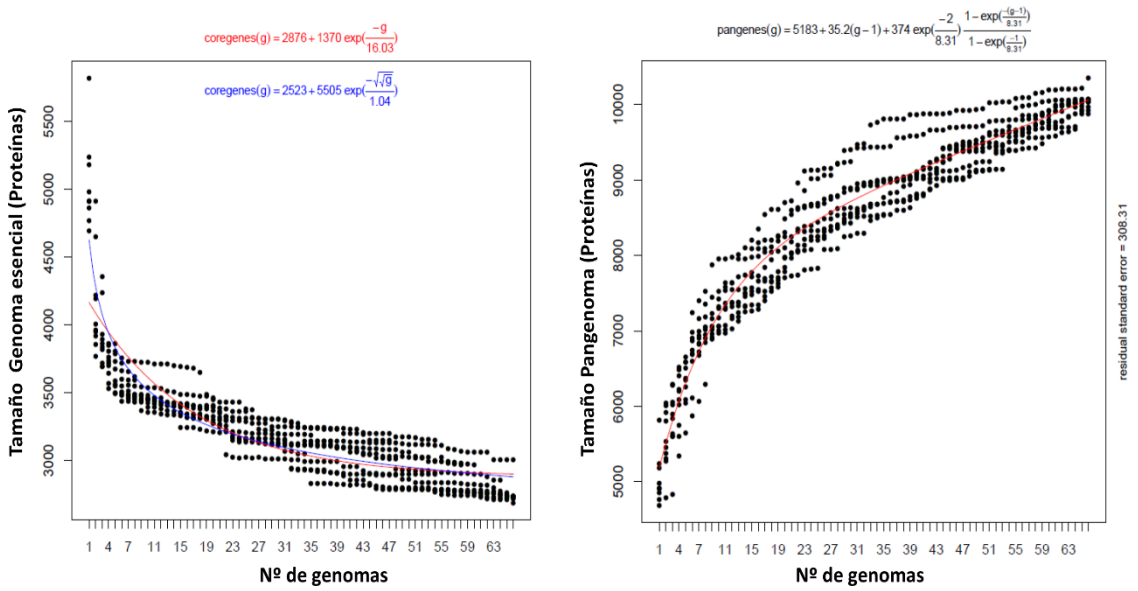


Figura 4.6. Curvas representativas de la proyección del genoma esencial y pangenoma del grupo *abscessus-chelonae-immunogenum*.

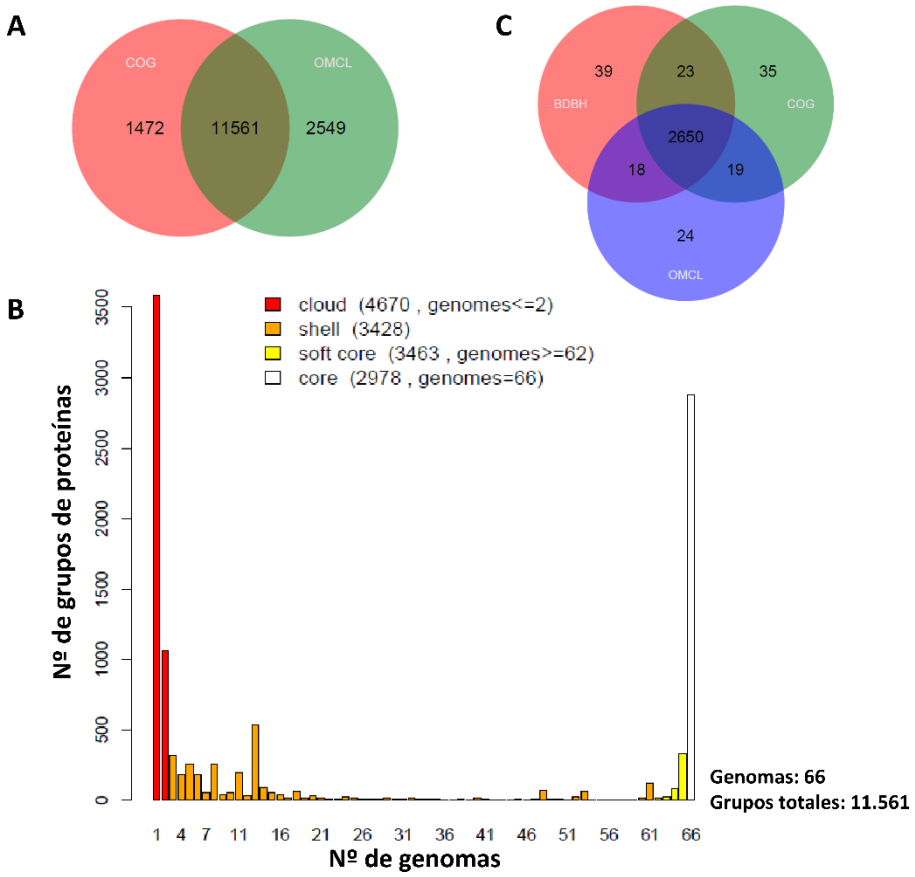


Figura 4.7. Número de grupos que conforman el A) pangenoma, B) clasificación de las distintas familias proteicas del pangenoma en genoma esencial, genoma esencial laxo, *Shell* y *Cloud* y C) genoma esencial monocopia. El número de familias proteicas del genoma esencial laxo representado en la leyenda es el resultado de la suma de grupos presentes en el 95% de los genomas y el genoma esencial estricto.

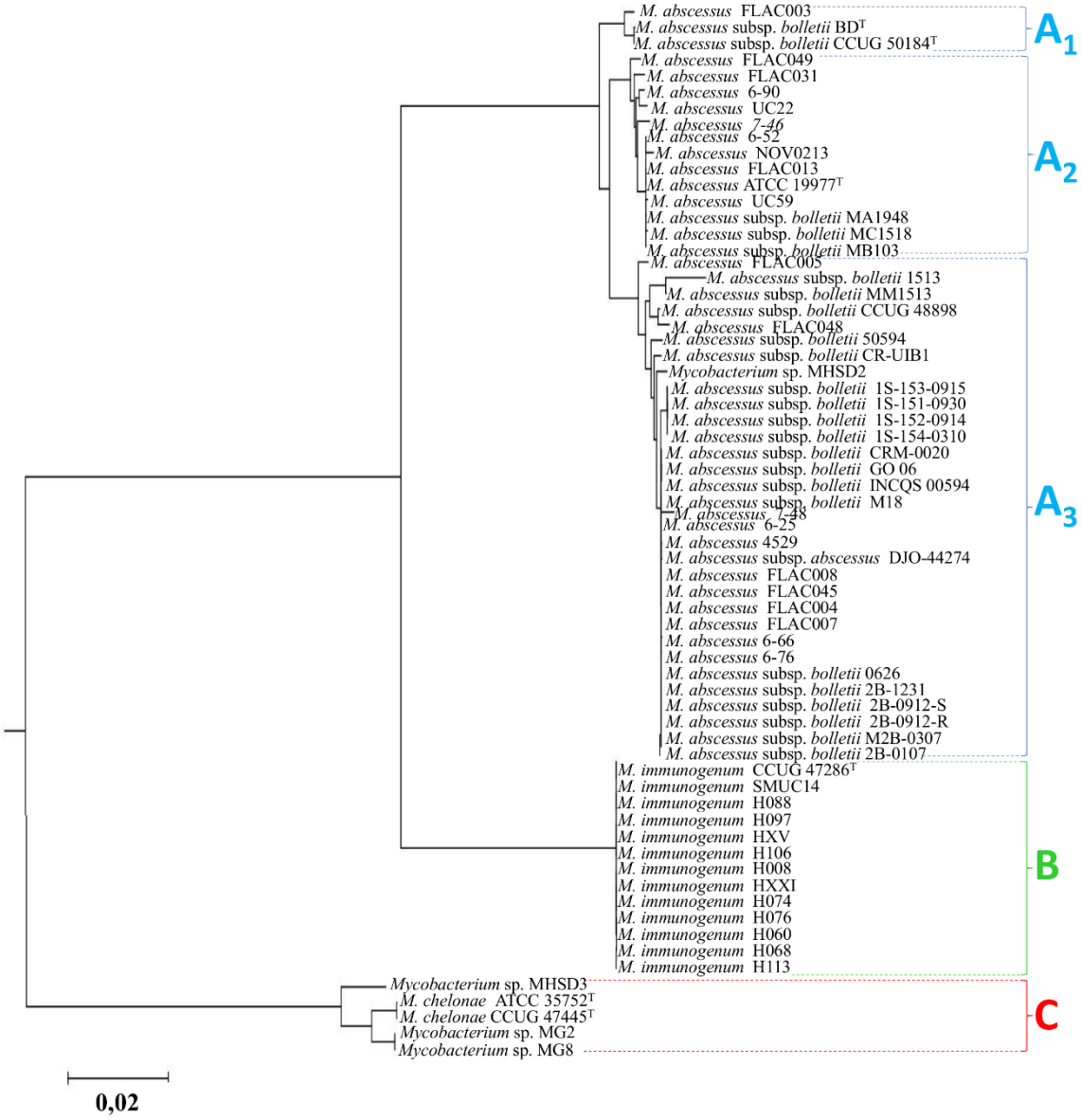


Figura 4.8. Árbol basado en las posiciones homólogas derivadas del genoma esencial monocopia. Para mas detalles relativos a los grupos obtenidos, ver texto.

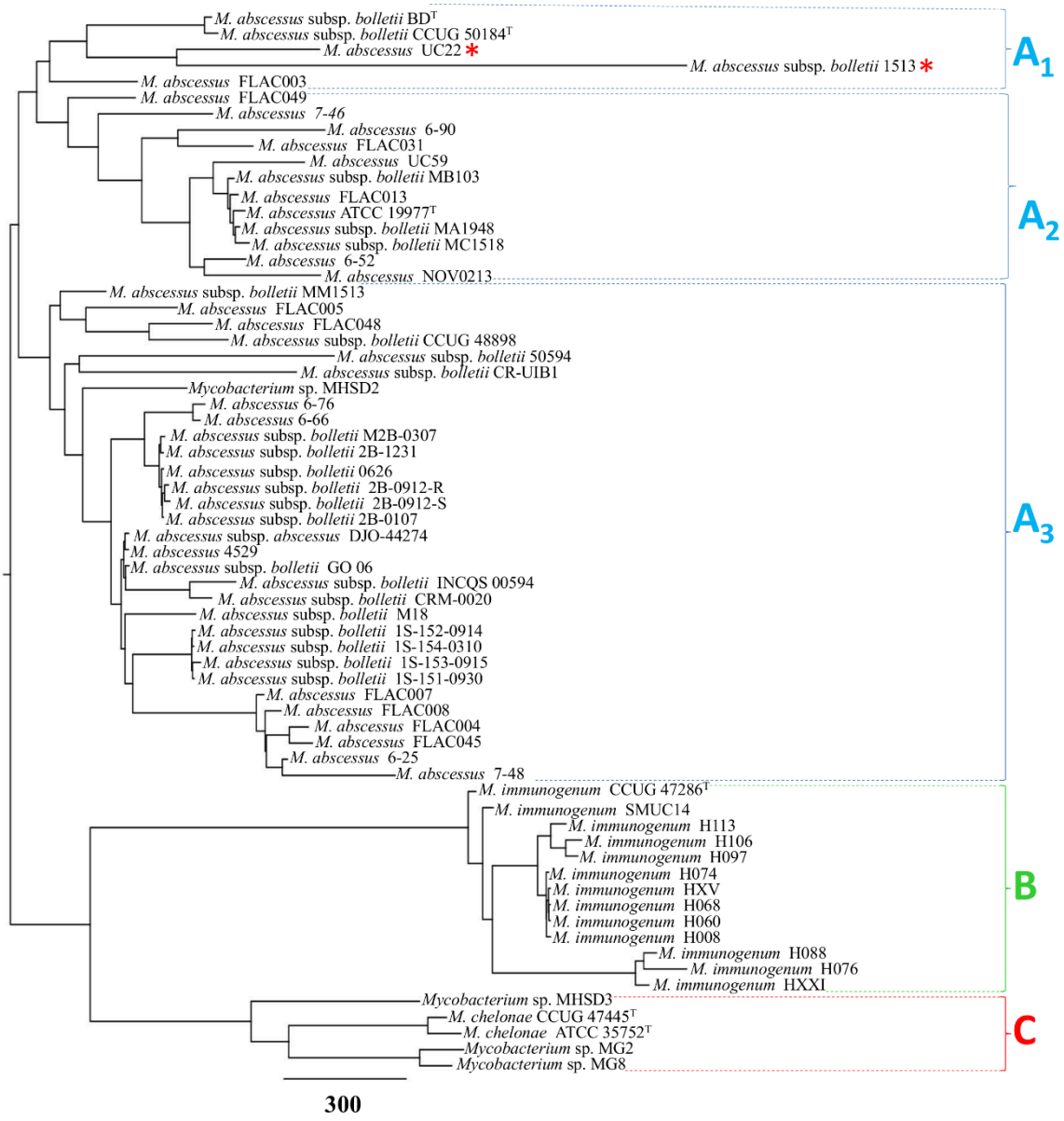


Figura 4.9. Dendrograma derivado de la matriz del pangenoma basada en la presencia/ausencia de genes. Para más detalles relativos a los grupos obtenidos, ver texto.

4.3.3. Genoma esencial y pangenoma de *Mycobacterium immunogenum*

Para este apartado se utilizaron los 13 genomas disponibles de *M. immunogenum*, ya utilizados en el estudio del genoma esencial y pangenoma del grupo *M. abscessus-chelonae-immunogenum* y que se incluyen en la Tabla suplementaria 2 del Anexo 1. Las cepas H008, H060, H068, H074, H076, H088, H097, H106, H113, HXV y HXXI tienen un origen ambiental, no clínico. Proceden de aislamientos realizados a partir de un sistema de distribución de aguas de consumo humano, el cual fue tratado con cloro y sometido a periodos de nitrificación (*Enviromental Protection Agency*, EEUU EPA, Gomez-Alvarez y colaboradores). Por su parte, las cepas SMUC14 y CCUG 47286^{T} tiene un origen clínico. Concretamente, la cepa SMUC14 fue aislada a partir de un absceso cerebral que presentaba una infección polimicrobiana [95] y la cepa CCUG 47286^{T} fue aislada de un lavado bronco-alveolar (Hospital Barner, St. Louis, Missouri, USA, 1990).

Las curvas obtenidas del genoma esencial y del pangenoma de *M. immunogenum*, utilizando como base para los cálculos las condiciones C50/S70, mostraron un descenso del número de grupos compartidos por todos los genomas, el cual se estabiliza en torno a las 5.200 familias proteicas en el genoma esencial. Por su parte, el pangenoma mostró una trayectoria ascendente a medida que se incorporaban más genomas, pero definiendo una curva que tendía claramente al efecto meseta por encima de las 6.200 familias proteicas (Figura 4.10), no observándose un incremento significativamente alto de nuevas familias proteicas cuando se incorporaban más de 8 genomas. Así pues, en este caso no se observó una tendencia a un crecimiento lineal del pangenoma sino a un ejemplo que tiende al cierre o estabilización del mismo.

Teniendo en cuenta la intersección de los grupos detectados por los algoritmos COG y OMCL, el pangenoma de esta especie resultó en un total de 6.202 grupos de proteínas (Figura 4.11A). De estos 6.202 grupos, 5.170 pertenecieron al genoma esencial estricto de los 13 genomas, 5.244 al genoma esencial laxo, mientras que *Shell* y *Cloud* se caracterizaron por 288 y 670 grupos respectivamente (Figura 4.11B). Estos datos reflejan una incorporación media de 79 proteínas nuevas por genoma añadido al análisis. En este caso el número de familias proteicas pertenecientes al *Cloud* sólo supone un 10,8 % del

tamaño del pangenoma calculado, mientras que el genoma esencial estricto supone un 83,36 % del mismo, más el 5,84 % restante del genoma accesorio.

La determinación del conjunto de proteínas codificadas por genes en monocopia del genoma esencial, utilizando el consenso de los tres algoritmos, reveló la presencia de 5.059 grupos que cumplen con esta premisa y la regla 50/70 (Figura 4.11C).

4.3.3.1. Relaciones basadas en el genoma esencial y el pangenoma de la especie *Mycobacterium immunogenum*

El análisis comparativo del genoma esencial de genes en monocopia, formado en este caso por 5.059 proteínas, dio lugar a un alineamiento constituido por 1.491.698 posiciones, de las cuales permanecen 1.483.980 después de la purga del mismo con GBLOCKS, lo que supone un 99,48 % de posiciones homólogas con respecto al alineamiento original. Sobre este alineamiento se computó el árbol del genoma esencial (Figura 4.12A). Por su parte, el dendograma del pangenoma se calculó sobre una matriz de presencia/ausencia de las 6.202 proteínas que conforman el mismo (Figura 4.12B).

Basándose en el árbol evolutivo obtenido para el genoma esencial, se observó que las 13 cepas están estrechamente relacionadas, salvo pequeñas diferencias que provocan la bifurcación del árbol en dos ramas principales, quedando las cepas CCUG 47286^{T} y SMUC14 en una rama y el resto de cepas agrupadas entre sí. Las diferencias, aunque pequeñas, dentro de esta gran rama permitieron definir cuatro agrupaciones diferentes: la formada por las cepas H068, H060 y H113; las cepas H106, HXV y H097; las cepas H076, H088 y HXXI; y por último la agrupación que incluye a las cepas H008 y H074. Por su parte, el árbol basado en el pangenoma mostró una distribución de cepas claramente diferenciada con respecto a la observada en la representación del genoma esencial, aunque las distancias observadas no son tan apreciables, y por tanto discriminativas, como las obtenidas en otros casos (ver pangenoma de MCR). Las discrepancias en este caso afectan a 1.042 proteínas que no pertenecen al genoma esencial estricto, un 16,8 % del total del pangenoma. Las cepas CCUG 47286^{T} y SMUC14, ambas de origen clínico, quedaron claramente agrupadas entre sí una vez más.

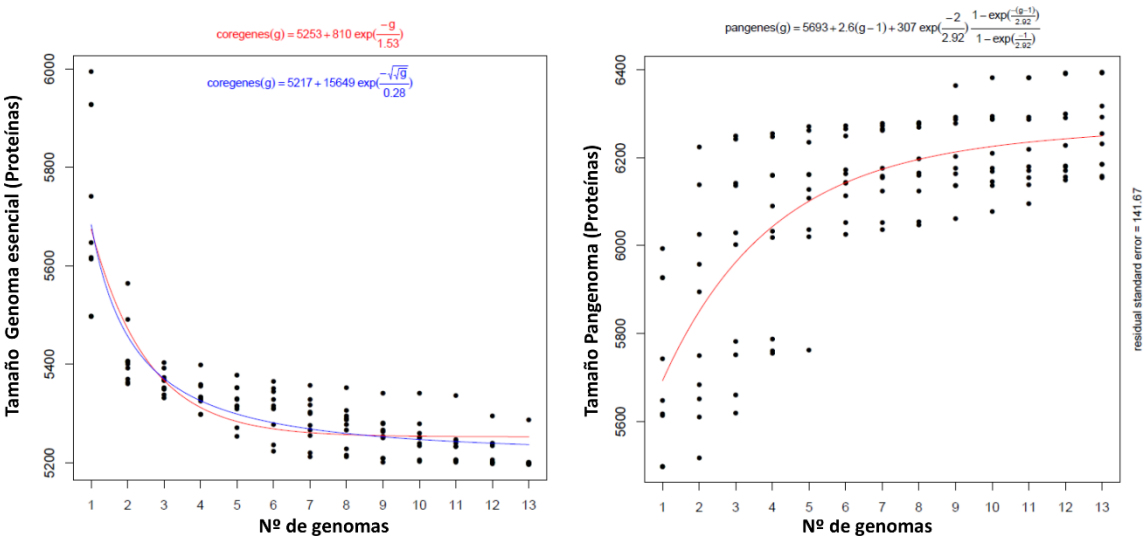


Figura 4.10. Curvas representativas de la tendencia del genoma esencial y pangenoma del grupo de cepas analizadas de *Mycobacterium immunogenum*.

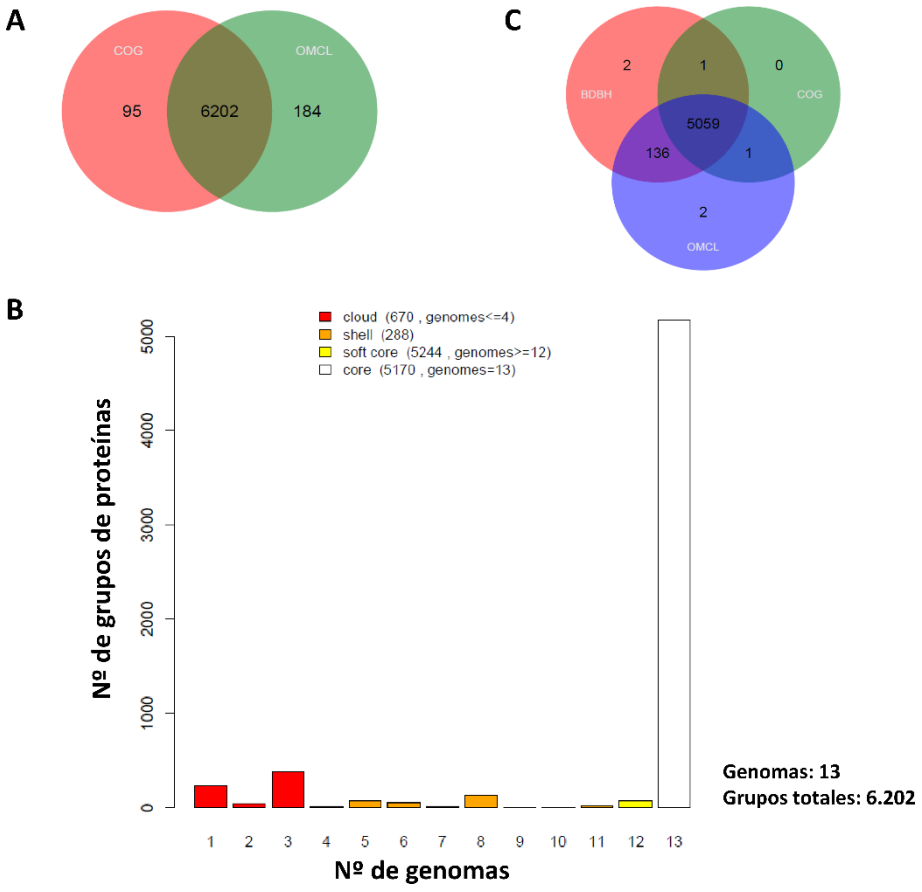


Figura 4.11. Número de grupos que conforman el A) pangenoma, B) clasificación de las distintas familias proteicas del pangenoma en genoma esencial, genoma esencial laxo, *Shell* y *Cloud* y C) genoma esencial monocopia. El número de familias proteicas del genoma esencial laxo representado en la leyenda resulta de la suma de grupos presentes en el 95 % de los genomas y el genoma esencial estricto.

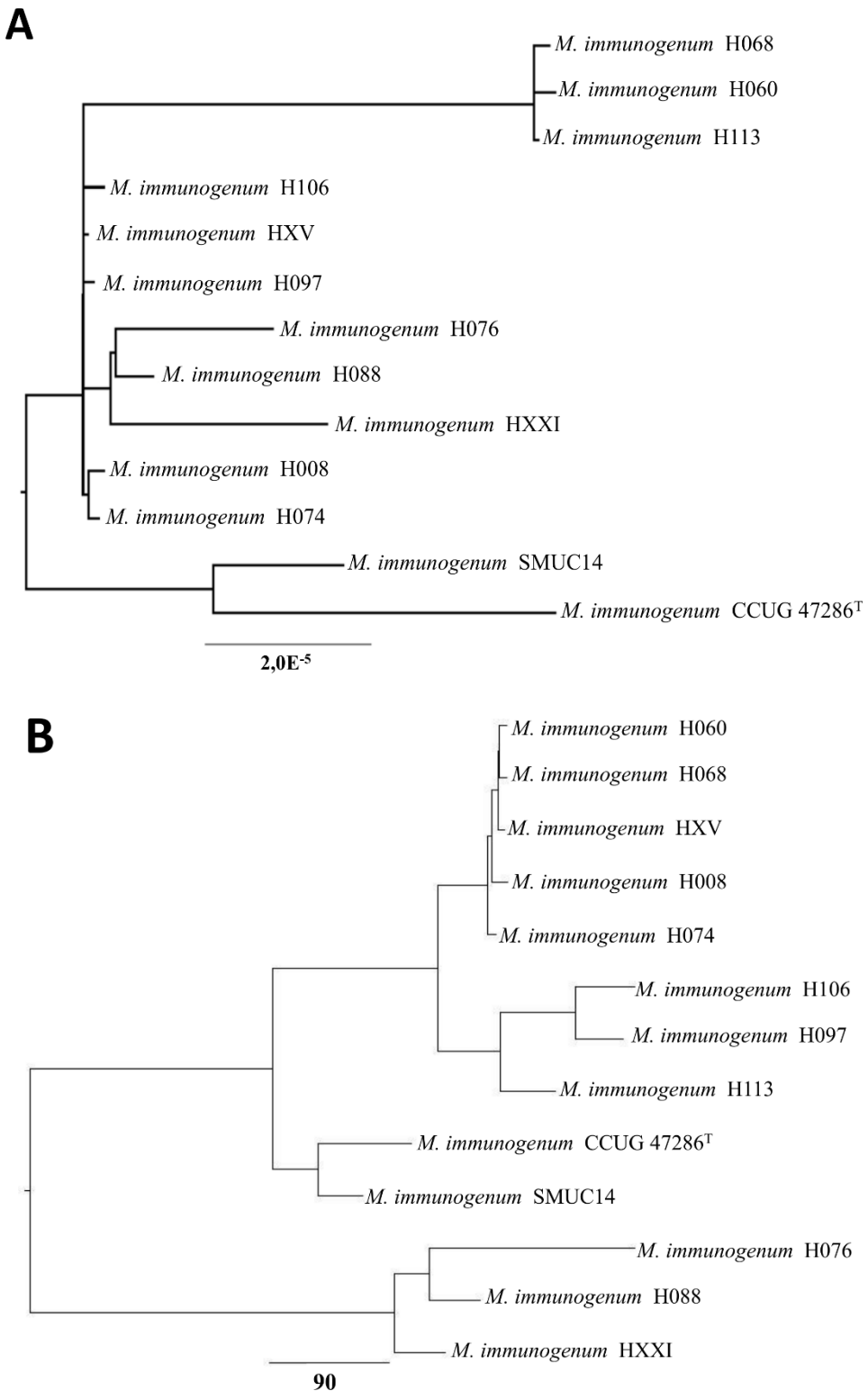


Figura 4.12. Representación del árbol basado en el genoma esencial monocopia (A) y el dendograma basado en la matriz de presencia/ausencia del pangenoma (B) del conjunto de cepas pertenecientes a la especie *Mycobacterium immunogenum*.

El estudio de la aportación de genes exclusivos de cada uno de los tres grupos surgidos en el dendrograma del pangenoma (Figura 4.12B) mostraron los elementos comunes de cada agrupación que les diferencian del resto y que explican la distribución obtenida (Tabla 4.2). Además, también se determinaron los elementos exclusivos de las dos cepas de origen clínico, así como la categorización funcional de cada grupo de elementos exclusivos.

Tabla 4.2. Proteínas exclusivas de las distintas agrupaciones de cepas de *Mycobacterium immunogenum.*

Cepas	Nº Proteínas exclusivas	Asignados a COGs	% Clasificados
Grupo A	122	5	4,10
Grupo B (Origen clínico)	1	1	100,00
Grupo C	358	46	12,9
SMUC14	21	4	19,05
CCUG 47286^{T}	17	3	17,65
Origen ambiental (A y C)	5	1	20,00

Las cepas de origen ambiental presentaron 5 proteínas comunes y ausentes en las cepas clínicas; mientras que por su parte estas sólo presentaron una proteína común que no está presente en ninguna de las 11 cepas ambientales. Además, las cepas clínicas SMUC14 y CCUG 47286^{T} presentaron 21 y 17 genes exclusivos respectivamente. Destaca también el hecho de que, a excepción del grupo B (origen clínico), sólo un pequeño porcentaje de genes exclusivos de cada determinación se puede asignar a una categoría funcional con un E-value ≤ 10^{-5} como punto de corte, siendo en su gran mayoría proteínas hipotéticas aquellas que quedan fuera de este rango o que directamente no han sido asignadas a ningún COG.

Solo una de las cinco proteínas exclusivas de las 11 cepas ambientales fue clasificada funcionalmente. Se trata de una proteína hipotética que se catalogó como IME4 (COG4725), dentro de la categoría funcional "mecanismos de transducción de señales" (T). La única proteína presente exclusivamente en las dos cepas de origen clínico se trata de una aldolasa. De las 21 proteínas exclusivas presentes en la cepa SMUC14, solamente cuatro fueron asignadas a COGs de acuerdo a los criterios de corte seleccionados. Entre ellas encontramos una ácido carboxilico reductasa, asignada al COG3320 (“biosíntesis,

transporte y catabolismo de metabolitos secundarios" (Q)); una NAD(P)H azoreductasa asignada al COG0702 ("biogénesis de la pared celular, membrana y envoltura" (M), "transporte y metabolismo de carbohidratos"(G)), una zinc metaloproteasa asignada al COG0750 ("biogénesis de la pared celular, membrana y envoltura" (M)) y, por último, un precursor de ramnosil O-metiltransferasa , el cual se asigna al COG3510 y que corresponde a una cefalosporina hidrolasa ("mecanismos de defensa" (V)). Por su parte, la cepa tipo CCUG 47286 presenta la chaperona DnaK (COG0443, "modificación postraduccional, recambio de proteínas, chaperonas" (O)), una proteína hipotética (COG4725, "mecanismos de transducción de señales" (T) y "transcripción" (K)) y un precursor de la enzima mureína DD-endopeptidasa MepH (COG0791, ("biogénesis de la pared celular, membrana y envoltura" (M)).

En cuanto a las dos agrupaciones de cepas ambientales, con origen en el mismo estudio, se detectó una notable discrepancia en lo que respecta a las proteínas diferenciales, observándose casi tres veces más en el grupo C que en el grupo A. Al realizar su categorización funcional, la diferencia observada también fue considerable (Figura 4.13). No obstante, la mayor parte de proteínas diferenciales no pudieron ser asignadas a una categoría funcional concreta, siendo en su mayoría proteínas hipotéticas.

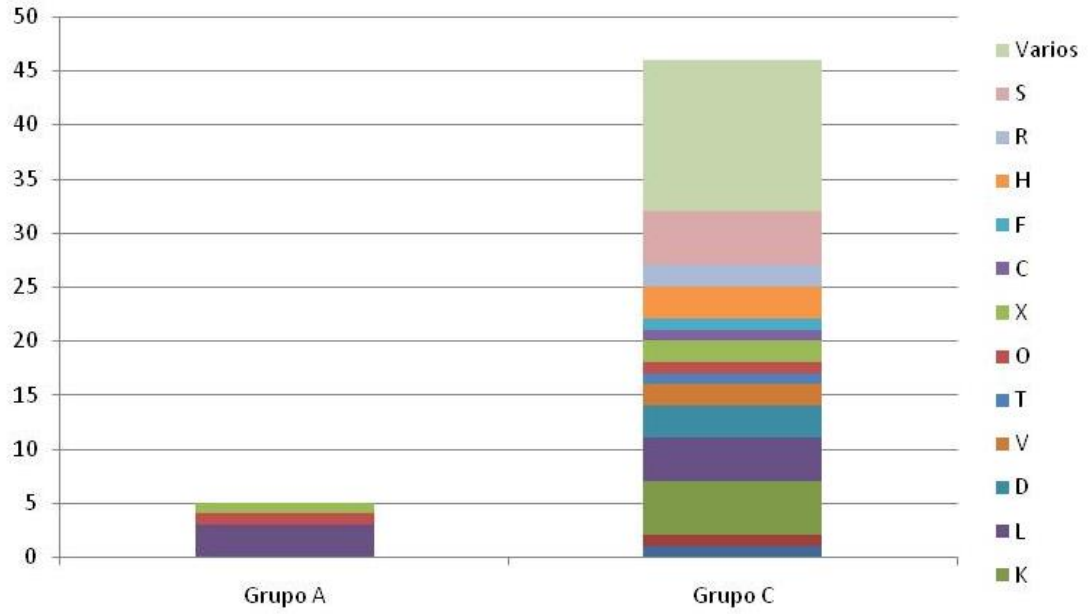


Figura 4.13. Categorización funcional de las proteínas específicas de los grupos A y C de cepas de *Mycobacterium immunogenum.* Se indica el número de proteínas que han podido ser asignadas a las distintas categorías funcionales en cada caso (Columna 3, Tabla 4.2).

Dentro de las proteínas asignadas a categorías funcionales bien definidas, es evidente que el grupo C dispone de una mayor variedad, siendo el grupo que mayor número de proteínas aporta al conjunto del pangenoma. Uno de los grupos más numeroso es el correspondiente a aquellas proteínas que pueden participar en distintos procesos celulares, ya que han sido asignadas a diferentes categorías funcionales a la vez, así como las proteínas relacionadas con transcripción (K) o las proteínas de función desconocida (S). En estas últimas se observaron dos potenciales proteínas de membrana (una de las cuales aparece anotada como proteína hipotética), una proteína putativa PPE aparentemente relacionada con la movilidad celular o la secreción, aunque sin función bien definida; y una proteína que contiene el dominio LysM, el cual parece estar relacionado con la unión a peptidoglicanos [96].

4.3.4. Genoma esencial y pangenoma de *Mycobacterium tuberculosis*

De los 3.627 genomas de esta especie presentes en GenBank, se utilizaron para este estudio los únicos 40 genomas completos disponibles, incluyendo genomas de cepas tipo. En este conjunto se incluyó en el análisis la cepa CR-UIB2 secuenciada y ensamblada durante el periodo de desarrollo experimental de la presente tesis.

El estudio realizado sobre el total de los 41 genomas considerados mostró una tendencia del genoma esencial caracterizada por una notable caída en el número de grupos de proteínas comunes a medida que se incorporaban más genomas, empezando por encima de las 4.000 familias proteicas hasta caer rápidamente por debajo de las 2.500, dibujando una curva que no parece tender a la estabilización. Por su parte, el pangenoma muestra una evolución completamente opuesta, presentando una pendiente muy pronunciada donde, nuevamente, no parece tender a la saturación sino a un crecimiento lineal muy evidente (Figura 4.14).

El tamaño del pangenoma obtenido a partir de los 41 genomas de *M. tuberculosis* estudiados es de 8.018 proteínas diferentes (Figura 4.15A), según el consenso de los algoritmos COG y OMCL, de los cuales 2.427 forman el genoma esencial estricto, 3.299 conforman el genoma esencial laxo, 489 pertenecen a *Shell* y 4.230 a *Cloud* (Figura 4.15B), calculándose una incorporación media de proteínas nuevas en torno a 136 por genoma añadido al análisis, siempre sin tener en cuenta el genoma esencial estricto. Por

su parte, el genoma esencial monocopia resultó en 2.313 familias proteicas codificadas por genes en copia única (Figura 4.15C). Destaca el número tan elevado de proteínas presentes en el *Cloud*, estableciéndose que el 52,75 % de los grupos detectados solamente están presentes en 1 ó 2 genomas como mucho, mientras que el genoma esencial , es decir, todos los elementos presentes en todos los genomas, sólo supuso el 30,27 %. El resto de familias proteicas representaron el 16,98 %.

En vista de estos resultados, se repitió el análisis reduciendo las restricciones de cobertura/identidad de 50/70 a 50/50 con el fin de comprobar si los resultados obtenidos eran consecuencia de un excesivo rigor en el análisis. En este caso se obtuvo un pangenoma de 8.026 agrupaciones distribuidas en genoma esencial (2.435), genoma esencial laxo (3.310), *Shell* (429) y *Cloud* (4.227); mientras que el genoma esencial monocopia resultó en 2.313 genes en monocopia consensuados por los tres algoritmos. Estas cifras apenas varían de las obtenidas con las condiciones 50/70. Por esta razón, y al tratarse de genomas pertenecientes a cepas de la misma especie, se consideró el análisis aplicando más rigor en términos de identidad como válido.

4.3.4.1. Relaciones basadas en el genoma esencial y pangenoma de *Mycobacterium tuberculosis*

En este caso, el dendrograma del genoma esencial constituida por proteínas codificadas por genes en monocopia se construyó a partir de un alineamiento con 652.756 posiciones homólogas, de los 687.756 originales (94,88 %), resultante del concatenado de las 2.313 proteínas en copia única del genoma esencial estricto (Figura 4.16). Por su parte, la representación gráfica de los datos del pangenoma en subcategorías se basó en la matriz de presencia/ausencia de las 8.018 proteínas que conforman el mismo (Figura 4.17).

Atendiendo a la distribución de ambas representaciones, se obtuvo un árbol basado en el genoma esencial cuyas distancias apenas mostraron diferencias entre la mayoría de genomas incluidos. Por su parte, el pangenoma mostró una diferenciación más evidente en función de la dotación de genes accesorios que presentó cada genoma. Sin embargo, en los casos de las cepas NITR204 y NITR202 quedaron claramente en ramas independientes con una distancia muy superior con respecto al resto.

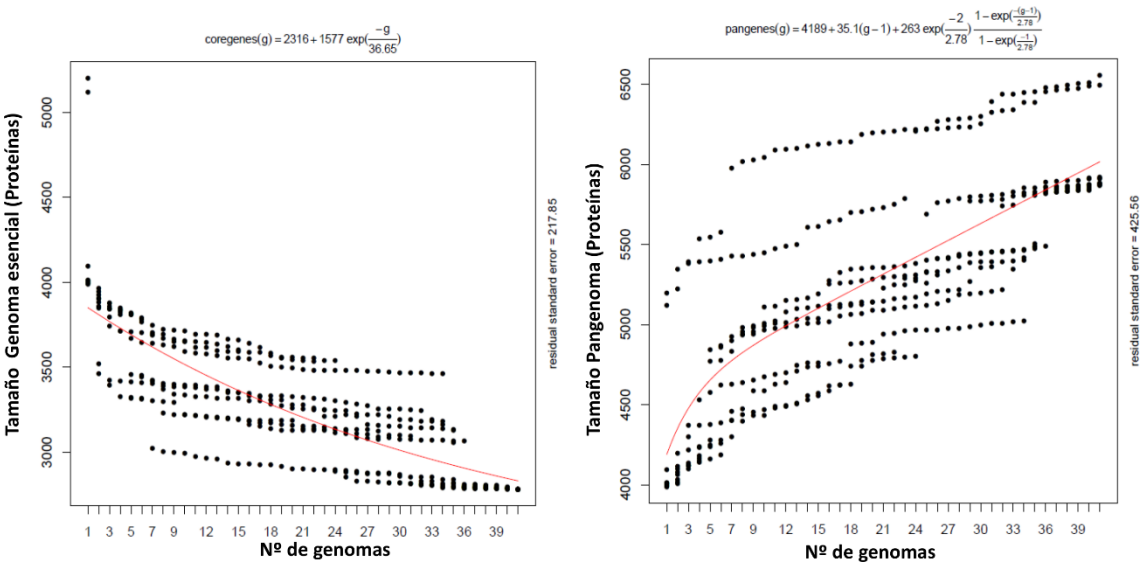


Figura 4.14. Curvas representativas de la proyección del genoma esencial y pangenoma del grupo de cepas de *Mycobacterium tuberculosis*.

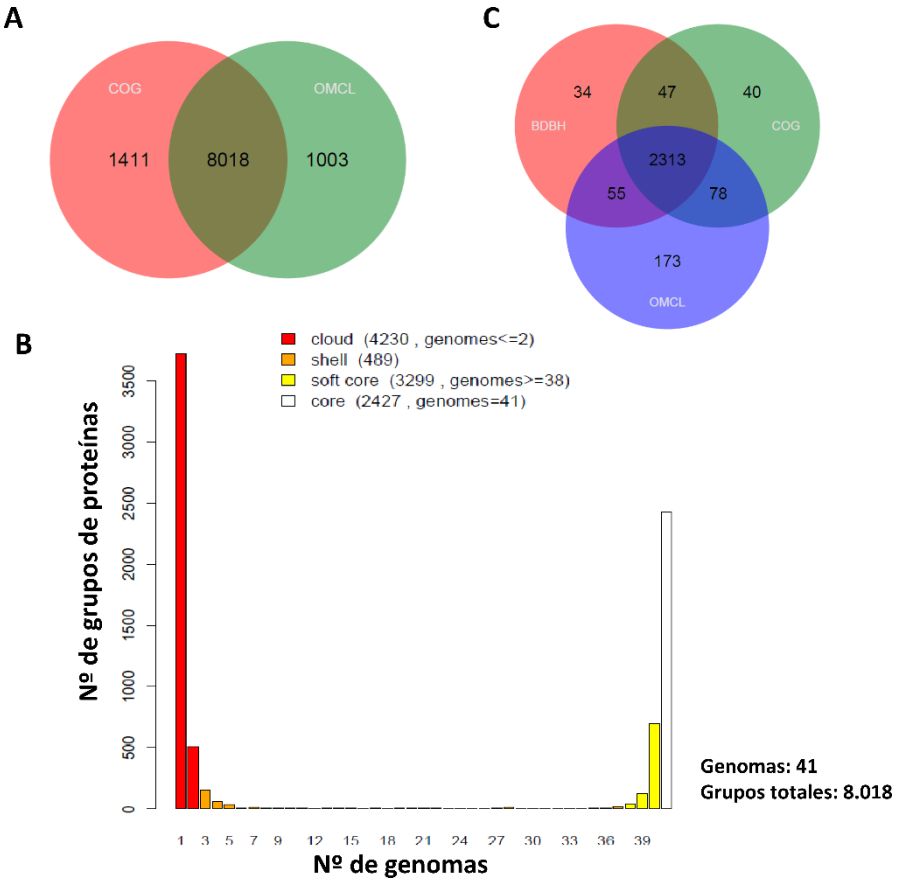


Figura 4.15. Número de grupos que conforman el A) pangenoma, B) clasificación de las distintas familias proteicas del pangenoma en genoma esencial, genoma esencial laxo, *Shell* y *Cloud* y C) genoma esencial monocopia. El número de familias proteicas del genoma esencial laxo representado en la leyenda resulta de la suma de grupos presentes en el 95 % de los genomas y el genoma esencial estricto.

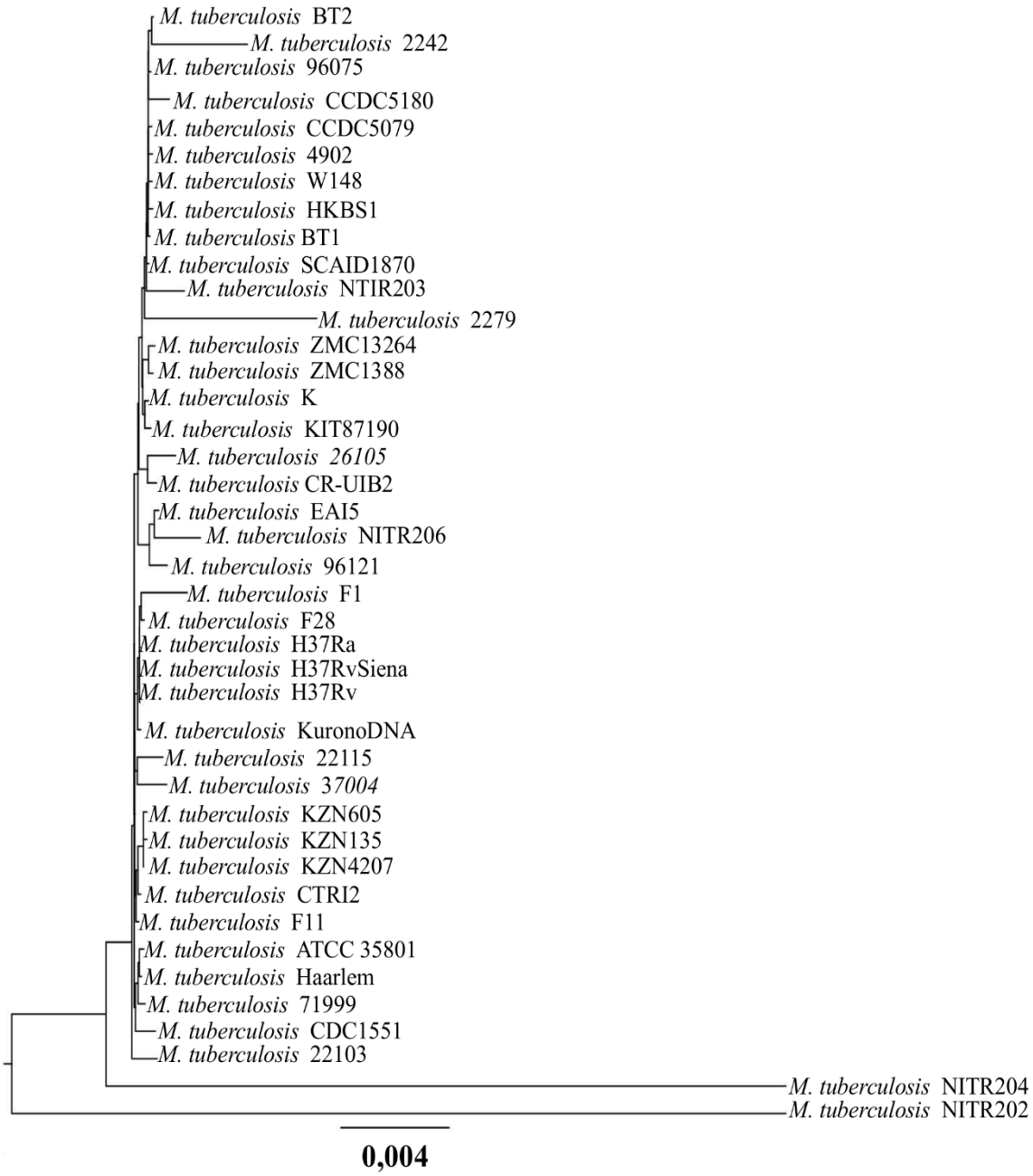


Figura 4.16. Árbol basado en las posiciones homólogas a partir del genoma esencial monocopia de los 41 genomas de *Mycobacterium tuberculosis.*

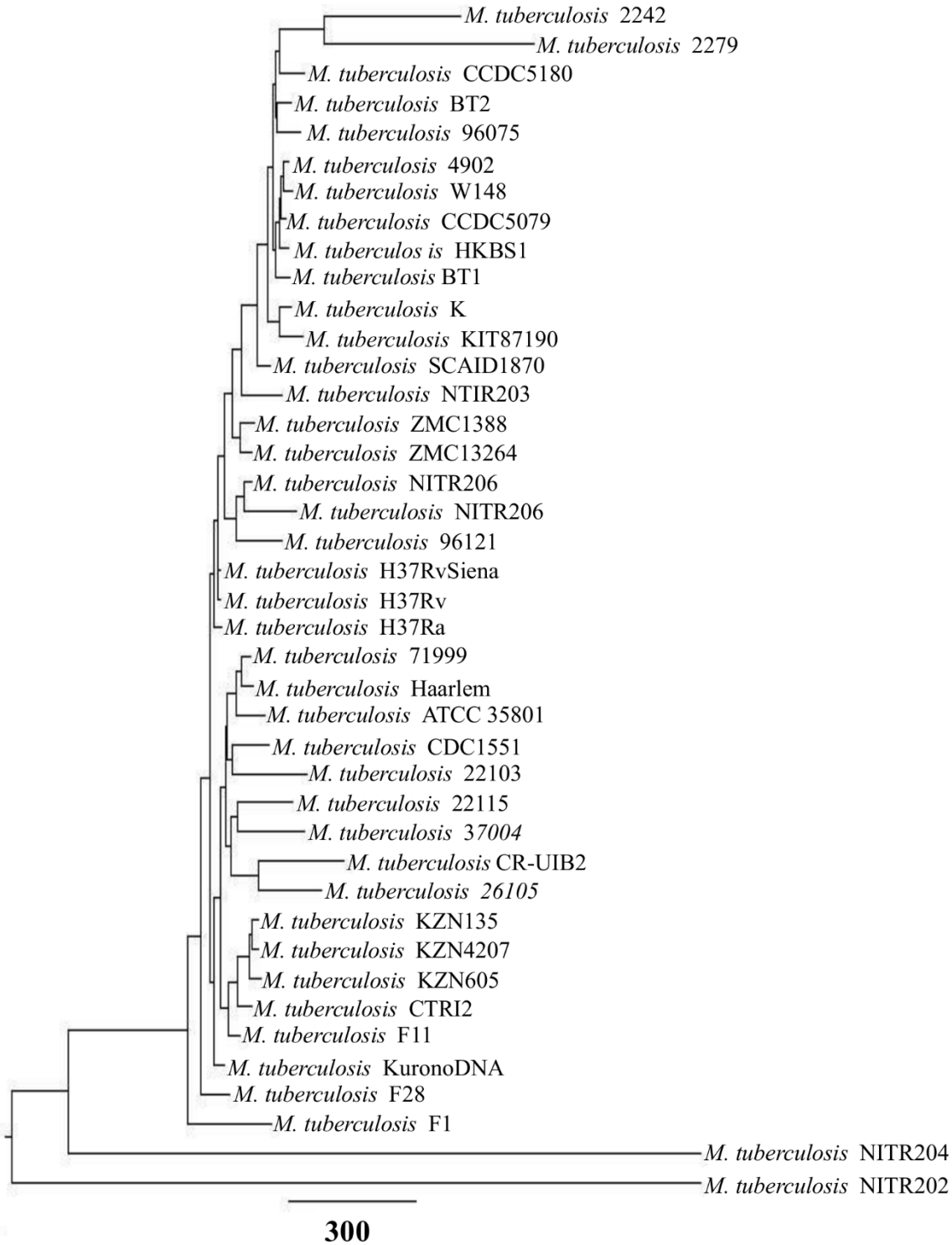


Figura 4.17. Dendograma basado en la presencia/ausencia de proteínas partir del pangenoma de los 41 genomas de *Mycobacterium tuberculosis*.

4.4. Discusión

4.4.1. Estudio del genoma esencial y el pangenoma del grupo MCR

Como se destacó en el apartado de resultados, se ha determinado un total de 21.056 proteínas diferentes a partir de 52 genomas representativos de hasta 17 especies diferentes del grupo MCR. De éstas, 1.253 proteínas representaron el genoma esencial, de las cuales 1.005 estaban codificadas por genes que se encontraban en monocopia. Esto supone que solo el 5,95 % de familias de proteínas del pangenoma están conservadas entre las distintas especies de MCR. Extrapolando este número a los CDS de un genoma, el genoma esencial representa entre un 17,5 % y un 27,4 %, un porcentaje no muy elevado, lógico si tenemos en cuenta de que hablamos de la comparación entre especies distintas, pero altamente conservado si tenemos en cuenta que después de procesar el concatenado de las proteínas codificadas por genes en monocopia del genoma esencial el 86,53 % de posiciones homólogas se mantienen.

El árbol obtenido es concordante con el árbol basado en el ADNr 16S, aunque con un poder discriminativo muy superior, especialmente en grupos complejos como el formado por las especies *M. abscessus*, *M. abscessus* subsps. *bolletii*, *M. chelonae* y *M. immunogenum.* En este conjunto se observó, por ejemplo, cómo las cepas MG2, MG8 y MHSD3 se rodearon de las cepas de *M. abscessus* y *M. abscessus* subsp. *bolletii*, mientras que el análisis comparativo basado en el genoma esencial las incluyó claramente como cepas de *M. chelonae*, hecho que concuerda con los estudios de MLSA de Gomila y colaboradores. [43] .

Además, los ADNr 16S para los genomas de *M. abscessus* y *M. abscessus* subsp. *bolletii* no fueron lo suficientemente discriminativos, apareciendo como una rama plana que indicaría una sola especie. El estudio del genoma esencial incorpora una mayor resolución al problema, separando perfectamente los genomas de las especies de *M. chelonae* y *M. immunogenum* y manteniéndolas separadas en dos ramas principales. En una de ellas hallamos a la cepa tipo de *M. abscessus* subsp. *bolletii* CCUG 50184^{T}, que se secuenció en la presente tesis, agrupada con la cepa *M. abscessus* FLAC003; mientras que la segunda rama se subdividió a su vez en dos más: una primera ramificación con la cepa tipo de *M. abscessus* ATCC 19977 ^{T}, predominando cepas anotadas como *M. abscessus*,

y una segunda agrupación donde prevalecen fundamentalmente cepas de *M. abscessus* subsp. *bolletii*. Es de destacar en este caso el hecho de que las cepas tipo de *M. abscessus* subsp. *bolletii* quedaran fuera, en una agrupación separada dentro del grupo, y no dentro de la ramificación donde predominan genomas supuestamente de la misma especie. Todo esto indicaría que, aunque los estudios del genoma esencial aportan una mayor resolución al problema, el caso concreto de estas dos especies sigue siendo complejo y requiere de un estudio más profundo.

Otro hecho destacable hace referencia a los genomas *M. smegmatis* JS623, *M. rhodesiae* NBB3, *M. chelonae* 1518 y *M. fortuitum* Z58. Estos genomas, ya en el estudio con ADNr 16S, quedaban alejados del hipotético lugar donde se deberían haber ubicado, es decir, con los demás genomas representantes de estas especies. Cabe recordar que para este estudio se utilizaron secuencias de ADNr 16S de cepas de referencia obtenidas de cada una de las cepas tipo. En este sentido, se consideró como criterio de pertenencia a una determinada especie el que las cepas quedaran en la misma rama del árbol obtenido mediante el análisis comparativo de sus secuencias de ADNr 16S. En base a este resultado, el análisis del genoma esencial no hizo más que confirmar la topología obtenida con el ADNr 16S de forma más concluyente. Esto es un reflejo de que, a pesar de que el ADNr 16S no es muy resolutivo en determinados grupos de cepas o especies debido a diversas limitaciones [97], sigue siendo una herramienta muy útil como punto de partida para establecer el contexto evolutivo adecuado y necesario en este tipo de estudios. Una secuencia de ADNr 16S igual o con un elevado porcentaje de identidad con otras puede ser indicativo, aunque no siempre, de la pertenencia de las correspondientes cepas a una misma especie. Sin embargo, unos valores de similaridad basados en la comparación de las secuencias de ADNr 16S suficientemente diferentes (< 97%) es indicativo inequívoco de que estamos ante especies diferentes. En consecuencia, lo más probable es que estos genomas no pertenezcan a la especie con cuyo nombre se describieron originalmente.

El pangenoma del grupo de MCR evaluado en su conjunto, y recordando que está constituido por especies diferentes, según las estimaciones realizadas se puede considerar como abierto; observándose un crecimiento prácticamente lineal que refleja la gran capacidad de incorporación de nuevas proteínas a medida que se añaden nuevos genomas. Esto unido al hecho de que el 65,5 % de las más de 20.000 proteínas diferentes

encontradas aparecían solo en 1 ó 2 genomas, refuerza el hecho de la gran capacidad adaptativa de este grupo de bacterias, hecho que corroboraría su amplia representatividad y distribución en el ambiente [98], a través de una capacidad de incorporación de las nuevas funciones necesarias para la supervivencia en ambientes concretos o en respuesta a condiciones adversas o variables puntuales o a más largo plazo. Esto no quiere decir que todas las especies presenten esta gran plasticidad genómica, sino que, en conjunto, estamos ante un género, *Mycobacterium*, con un gran potencial de hacerlo.

La agrupación que se obtuvo a través del análisis genómico comparado, y que se observó en el árbol basado en el genoma esencial, nos proporciona por tanto una información que en términos evolutivos relaciona de forma consistente a todas las cepas o especies estudiadas. Por su parte, el dendograma del pangenoma ofrece un punto de vista diferente al partir de qué genes o proteínas comparten o no los diferentes genomas comparados. Es por esto que, en este caso, la información basada en la presencia/ausencia de genes no puede ser interpretada desde el punto de vista evolutivo. Aunque en líneas generales los 8 grupos perfilados en la representación del genoma esencial se mantuvieron también en la representación del pangenoma, posiblemente por el peso específico que implica la inclusión de las proteínas del genoma esencial, su distribución varió en algunos casos. Por ejemplo, *M. llatzerense* presentaba según el genoma esencial como especies evolutivamente más cercanas a *M. smegmatis*, *M. fortuitum* y, especialmente, *M. neoaurum*. Sin embargo, parece compartir más genes o proteínas con las especies *M. marinum* y el grupo *abscessus-chelonae-immunogenum*, por lo que en el dendograma del pangenoma apareció agrupada con estas últimas en lugar de las mencionadas anteriormente. De hecho, comparte más genes con el grupo comentado que la propia *M. marinum*, la cual es la especie más cercana al mismo. Cabe destacar que la cepa tipo de *M. llatzerense* MG13^{T} fue aislada compartiendo el mismo nicho ecológico que las cepas MG2, MG8 de *M. chelonae* [43,99], por lo que el hecho de compartir hábitat podría haber favorecido a la transferencia, captación de genes o el desarrollo de mecanismos similares entre estas especies, y que ha conducido a este resultado. Un caso similar ocurre con las especies *M. phlei* y *M. hassiacum*, las cuales comparten más elementos con las especies comentadas en el caso anterior, que con las especies filogenéticamente más cercanas a ellas (*M. vanbaleenii*, *M. vaccae, M. phlei* y *M. chubbuense*). De hecho, con quien más

genes parecen compartir es con *M. thermoresistibile*, una especie muy alejada de éstas según el análisis comparativo basado en el genoma esencial.

Sobre el dendrograma del pangenoma queda de manifiesto la ganancia de genes que ha llevado a las distintas bifurcaciones del mismo, genes que se han ido incorporando en función de las condiciones o necesidades presentadas en cada caso. Un punto interesante de este análisis es el que afecta a la determinación de las proteínas exclusivas de una especie determinada, teniendo en cuenta todas las cepas utilizadas en cada caso. Como se pudo comprobar en el apartado de resultados, el rango de proteínas exclusivas calculado varió entre las 18 para las especies *M. abscessus* y *M. abscessus* subsp. *bolletii*, hasta las 1.071 proteínas exclusivas de *M. rhodesiae*. En este último caso destacó el hecho de que se utilizó el único genoma disponible de *M. rhodesiae* correctamente identificado. A pesar de poder identificar las proteínas exclusivas del conjunto de cepas de cada especie, sólo un pequeño porcentaje de ellas (del 13 al 39 % según la especie) pudo asociarse a un COG determinado y, por lo tanto, claramente a una categoría funcional. El resto, cuyo E-value era mayor a 10^{-5} (menos negativo), fueron en su gran mayoría proteínas hipotéticas que podrían suponer una buena fuente de nuevas funciones todavía por determinar dentro del grupo MCR. Al mismo tiempo, muchas proteínas anotadas inicialmente como hipotéticas pudieron ser asignadas a una función determinada a través de este procedimiento.

En líneas generales la variedad metabólica observada entre los diferentes conjuntos de proteínas exclusivas fue elevada, aunque evidentemente menor en aquellas especies donde el número de proteínas exclusivas ha sido bajo. Durante su proceso evolutivo las distintas especies se han visto probablemente sometidas a diferentes presiones selectivas o condiciones definidas específicas en sus nichos ecológicos, que junto con las especies con las que lo compartían y con las que hayan podido intercambiar material genético, han ayudado a la potencial ganancia o pérdida de material genético para responder a estos fenómenos [100]. En base a dichas condiciones se favorecen determinadas funciones sobre otras haciendo que, evidentemente, las distintas categorías funcionales observadas en la gráfica (Figura 4.5) no tengan el mismo peso en las distintas especies, en gran medida debido a este proceso de adaptación.

4.4.2. Estudio del genoma esencial y el pangenoma del grupo *abscessus-chelonae-immunogenum*

Las especies *M. abscessus* (incluidas subespecies), *M. chelonae* y, en menor medida, *M. immunogenum*, son especialmente importantes en cuanto a infecciones oportunistas protagonizadas por las denominadas micobacterias no tuberculosas (MNT) [11,21,31,38,39]. Si tenemos en cuenta el árbol del genoma esencial generado para el grupo MCR, las MNT son especies estrechamente relacionadas entre sí desde el punto de vista evolutivo. Al centrarnos en este grupo concreto se observó que el genoma esencial, a pesar de solo representar un 25,8 % del total del pangenoma, está muy conservado (96 % de posiciones homólogas conservadas). En este caso se aumentó el número de genomas de las especies *M. abscessus* y *M. abscessus* subsp. *bolletii*, para intentar incrementar la resolución entre estas dos especies, así como también los 11 genomas adicionales de *M. immunogenum* disponibles en el momento del estudio.

El árbol del genoma esencial de genes presentes en monocopia obtenido a partir de las posiciones homólogas de 2.650 proteínas separó nuevamente esta agrupación de especies fenotípicamente diferentes en tres grandes ramas: una ramificación para las cepas de la especie *M. chelonae*, una segunda para *M. immunogenum*, y una tercera rama que engloba el resto de los genomas de las otras dos especies. El resultado esperado era obtener en esta rama dos subramas con distancias evolutivas cortas entre ellas, pero donde se conseguiría separar por una parte los genomas de *M. abscessus* subsp. *bolletii* y los de *M. abscessus* por otra. En su lugar se obtuvo una división en dos subramas. En la primera se agruparon los dos genomas de las cepas tipo de *M. abscessus* subsp. *bolletii*, prácticamente idénticos, y la cepa FLAC003 de *M. abscessus*. La segunda, por su parte, se subdividió a su vez en dos nuevos grupos, donde uno de los cuales aparece claramente dominado por genomas de *M. abscessus*, incluido el de la cepa tipo ATCC 19977, mientras que en el segundo predominan genomas de *M. abscesus* subsp. *bolletii*. Si tomamos las cepas tipo como referencia del estudio, tendríamos solo tres genomas pertenecientes a *M. abscessus* subsp. *bolletii* (dos de las cuales son las cepas tipo), y un grupo de genomas todos ellos pertenecientes a la especie *M. abscessus* o bien un grupo de genomas agrupados en torno a la cepa tipo de la misma, que si pertenecerían a dicha especie; además de un tercer grupo de genomas que podrían pertenecer a una tercera

subespecie. En cualquier caso, sigue poniéndose de manifiesto la enorme problemática taxonómica todavía existente incluso en la era genómica dentro de este grupo. En un reciente estudio centrado específicamente en la resolución de la taxonomía de este complejo grupo se propuso la separación de la especie *M. abscessus* en tres subespecies distintas basándose en 1) matrices de distancias obtenidas a partir de valores de ANIs y 2) la respuesta diferencial a macrólidos como una característica fenotípica clave desde el punto de vista clínico: *M. abscessus* subsp. *abscessus* (definida por la cepa tipo ATCC 19977^T), *M. abscessus* subsp. *bolletii* (definida por la cepa tipo CCUG 50184^T) y *M. abscessus* subsp. *masiliense* (definida por la cepa tipo CCUG 48898^T). En dicho estudio se establece que existen notables diferencias y evidencias como para dividir la especie *M. abscessus* en tres subespecies, pero no suficientes para considerarlas especies totalmente distintas [101]. Aunque los genomas utilizados en el presente trabajo no son los mismos a los del mencionado estudio, la topología del árbol que se presenta (Figura 4.8) define estas tres grandes agrupaciones, cada una representada por la cepa tipo correspondiente; por lo que ambos resultados se complementan, aunque utilizando puntos de vista completamente distintos. Cabe destacar que, en el presente estudio, el genoma de la cepa tipo CCUG 48898^T venía definido por GenBank como un genoma perteneciente a *M. abscessus* subsp. *bolletii*, probablemente debido a los constantes procesos de reclasificación que han sufrido estas especies. Sin embargo, realmente representa el genoma de la cepa tipo para *M. abscessus* subsp. *massiliense*. En cualquier caso, reflejada esta aclaración y para no confundir términos en lo expuesto hasta el momento, en el presente trabajo se seguirá tratando a dichos genomas con el nombre a partir del cual vinieron definidos en las bases de datos.

El tamaño del pangenoma determinado para estas tres especies tan cercanas fue de 11.081 proteínas y presentó una tendencia abierta, lo cual es más significativo teniendo en cuenta que 48 de los 66 genomas utilizados son de las especies *M. abscessus*, y *M. abscessus* subsp. *bolletii*. Es decir, que estas dos especies son las que mayor peso tienen en este pangenoma, y aun así este sigue siendo abierto. De hecho, el 42 % de las proteínas determinadas se englobaron en la sección del pangenoma llamada *Cloud*, por lo que solo se presentan en uno o dos genomas. Analíticamente esta tendencia refleja una gran capacidad para la incorporación de nuevas proteínas a medida que si se añaden nuevos genomas al estudio. Es decir, la plasticidad general de este conjunto de genomas,

especialmente los de las dos últimas especies mencionadas, parece elevada y se caracteriza por una gran capacidad de ganar o perder genes de acuerdo a las necesidades circunstanciales. Al ser un grupo especialmente implicado en infecciones oportunistas este hecho es de gran relevancia, ya que su capacidad de adaptarse a las distintas condiciones adversas, tratamientos o medidas preventivas es también potencialmente superior. Al ser las especies *M. abscessus* y *M. abscessus* subsp. *bolletii* las que mayor peso tienen en este estudio, es de suponer que gran parte de la apertura del pangenoma, y por tanto de la plasticidad genómica, es debida a ellas. Esto estaría de acuerdo con el hecho de que de las cuatro especies son, junto con *M. chelonae*, las que mayor número de infecciones provocan y las que mayor resistencia presentan a elementos adversos externos.

En las representaciones del genoma esencial y pangenoma la distribución de las cepas fue muy similar, formándose las tres grandes agrupaciones de especies (A, B y C) y tres subgrupos dentro de la rama *abscessus* - *bolletii* (A1, A2 y A3), donde sólo las cepas UC22 y 1513 cambiaron su posición a un subgrupo diferente (de A2 a A1 y de A3 a A1 respectivamente). Esta concordancia entre ambas representaciones es indicativa de que los genomas agrupados en el análisis comparativo del genoma esencial también presentan una gran similitud en cuanto a las proteínas que comparten.

4.4.3. Estudio del genoma esencial y el pangenoma de la especie *Mycobacterium immunogenum*

Al de intentar identificar diferencias entre aquellas cepas aisladas de fuentes más bien ambientales y aquellas aisladas directamente de pacientes como agentes infecciosos, el contraste resultante fue como mínimo interesante. El genoma esencial de esta especie, en base a los 13 genomas disponibles en el momento del análisis, mostró un alto grado de conservación al representar el 83,34 % del total de proteínas determinadas en el pangenoma; de las cuales 5.059 de los genes se encontraban en monocopia. Este alto grado de conservación también se trasladó al alineamiento de las correspondientes proteínas únicas, que alcanzaron un 99,48 % de identidad. La incorporación de ocho genomas mostró que la tendencia del genoma esencial estabilizaba claramente su descenso, sin apenas variar hasta el final de la estimación. Esto apuntó a que el hecho de incorporar más genomas al análisis no repercutía en un descenso muy significativo. Hay

que destacar que, tal y como se estableció en el apartado de materiales y métodos, estos estudios dentro de una sola especie se hicieron más estrictos subiendo la similitud de 50 a 70%, por lo que el resultado adquiere mayor peso ya que sigue mostrando un nivel de conservación muy elevado a pesar de aumentar el rigor comparativo.

El árbol obtenido a partir del genoma esencial de genes monocopia mostró dos grandes agrupaciones: la correspondiente a los aislamientos ambientales y la de las dos cepas clínicas. Esto implica que las dos cepas clínicas son ligeramente más parecidas entre si en el genoma esencial que con respecto al resto de cepas. Sin embargo, en líneas generales las distancias evolutivas observadas en este árbol fueron bajas por lo que, salvando pequeñas diferencias detectadas, se podría afirmar que son prácticamente iguales. Esto es consecuencia del alto grado de conservación observado anteriormente en el concatenado, donde prácticamente resultaron ser idénticas.

La tendencia del pangenoma calculado fue cerrada, dibujando una curva que tiende claramente a la saturación prácticamente desde la incorporación incluso del quinto genoma. La conclusión lógica a raíz de este dato sería que esta especie no tendría tanta capacidad de incorporar o perder genes en función de las necesidades existentes en el ambiente o no lo ha necesitado. Como se ha destacado en el caso anterior, prácticamente todo el pangenoma determinado se agrupó dentro del genoma esencial y solo el 10,8 % de proteínas se clasificaron dentro del *Cloud*, por lo que la cantidad de nuevas proteínas incorporadas por cada genoma es baja y no comparable con los casos expuestos anteriormente. La representación del árbol de pangenoma mostró una distribución diferente respecto a la del genoma esencial. Sin embargo, las cepas de origen clínico siguieron agrupándose conjuntamente, lo que significa que presentan más elementos en común entre ellas que con el resto. Estos dos genomas realmente sólo exhibieron una proteína común diferente al resto de genomas. Concretamente una aldehido deshidrogenasa (*ald*A) aunque aportan 21 (SMUC14) y 17 (CCUG 47286^{T}) proteínas exclusivas. En la cepa SMUC14 destacó el precursor de ramnosil O-metiltransferasa, asignada al COG3510, representado por una cefalosporina hidrolasa, por lo que implicaría un mecanismo de defensa frente antibióticos. Por su parte, en la cepa tipo CCUG 47286^{T} destacó el precursor de mureina DD-endopeptidasa MepH, asignada al COG0791 (“biogénesis de la membrana celular, membrana externa”). Los elementos

implicados en la modificación de la membrana podrían jugar un papel importante en el proceso de infección. El representante de este COG es la proteína SRP, una hidrolasa asociada a membrana que estaría relacionada con procesos de invasión celular. En ambos casos se trataría de dos proteínas características que encajarían con el origen clínico de ambas cepas.

4.4.4. Estudio del genoma esencial y el pangenoma de *Mycobacterium tuberculosis*

A causa de los graves problemas de salud que ha generado históricamente en el ser humano, *M. tuberculosis* es uno de los ejemplos más característicos de representante patógeno dentro del género *Mycobacterium* [6,8]. Es un microorganismo bien conocido por la gravedad y persistencia de las infecciones que produce, las múltiples variables existentes y los múltiples mecanismos de los que puede valerse para enfrentarse a todo tipo de condiciones adversas [102]. En estudios previos del genoma esencial y el pangenoma realizados sobre esta especie (y otras pertenecientes al complejo tuberculosis) se obtuvieron similares resultados a los que se presentan aquí [103], no obstante en la presente tesis se decidió centrar el estudio en genomas nombrados como *M. tuberculosis* propiamente dicho y utilizar su ejemplo como contraste con los demás realizados dentro de este proyecto.

La media de genes codificados por un genoma de tuberculosis es de unos 4.400 genes. El genoma esencial determinado para los 41 genomas estudiados fue de 2.313 proteínas, es decir, que aproximadamente el 50 % de los genes de un genoma del agente causal de la tuberculosis son comunes a la especie. Esto supone un genoma esencial altamente conservado, aunque no tanto como por ejemplo el de *M. immunogenum*, donde un 83,36 % de proteínas se englobaron en el genoma esencial con un grado de conservación del 99,48 % de posiciones homólogas. Sin embargo, también supone que en las cepas de *M. tuberculosis* el otro 50 % restante es más variable y, de hecho, el 52,75 % de las proteínas determinadas en el pangenoma pertenecieron al *Cloud*, es decir, son específicas de 1 ó 2 genomas. Esto supone que el contenido de genes de un genoma a otro puede variar enormemente y esto se vio bien reflejado en las curvas de genoma esencial y pangenoma (Figura 4.14), donde el genoma esencial prácticamente delimitó una línea recta

descendiente, sin llegar a estabilizarse a medida que se incorporaban nuevos genomas. Por su parte, el pangenoma dibujó una tendencia completamente abierta, sin síntomas de ajustarse a una curva con saturación. Esto puede ser un reflejo de la enorme plasticidad genómica de este patógeno, con gran capacidad de variar su repertorio genético o introducir mutaciones con el fin de adoptar o perder capacidades en función de sus necesidades.

La representación gráfica derivada del análisis del genoma esencial mostró un conjunto de genomas que apenas se diferenciaban unos de otros, a excepción de las cepas NITR204 y NITR202, ambas secuenciadas en el mismo estudio [104], las cuales incluso aparecieron como dos referencias externas para el resto de genomas. En el caso del pangenoma la distribución fue diferente en función de los genes que compartían, por lo que no se formaron grupos bien definidos como en casos anteriores, no obstante, volvió a ocurrir lo mismo con las cepas NITR204 y NITR202, consecuencia de las notables diferencias con respecto al contenido de genes. A nivel mundial existen diferentes linajes de *M. tuberculosis* [105], y NITR204 se describe como perteneciente al linaje CAS (linaje 3, también conocido como CAS/Delhi), mientras que NITR202 correspondería al linaje Haarlem (linaje 4), tal y como se puede comprobar a través de sus números de acceso (CP004886 y CP005386 respectivamente). En el caso de la cepa NITR202 se observa como otro genoma representativo del linaje Haarlem quedó completamente separado de dicha cepa, por lo que a nivel de genoma esencial y pangenoma quedaría patente que este genoma realmente no pertenecería a ese linaje, si atendemos a los resultados presentados aquí. Sin embargo, estudios más discriminativos o específicos serían necesarios para determinar si se trata de un nuevo linaje o de una cepa perteneciente a otra especie del complejo *M. tuberculosis*, ya que las diferencias son muy evidentes con respecto al resto. En cuanto a la cepa NITR204, aunque las diferencias también fueron notables, en este caso no se dispone de un genoma bien definido como perteneciente al linaje CAS en este estudio, por lo que desde el punto de vista comparativo es difícil determinar si realmente pertenece a él. Lo que sí es evidente es que las diferencias a nivel de genoma esencial y pangenoma apuntarían a las mismas sospechas planteadas con la cepa NITR202.

Un pangenoma abierto es una peculiaridad que hace especialmente peligroso a un patógeno por la potencial capacidad de intercambio del repertorio genético, pudiendo conducir a la adquisición de nuevas resistencias o incorporar nuevos mecanismos de virulencia e infección. Además, un pangenoma abierto es un reflejo de una rápida velocidad de evolución. Si comparamos esta característica a lo visto, por ejemplo, en *M. immunogenum* veríamos los dos casos opuestos que corresponden a los dos tipos de pangenoma (abierto y cerrado). Si lo comparamos con *M. abscessus* subsp. *bolletii* y *M. abscessus*, especies de difícil diferenciación, el pangenoma sería similar al pangenoma de *M. tuberculosis*. Siendo estas dos especies, y especialmente *M. abscessus*, las MNT que más importancia tienen como patógenos oportunistas, es posible que el potencial beneficio que le aporta a *M. tuberculosis* tener un pangenoma abierto sea similar al que les aporta a estas dos especies no tuberculosas en cuanto a su capacidad patogènica. Es cierto que el pangenoma del grupo MCR es también abierto, pero no olvidemos que en este caso se comparan hasta 17 especies diferentes, por lo que en cualquier caso es un posible reflejo de la capacidad adaptativa del género, mientras que un pangenoma abierto de una especie en concreto es reflejo de su capacidad adaptativa como tal.

5. Capítulo 3: Adaptación y patogenicidad

5.1. Introducción

Una vez abordada la configuración del grupo MCR desde el punto de vista ecológico, a través de la descripción basada en los resultados del cálculo del genoma esencial y el pangenoma, centrando paulatinamente el estudio sobre el grupo formado por las especies *M. chelonae, M. abscessus, M. abscessus* subsp. *bolletii* y *M. immunogenum*, es el momento de concentrar el estudio en los aspectos clínicos de estas especies, debido a su constante y creciente implicación en infecciones nosocomiales oportunistas.

El estudio clínico del género se ha centrado tradicionalmente en especies puramente patógenas como *M. tuberculosis* o *M. leprae,* debido a los graves problemas clínicos que generan las infecciones que produce en el ser humano. Fruto de estos estudios, cada vez se ha conocido más el funcionamiento de los mecanismos de los que se sirven dichas especies para su proceso de infección y cómo, durante el proceso, despliegan toda una serie de recursos para sortear las defensas del sistema inmunológico humano, reforzando enormemente su virulencia. Además, tal y como se ha visto en los resultados del pangenoma de *M. tuberculosis*, aparentemente su capacidad para intercambiar su repertorio genético es elevada y esto la capacita para seguir desarrollando nuevas estrategias de infección. Por ello, es estrictamente necesario proseguir con los estudios para comprender mejor el funcionamiento de sus mecanismos de patogenicidad y, más importante, la identificación de nuevas y potenciales dianas terapéuticas para combatir la infección. Este último hecho es especialmente importante debido al creciente problema de las resistencias a antibióticos que se están extendiendo a gran velocidad en el mundo microbiano. Este problema es especialmente preocupante en el caso de *M. tuberculosis,* dada la complejidad de su tratamiento y el reducido espectro de antimicrobianos que se dispone contra este microorganismo.

La focalización apuntada del estudio del género en las especies patógenas ha ido en detrimento del estudio del resto de especies de micobacterias como se ha destacado anteriormente, al ser consideradas en muchos casos como microorganismos meramente ambientales. Pero este último grupo de bacterias, y en especial el de las MCR, han ido cobrando mayor importancia con el paso de los años al detectarse un incremento de infecciones nosocomiales producidas por estas micobacterias [11]. Si bien es cierto que

estas infecciones no son tan peligrosas como las provocadas, por ejemplo, por *M. tuberculosis,* siguen siendo infecciones difíciles de tratar, principalmente por la resistencia intrínseca que de forma natural tienen muchos representantes de este género a agentes biocidas como antibióticos o desinfectantes [16,106]. Por ello, al desencadenarse este tipo de infecciones en pacientes con problemas de salud de base, su situación se puede ver seriamente comprometida, hasta el punto de que pueden ser incluso mortales [5].

Las especies *M. chelonae*, *M. abscessus*, *M. abscessus* subsp. *bolletii* y *M. immunogenum* son, por definición, especies ambientales que pueden encontrarse en varios nichos ecológicos como aguas o suelos. Pero la evidente capacidad de infección que presentan ha provocado que se les considere como patógenos oportunistas significativos, bacterias que en situaciones normales no causan problemas de salud pero que, si se dan las condiciones adecuadas (como una enfermedad subyacente) son capaces de desencadenar una infección. Llegados a este punto, surge la pregunta de cómo son capaces de hacerlo, de qué mecanismos disponen, si son similares o no a los observados, por ejemplo, en *M. tuberculosis.* Responder a estas cuestiones es el objetivo principal en torno al cual gira este capítulo. Para ello se ha optado por realizar un exhaustivo estudio de los genomas de las especies de MCR secuenciadas en este trabajo. El principal propósito gira en torno a describir con la mayor rigurosidad posible los potenciales mecanismos que pueden influir en su patogenicidad (resistencias a antibióticos u otros agentes externos, factores de virulencia, proteínas reguladoras, elementos móviles y mecanismos de comunicación con el ambiente) y que las capacita para el proceso infeccioso. Evidentemente, es una descripción teórica basada en la información obtenida a partir del análisis metódico de la anotación genómica. Para ello, se confeccionará un catálogo lo más completo posible de todos estos factores, que evidentemente requerirían de su demostración experimental para su confirmación.

5.2. Material y métodos

5.2.1. Determinación del resistoma

Las potenciales resistencias a antibióticos de los genomas secuenciados se determinaron utilizando la herramienta en línea *Resistance Gene Identifier* (RGI) de la base datos *Comprehensive Resistance Data Base* (CARD) [107], utilizando la secuencia de los genomas en formato FASTA. Se aceptaron valores con E-value $\leq 10^{-5}$ y con una identidad $\geq 50\%$. Los elementos identificados se contrastaron con la anotación obtenida con Prokka v1.10 [75], además de realizar una nueva búsqueda por nombre clave en la misma para encontrar potenciales elementos no contemplados en el análisis. Las resistencias a otros elementos biocidas, como metales pesados, se buscaron mediante la prospección directa de las anotaciones de los genomas. Basándose en protocolos previamente descritos [108], la confirmación de producción de metalo-β-lactamasas (MBL) se realizó mediante la obtención de un cultivo confluyente sobre placas de R2A por duplicado a partir de una suspensión McFarland 0,5 en suero fisiológico. En cada placa se dispuso un disco de imipenem de 10 μg (Bio-Rad, Hercules, California, EEUU) y en uno de los duplicados el disco se suplementó con 1 mg de EDTA a partir de una solución de EDTA 0,5 M a pH 8. Las placas se incubaron a 30 ºC durante 4 días, midiendo posteriormente los respectivos halos de inhibición.

5.2.2. Determinación de los factores de virulencia

La determinación de los potenciales factores de virulencia se realizó utilizando la base de datos *Virulence Factor Data Base* (VFDB) [109]. A partir de la VFDB se descargó el archivo multi-FASTA (archivo con múltiples secuencias en formato FASTA) de secuencias proteicas correspondiente al set A (2.585 secuencias de ADN y sus respectivas proteínas curadas manualmente). Con este conjunto de secuencias se creó una base de datos local curada para realizar las correspondientes búsquedas con BLAST, implementado en UGENE v1.16.1 [110], utilizando como consulta las secuencias de proteínas de los archivos multi-FASTA propios de cada genoma. Se aceptaron como válidos aquellos resultados con un E-value $\leq 10^{-5}$, siendo estos contrastados posteriormente con la anotación.

5.2.3. Determinación del reguloma

La determinación de las proteínas reguladoras codificadas por un genoma (reguloma) se realizó con el protocolo en línea P2RP (del inglés *Predicted Prokaryotic Regulatory Proteins*) [111], utilizando en cada caso un archivo multi-FASTA incluyendo las secuencias de todas las proteínas codificadas por cada genoma. La identificación de proteínas reguladoras se basa en el análisis de dominios con el programa RPSBLAST. P2RP utiliza para ello una base de datos obtenida a partir de dominios seleccionados manualmente de las bases de datos Pfam [112] y SMART [113]. Basándose en parámetros de similitud, así como la arquitectura que presentan los dominios sobre las proteínas, estas son asignadas a una familia determinada y clasificadas como factores de transcripción (FT), reguladores de respuesta (RR) o factores sigma (FS); identificándose además los sistemas de dos componentes (SDC) formados por una histidina quinasa (HK) y un RR. Para reducir el número de falsos positivos se realizó un post análisis en el que se eliminaron proteínas erróneamente clasificadas o se las asignó al grupo denominado "Otras Proteínas de unión al ADN" (OPA).

5.2.4. Estudio del mobiloma (GI, Integrasas, transposasas y profagos)

La catalogación de los elementos móviles de los genomas se realizó por partes. La determinación de las islas genómicas se realizó analizando el genoma en IslandViewer [114,115], el cual utiliza tres algoritmos distintos (SIGI-HMM, IslandPick y IslandPath-DIMOB), y siguiendo las recomendaciones de la herramienta. La presencia de integrasas y transposasas se determinó a través de la prospección directa de la anotación de los genomas. La presencia de potenciales profagos se determinó utilizando la herramienta PHAST (del inglés *PHAgue Search Tool*) [116].

5.2.5. Estudio de los elementos implicados en la percepción del Quórum o *Quorum Sensing* (QS)

Los archivos FASTA de proteínas obtenidos durante la anotación de los respectivos genomas se utilizaron para rastrear la base de datos KEGG (http://www.genome.jp/kegg) con ayuda de BLASTKOALA [117]. A partir de la clasificación obtenida, se seleccionaron manualmente aquellas proteínas incluidas en la categoría “Percepción del

Quórum" (QS). Estas proteínas fueron, a su vez, contrastadas con la anotación y analizadas en la base de datos STRING [118,119] con el fin de proceder a la identificación de elementos colindantes y potencialmente relacionados con ellos, especialmente en términos de coexpresión.

5.3. Resultados

5.3.1. Resistoma

El análisis de elementos potencialmente implicados en la resistencia a antibióticos en los genomas secuenciados dentro del grupo *abscessus-chelonae-immunogenum*, permitió la identificación desde un punto de vista genético de una gran variedad de posibles mecanismos de resistencias (Tabla 5.1). Adicionalmente en la cepa tipo de *M. abscessus* subsp. *bolletii* se encontró el gen *erm*41, el cual proporcionaría el fenotipo MLSb (Macrolidos-Licosamida-Streptogramina B resistente). Concretamente, Erm41 se ha descrito en la bibliografía como la proteína responsable de aportar resistencia inducida a macrólidos en *M. abscessus* [120]. Además, tal como se puede apreciar en la Tabla 5.2, la simple búsqueda directa sobre la anotación de los genomas permitió el hallazgo de lactamasas de subclase B, es decir, MBL.

Con la finalidad de catalogar los diferentes tipos de MBL detectados en los genomas objeto de estudio, se realizó un análisis comparativo con todas las secuencias proteicas derivadas de estos genes detectadas en los genomas secuenciados en la presente tesis, así como todas las secuencias de MBL incluidas en el subgrupo de proteínas ARO_3000004 (del inglés *Antibiotic Resistance Ontology*) de la base de datos CARD, donde se incluyen todas las secuencias de MBL contempladas en dicha base de datos (Figura 5.1). En dicha figura resulta evidente que las MBL procedentes de MCR se englobaron en tres grandes grupos bien definidos. Además, los elevados valores de bootstrap obtenidos refuerzan su validez. El primer grupo, y a excepción de las MBL obtenidas de bases de datos como la del genoma de *M. immunogenum* SMUC14 y otra procedente de *M. immunogenum* hallada en la base de datos UniProt, apareció constituido exclusivamente por MBL halladas en el actual estudio. El segundo gran grupo apareció como una rama independiente formada exclusivamente por MBL procedentes de las MCR secuenciadas

en el actual trabajo, siendo una de las MBL el de la cepa tipo de *M. llatzerense*, sustancialmente diferente al resto.

Tabla 5.1. Inventario de los elementos de resistencia a antibióticos identificados por RGI. Se indica el tipo de antibiótico contra el que se han encontrado resistencias, así como el número de proteínas detectadas en cada caso, tanto para las cepas tipo (^{T}) como para el resto de cepas.

	CCUG 47445^T	**CCUG 47286^T**	**CCUG 50184^T**	**MG2**	**MG8**	**MHSD2**	**MHSD3**	**CR-UIB1**
Aminoglicosidos	1	2	1	1	1	1	1	1
β-lactámicos	6	3	2	4	4	5	3	4
Isoniacida	1	2	2	2	2	2	2	2
Péptidos antimicrobianos	1	1	2	1	1	1	1	1
Fluoroquinolonas	1	1	1	1	1	1	1	1
Trimetoprima	1	1	1	1	1	1	1	1
Bombas de expulsión	14	13	13	13	13	13	13	13
Etambutol	1	1	1	1	1	1	0	0
Glicopéptidos	1	1	1	1	1	1	2	1
Rifampicina	2	2	3	3	3	2	3	2
Pirazinamida	0	0	1	0	0	0	0	1
Daptomicina	0	0	1	0	0	1	0	1
Daunorubicina	0	4	3	3	3	3	3	3
Ethionamida	0	0	0	0	0	0	0	0
Bleomicina	0	1	1	2	2	1	2	2
Polimixina	0	0	0	0	0	0	0	0
Bacitracina	0	0	0	0	0	0	0	0
Fosmidomicina	0	1	0	0	0	1	0	1
Biciclomicina	0	1	1	1	1	1	2	1

.

Tabla 5.2. Potenciales MBL identificadas en los genomas secuenciados. Se indica el número de genes encontrados en cada caso.

	CCUG 47445^T	**CCUG 47286^T**	**CCUG 50184^T**	**MG2**	**MG8**	**MHSD2**	**MHSD3**	**CR-UIB1**
Superfamilia MBL	1	0	2	1	1	2	1	2
MBL precuror L1	0	1	0	1	1	0	1	0

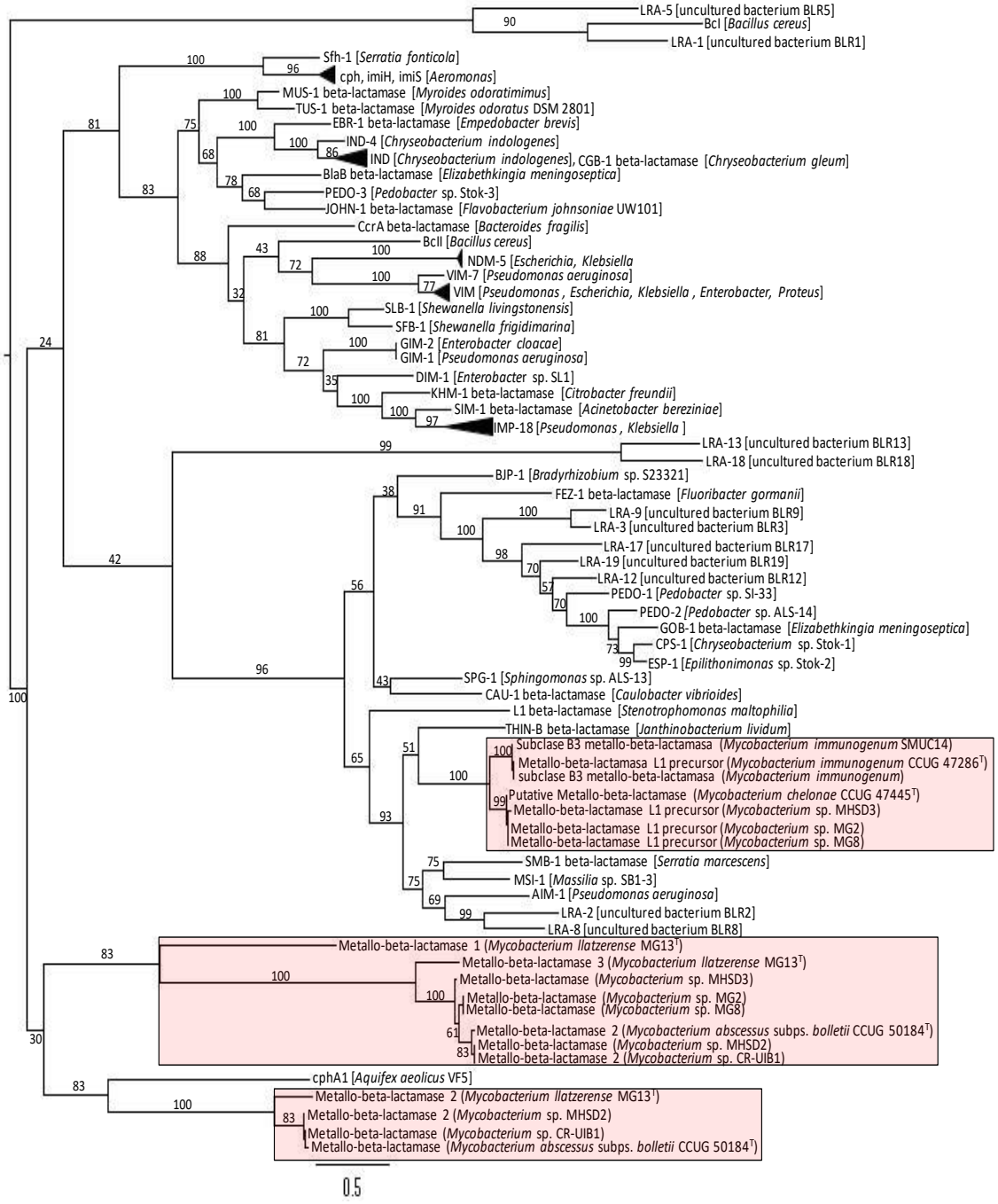


Figura 5.1. Representación de las diferentes agrupaciones de MBL por comparación de secuencias, obtenidas con el algoritmo MLE (bootstrap=100).

Utilizando como control positivo una cepa de la especie *Pseudomonas monteilii* que presenta una MBL, se comprobó si estas cepas eran realmente productoras de MBL mediante la inhibición de su acción con EDTA. Los resultados obtenidos al medir los halos de inhibición demostraron que, en todos los casos, aparentemente si hay producción de este tipo de enzimas, ya que su sensibilidad al imipenem aumenta en mayor o menor medida dependiendo de la cepa (Tabla 5.3).

Tabla 5.3. Diámetro de los halos de inhibición frente a Imipenem (Imp) en ausencia y presencia de EDTA. Se indica el porcentaje de incremento de halo observado en cada caso.

Cepa	Ø Imp (mm)	Ø Imp + 1 mg EDTA (mm)	Incremento de halo (%)
P. monteilii	0	30	30 %
CCUG 47445^{T}	50	66	32 %
CCUG 47286^{T}	21	55	162 %
CCUG 50184^{T}	36	48	33,3 %
MG2	30	37	23,3 %
MG8	26	35	34,6 %
MHSD2	35	50	42,8 %
MHSD3	20	25	25 %
CR-UIB1	37	55	48,6 %

Las cepas MG2 y MHSD3 fueron las que presentaron el porcentaje de incremento de sensibilidad más bajo según los halos de inhibición determinados. Por otra parte, la cepa tipo de *M. chelonae* es la que mayor diámetro de halo mostró, incluso sin suplementar el disco con EDTA. El efecto de inhibición más evidente de la acción de las MBL producida por la acción del EDTA se observó en *M. immunogenum* CCUG 47286^{T} (Figura 5.2). Además, también se encontraron genes implicados en fenotipos de resistencia a metales pesados, así como a otros agentes biocidas (Tabla 5.4).

Finalmente, destacar que en la cepa tipo de *M. chelonae* se encontraron 6 componentes que corresponderían a una bomba de cationes implicada en la homeostasis del pH, un gen de resistencia a feomicina y otro para sulfonamida, así como dos elementos potencialmente implicados en la resistencia al arsénico y al alcanfor.

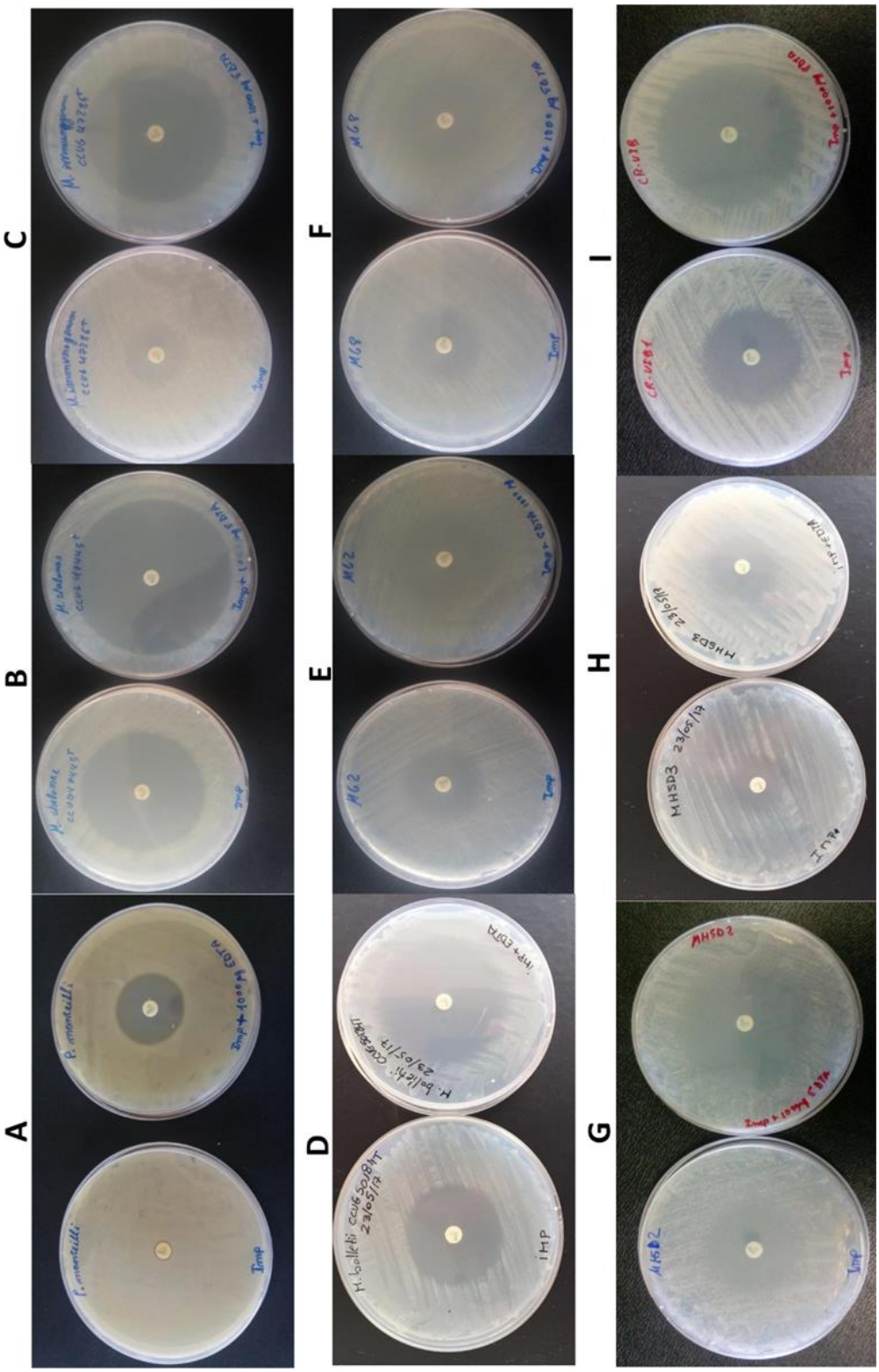


Figura 5.2. Ensayo experimental de la inhibición de MBL por EDTA en A) *P. monteilli* (control positivo), B) *M. chelonae* CCUG 47445^{T}, C) *M. immunogenum* CCUG 47286^{T}, D) *M. abscessus* subsp. *bolletii* CCUG 50184^{T}, E) *Mycobacterium* sp. MG2, F) *Mycobacterium* sp. MG8, G) *Mycobacterium* sp. MHSD2, H) *Mycobacterium* sp. MHSD3, I) *Mycobacterium* sp. CR-UIB1.

Tabla 5.4. Número de proteínas relacionadas con la resistencia a otros elementos externos diferentes a antibióticos representadas en el proteoma de las cepas estudiadas.

	CCUG 47445^T	CCUG 47286^T	CCUG 50184^T	MG2	MG8	MHSD2	MHSD3	CR-UIB1
Cobre	3	2	1	1	1	1	1	1
Tellurio	2	0	1	1	1	1	1	1
Cadmio	0	1	0	0	0	0	0	0
Mercurio	0	1	1	0	0	1	0	1
Cobalto-Zinc-Cadmio	0	1	0	1	1	0	1	0
Hidroperóxido orgánico	1	3	3	2	2	3	2	3
6-N-Hidroxilaminopurina	1	1	1	1	1	1	1	0
Ácido Furásico	0	2	0	1	0	0	1	0
Amonio cuaternario	1	1	2	1	1	1	1	1
PM res. acido	1	2	2	1	2	2	1	1
Antisépticos	3	2	1	4	2	2	3	2

5.3.2. Factores de virulencia

Los resultados de la búsqueda de genes implicados en la virulencia se recogen en la Tabla 5.5. En dicha Tabla se refleja la presencia/ausencia de un determinado factor de virulencia en la lista final que se obtuvo después del cribado y la comparación con la anotación de los respectivos genomas para confirmar la presencia de cada gen, así como la presencia o ausencia de genes relacionados en aquellos casos de factores de virulencia que incluyen un conjunto de elementos necesarios para su funcionamiento. Este es el caso del sistema de secreción tipo VII ESX-3, constituido por los genes *ecc*A3, *ecc*B3, *ecc*C3, eccD3, *ecc*E3, *esp*G3, *esx*G, *esx*H, *myc*P3, PE5, PPE4; los genes que conforman el complejo antígeno 85 (*fbp*A, *fbp*B, *fbp*C, *fbp*C1); o los genes implicados en la constitución de sideróforos (*mbt*A, *mbt*B, *mbt*C, *mbt*D, *mbt*E, *mbt*F, *mbt*G, *mbt*H). En este análisis se utilizó la información derivada de las bases de datos KEGG y GeneBank con el fin de confirmar, desde el punto de vista genómico, los diferentes elementos detectados.

Los elementos relacionados con los sistemas de secreción tipo VII se encontraron representados en todos los genomas analizados y responden a una misma organización (Figura 5.3), por lo menos en aquellos genomas completos o más continuos. En el caso de genomas incompletos como, por ejemplo, *Mycobacterium* sp. MG2 o *Mycobacterium* sp. MG8 algunos elementos no se encontraron presentes (Figura 5.3A). Aunque en la anotación los elementos se describen como pertenecientes al sistema ESX-1, se encontraron en una organización que respondería a la del sistema de secreción ESX-3 de *M. smegmatis*. En esta organización se hallaron una proteasa, que corresponde al gen *myc*P3, y dos genes codificantes para proteínas hipotéticas que se identificaron por UniProt como las proteínas EccE3 y EsxS, ambas características de estos sistemas. Además, en cuatro de los genomas también se encontró un homólogo del gen *myc*P5, el cual se localizó asociado con otros genes anotados como componentes de este tipo de sistemas y siguiendo un patrón análogo en los 4 genomas en cuanto a su ordenación (Figura 5.3B). Aunque los análisis con KEGG no dieron ningún resultado, las búsquedas con BLAST en GenBank confirmaron su posible relación con los sistemas de secreción.

En cuanto al metabolismo del hierro, se hallaron toda una serie de genes implicados en este proceso. En primer lugar, los genes *mbt* constituyen importantes sistemas de internalización de hierro en la célula (sideróforos). En el género *Mycobacterium* se pueden encontrar dos tipos*:* las micobactinas y las exoquelinas. La organización de esos genes no se encontró tan conservada como en los ejemplos mencionados anteriormente, siendo especialmente diferente entre las cepas *M. chelonae* CCUG 47445^{T} y *M. immunogenum* CCUG 47286^{T} (Figura 5.4). En todos los casos se encontró, además, por lo menos uno de los genes *irt*A o *irt*B o ambos asociados a dichos sideróforos. Los genes *irt* constituyen transportadores regulados por hierro. Algunos representantes de los genes *mbt* como *mbtC*, *mbt*B y *mbt*N aparecieron alejados en el cromosoma del conjunto y no están presentes en todos los genomas. Por último, en todos los casos también se identificó un importante regulador, codificado por el gen *ide*R, y que estaría implicado en la modulación de toda una serie de genes dependientes de los niveles de hierro, entre ellos los genes relacionados con la síntesis de micobactinas comentados anteriormente [121].

Por su parte, el conjunto de genes *fbp*ABC (del inglés *fibronectin-binding protein*) apareció en todos los genomas estudiados y en una misma configuración (Figura 5.5A).

En todos los genomas analizados se detectaron también los genes de *ure*ABG. Además, al revisar la anotación de estos genomas se pudo comprobar la presencia de los genes *ure*ABCFGD formando un único grupo en todos ellos (Figura 5.5B). Otro grupo de genes anotados como factores de virulencia y puestos de manifiesto en las cepas analizadas es el constituido por *phz*E1, *phz*D1 y *phz*C1. La presencia y la organización, apareciendo los mismos elementos asociados tanto corriente arriba como corriente abajo, es análoga en todas las MCR analizadas (Figura 5.5C). Asimismo, los tres genes fueron asignados por KEGG a la ruta de síntesis de este tipo de compuestos.

El conjunto *pho*P-*pho*R consituye un SDC que se encontró en todos los genomas. Este sistema respondería a una situación de privación de Mg^{2+} y es capaz de modular la expresión de toda una serie de factores de virulencia [122].

Por último, en lo que respecta a proteínas de membrana, se encontró la proteína codificada por el gen *hbh*A. A esta proteína se le supone una importante función de adherencia a células del hospedador en patógenos importantes como la propia *M. tuberculosis* [123].

En el contexto de la capacidad de respuesta frente a diferentes condiciones de estrés, para favorecer la supervivencia celular, se encontraron diversos representantes. En primer lugar, los genes *clp*P y *clp*C hallados en todos los genomas codificarían para serina proteasas (hidrolasas que degradan enlaces peptídicos de péptidos y proteínas y que poseen en su centro activo un aminoácido serina esencial) que responden a condiciones de estrés. Al contrastar la anotación con el resultado de BLAST obtenido, *clp*P apareció anotado como la subunidad 1 de una serina proteasa dependiente de ATP. Junto a este gen se encontró anotada la subunidad 2 y el gen de la chaperona acompañante *clp*X (Figura 5.5D). Por su parte, ClpC correspondió al gen *clp*C1, descrito como la subunidad de unión al ATP de una proteasa Clp. En segundo lugar, otros factores de virulencia activados por estrés hallados fueron LipF, una lipasa cuya expresión esta modulada por niveles bajos de pH [124] y KatA, una catalasa de gran importancia para la supervivencia intracelular de microorganismos como *Legionella pneumophila* [125]. Otro factor de virulencia encontrado en todos los casos hace referencia a RelA, o PPGPP sintasa, una proteína implicada en la síntesis de esta alarmona (molécula de señalización intracelular), cuya presencia modula la expresión de toda una serie de genes implicados en la

supervivencia a largo plazo en situaciones de falta de nutrientes en el medio [126]; y la isocitrato liasa , o *icl*, cuya acción permite a la célula utilizar ácidos grasos como fuente de carbono alternativa [127].

Por su parte, el gen que codifica para el factor de virulencia CtrD se ha descrito como parte de una agrupación de genes dividida en cinco bloques (A, B, C, D, D') implicados en la formación de cápsula. Concretamente se encuentra en la región C, donde los genes *ctr* codificados están implicados en el transporte de polisacáridos [128]. Aparentes homólogos de este factor de virulencia fueron detectados en *M. chelonae* CCUG 47445^{T}, *Mycobacterium* sp. MG8 y *Mycobacterium* sp. MHSD3. Sin embargo, el análisis en KEGG de esta proteína y de las proteínas adyacentes las identificó como un sistema de transporte completo de aminoácidos de cadena ramificada y no de polisacáridos, desempeñando una función que según dicha base de datos podría estar más bien relacionada con la QS.

Por último, en todos los casos analizados se encontraron representantes de lo que se conoce como operones MCE (del inglés *Mammalian Cell Entry proteins*) y que estarían constituidos por un conjunto de genes *mce* precedidos de dos genes que codifican para proteínas asociadas a membrana. Los homólogos de estas proteínas se hallaron organizados en operones en todos los genomas estudiados (Figura 5.6). Los genomas de *M. chelonae* CR-UIB1, *M. immunogenum* CCUG 47286^{T}, *M. abscessus* subsp. *bolletii* CCUG 50184^{T} y *Mycobacterium* sp. MG8 presentaron siete operones completos cada uno, nueve en el caso de *Mycobacterium* sp. MHSD3 y *M. abscessus* subsp. *bolletii* CCUG 50184^{T}, ocho en *Mycobacterium* sp. MG2 y hasta cinco en el genoma de *Mycobacterium* sp. MHSD2. En algunos de estos operones algunas de las proteínas constituyentes se catalogaron como hipotéticas, que al ser analizadas con BLAST en UniProt se correspondieron con la proteína esperada en ese lugar en comparación con otros operones. En todos los genomas aparece, además, un conjunto de genes *mce* precedidos por una acil coenzima A (*acil-CoA*) sintetasa (FadD13) y una enoil-*CoA* hidratasa. En las cepas CR-UIB1, MG8 y MHSD2 se observó un operón truncado en el extremo de un *contig*. En el caso de la cepa tipo de *M. immunogenum* se halló un conjunto de dos genes *mce* precedidos de un único transportador de membrana.

Tabla 5.5. Factores de virulencia encontrados a partir de los proteomas de las cepas estudiadas. Se indica la presencia (+) o ausencia (-) de los factores encontrados en cada caso. Se destaca la ausencia en algunas agrupaciones de los componentes A) eccD3, B) mbtE y C) irtA.

	CCUG 47445^T	CCUG 47286^T	CCUG 50184^T	MG2	MG8	MHSD2	MHSD3	CRUIB1
fbpABC	+	+	+	+	+	+	+	+
ureABG	+	+	+	+	+	+	+	+
phzE1D1C1	+	+	+	+	+	+	+	+
mycP3	+	+	+	-	-	-	-	+
eccA3-B3-C3-D3	+	+	+	+	+	+	+	+
espG3	+	+	+	+	+	+	+	-
es+HG	+	+	+	+	+	+	+	-
PE5	+	+	+	+	+	+	+	+
mbtAEGHI	+	+	+^{B}	+^{B}	+^{B}	+^{B}	+	+^{B}
irtAB	+	+	+	+	+^{C}	+^{C}	+	+
mbtC	-	+	-	-	+	-	-	-
mbtB	-	+	+	+	+	+	+	+
mbtN	-	-	-	-	-	-	+	-
phoP-phoH	+	+	+	+	+	+	+	+
hbhA	+	+	+	+	+	+	+	-
htpB	+	+	+	+	+	+	+	+
clpC	+	+	+	+	+	+	+	+
clpP	+	+	+	+	+	+	+	+
lipF	+	+	+	+	+	+	+	+
katA	+	+	+	+	+	+	+	-
ctrD	+	-	-	-	+	-	+	-
relA	+	+	+	+	+	+	+	+
ideR	+	+	+	+	+	+	+	+
icl	+	+	+	+	+	+	+	+
mycP5	-	-	+	+	+	-	-	+

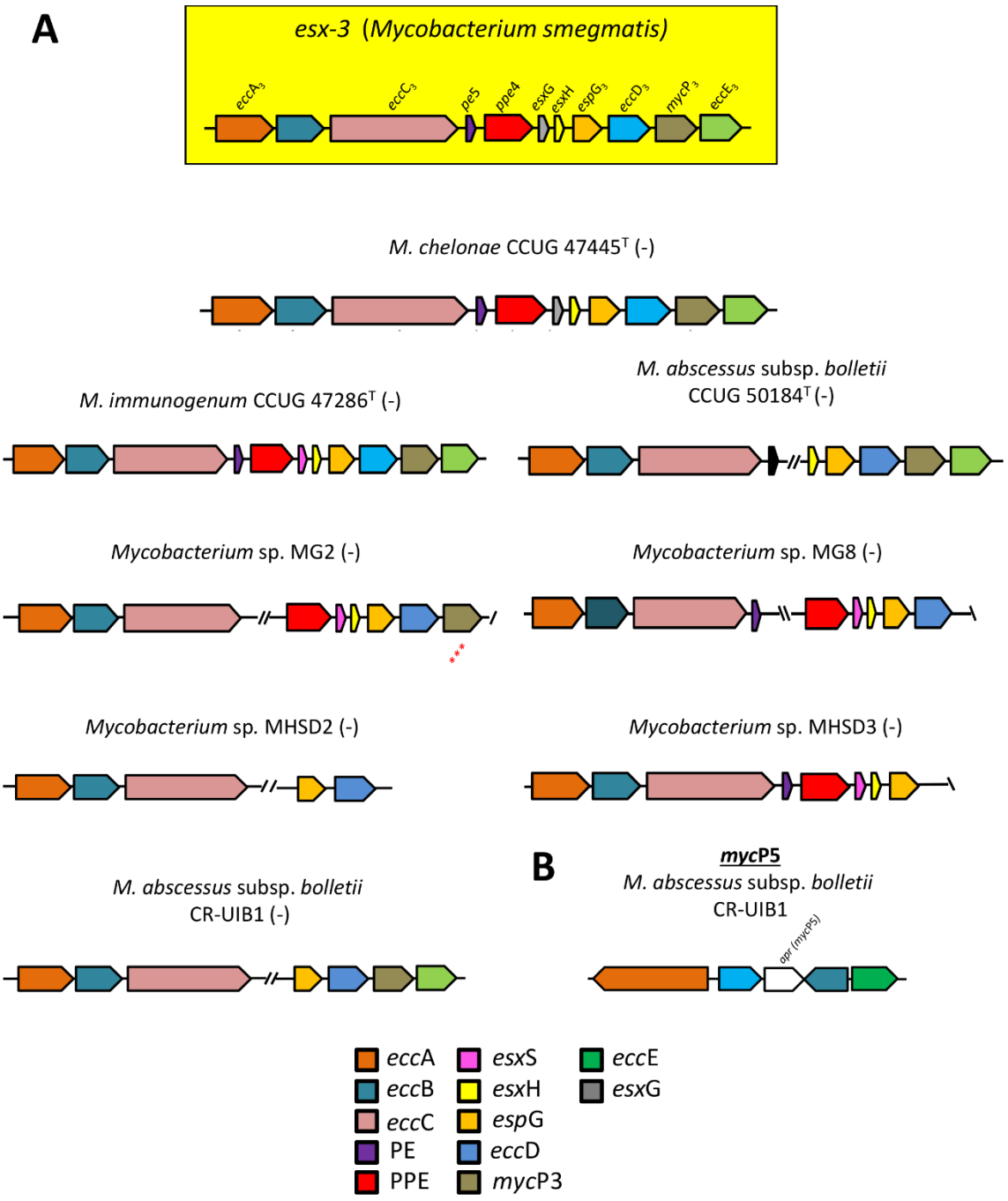


Figura 5.3. A) Sintenia del sistema de secreción tipo VII esx-3 entre las distintas cepas, utilizando como modelo el sistema de *M. smegmatis*. Entre paréntesis se indica la orientación encontrada en el respectivo genoma (-). B) Proteasa mycP$_5$ y los genes adyacentes.

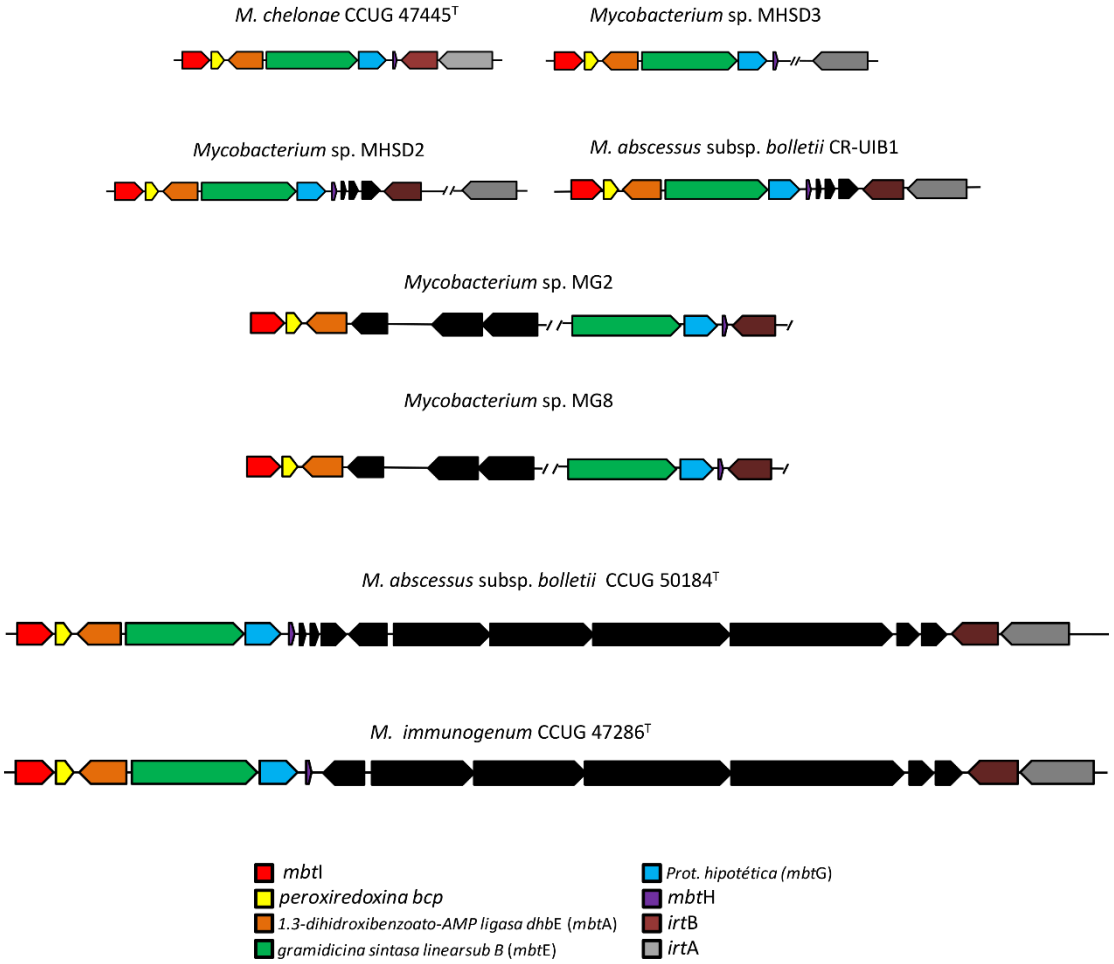


Figura 5.4. Sintenia resultante entre los sistemas de captación de hierro encontrados en las distintas cepas. La orientación reflejada en la figura es la encontrada en los respectivos genomas.

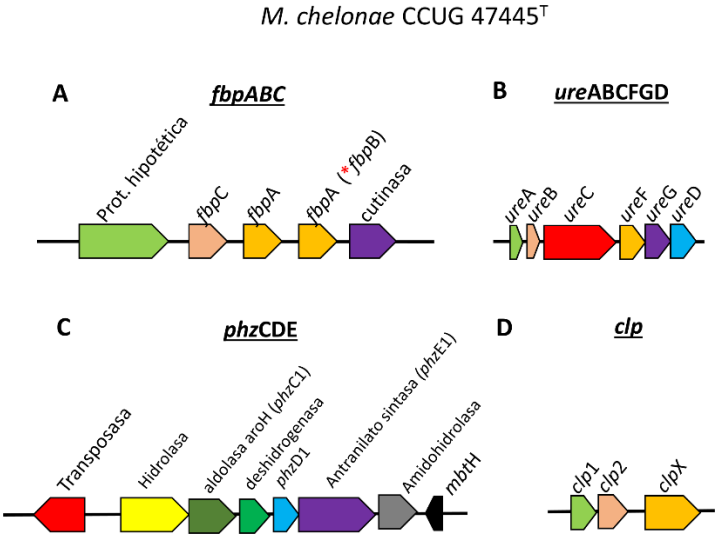


Figura 5.5. Factores de virulencia hallados en *Mycobacterium chelonae* CCUG 47445^T con la misma organización en las diferentes cepas: A) antígeno 85 (*fbp*ABC), B) subunidades de la enzima ureasa, C) elementos implicados en la sisntesis de fenazinas y D) subunidades de la proteasa *clp*.

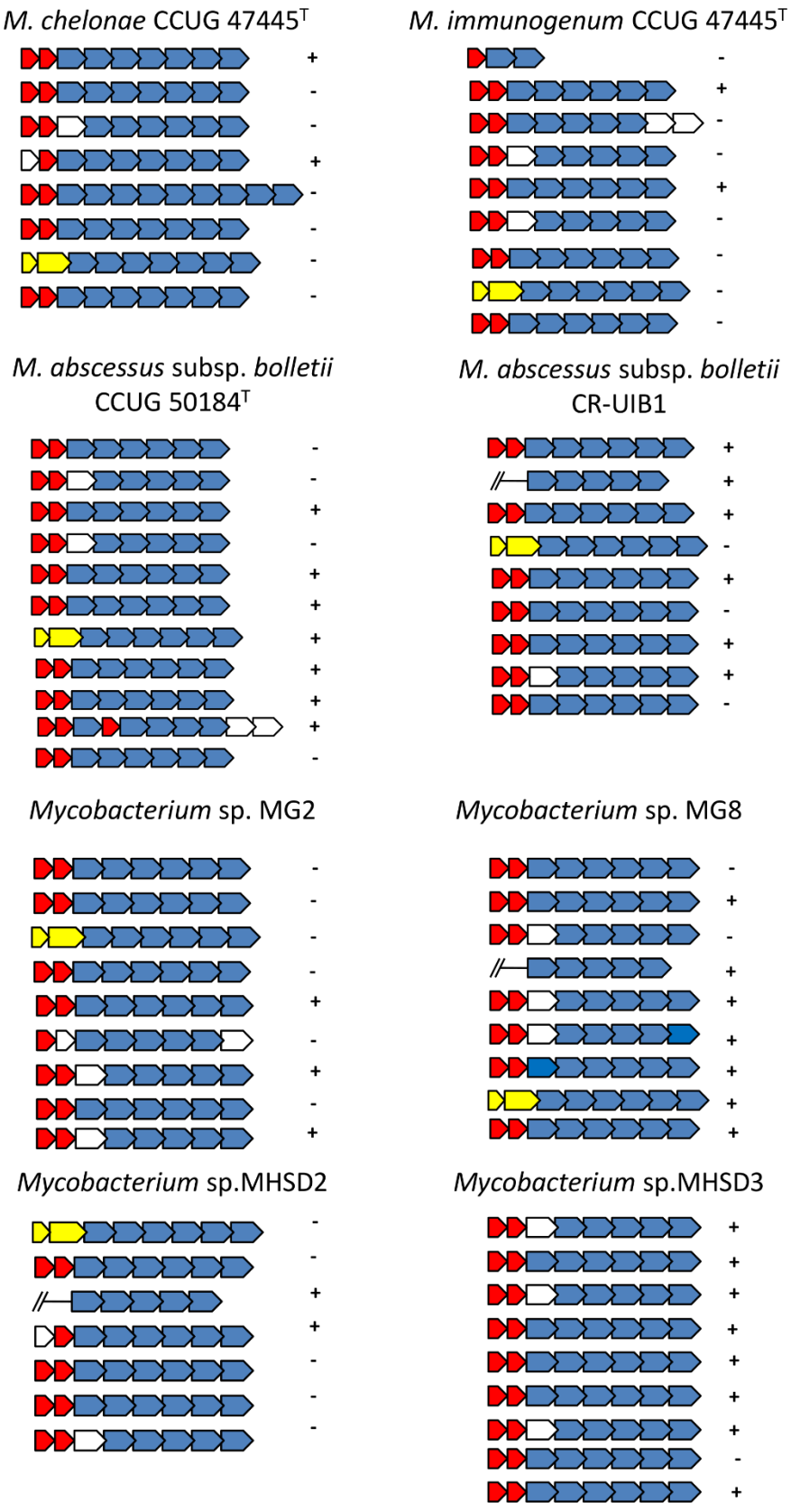


Figura 5.6. Organización de los operones *mce* encontrados en las cepas estudiadas. Se puede observar la disposición de los transportadores de membrana (rojo), genes *mce* (azul), proteínas hipotéticas (blanco) y genes de otro tipo (amarillo).

5.3.3. Reguloma

Factores de transcripción

El número de factores de transcripción (FT) hallados en los genomas analizados osciló entre 323 y 393, siendo *M. immunogenum* CCUG 47286^T y *Mycobacterium* sp. MG2 las cepas que respectivamente presentaron mayor y menor número. Los factores de transcripción hallados se clasificaron en reguladores transcripcionales (RT) (Tabla 5.6), reguladores de respuesta (RR) (Tabla 5.7) y factores sigma (FS) (Tabla 5.8). El mismo análisis se aplicó sobre las cepas de *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv detectando en estos casos la mitad de proteínas reguladoras en total, aproximadamente (Tablas suplementarias 2, 3, 4, 5 y 6, Anexo 2).

En el caso de los RT, se detectaron un total de 35 familias diferentes repartidas entre todos los genomas, así como una serie de reguladores no clasificados. Las familias potencialmente relacionadas con la patogenicidad del microorganismo se incluyen en la Tabla 5.6.. El listado completo de todas las familias de RT encontrados se recogen en la Tabla suplementaria 1 del Anexo 2. En general, las familias detectadas suelen incluir representantes muy diversos y que pueden estar implicados en la regulación de procesos muy dispares. La familia de RT más abundante en los genomas estudiados fue TetR, con un número comprendido entre 138 y 155 FT. Por su parte, se encontraron hasta tres tipos de RR (Tabla 5.7), siendo el correspondiente a la familia IclR el menos distribuido, mientras que la familia OmpR no sólo resultó estar presente en todos los genomas, sino que además fue el más abundante en cada uno de ellos

Factores Sigma

En lo que se refiere al catálogo de FS presentes en las cepas (Tabla 5.8), a excepción del FS SigF, SigK y SigX; el resto de FS identificados se hallaron en todas las cepas, detectándose 13 ó 14 tipos de FS en cada caso. Además, cabe mencionar la presencia de FS sin clasificar, así como de los elementos reguladores (Anti-Sig) de SigF, SigE, SigM, SigD, SigI, SigK y SigH (Tabla 5.9). Todos los FS identificados pertenecieron a la familia Sigma-70 (σ^{70}), la cual se divide en cuatro grandes grupos en función del grado de conservación de las cuatro regiones clave de para su funcionamiento. SigA pertenece al grupo 1 (FS esenciales), SigB pertenece al grupo 2, SigF al grupo 3, y el resto de FS

estarían incluidos en el grupo 4, en una subfamilia caracterizada por ser muy diversa y englobar los FS con función extracitoplasmática (ECF, del inglés *Extra Cytoplasmátic Function*). El número total de FS hallados en *M. tuberculosis* oscila (tanto en la cepa tipo como en la cepa CR-UIB2 como en la cepa H37Rv) en torno a los 13 (Tabla suplementaria 4, Anexo 2), mientras que en los genomas de MCR estudiados la cantidad global es algo superior, oscilando entre 18 y 20 FS. En la mayoría de los casos se hallaron representantes de los distintos tipos de FS en todos los genomas, a excepción de SigF, SigK y SigX.

Tabla 5.6. Familias de reguladores transcripcionales potencialmente relacionados con la patogenicidad encontradas en los diferentes genomas analizados. Se indica el número de representantes encontrados en cada caso.

	CCUG 47445^T	**CCUG 47286^T**	**CCUG 50184^T**	**MG2**	**MG8**	**MHSD2**	**MHSD3**	**CR-UIB1**
AraC	13	22	17	13	14	19	14	20
ArsR	14	11	13	14	14	13	17	14
AsnC	4	4	4	5	5	4	5	5
Crp	1	1	2	1	1	2	1	3
DtxR	1	2	2	2	2	2	1	2
FeoC	0	0	0	0	0	0	0	1
Fur	3	2	3	3	3	2	3	2
GntR	12	15	14	10	11	16	13	16
HrcA	1	1	1	1	1	1	1	1
IclR	7	9	9	6	6	9	7	14
LexA	1	1	1	1	1	1	1	1
LuxR	3	2	2	3	3	2	2	2
LysR	16	22	21	14	15	16	17	21
MarR	21	27	21	19	19	21	20	21
MerR	8	8	7	8	8	6	7	7
Mga	0	0	0	0	0	0	0	1
Rrf2	2	1	1	1	1	1	2	2
TetR	144	155	149	143	144	146	145	138

Tabla 5.7. Familias de reguladores de respuesta encontrados en los genomas analizados. Se indica el número de representantes encontrados en cada caso.

	IclR	NarL	OmpR	Total
CCUG 47445^T	1	6	9	16
CCUG 47286^T	0	7	10	17
CCUG 50184^T	0	6	11	17
MG2	1	6	8	15
MG8	1	6	8	15
MHSD2	0	7	10	17
MHSD3	1	6	9	16
CR-UIB1	0	5	11	16

Tabla 5.8. Factores sigma identificados en el análisis del reguloma de los genomas secuenciados. Se indica el número de representantes encontrados en cada caso.

	CCUG 47445^T	CCUG 47286^T	CCUG 50184^T	MG2	MG8	MHSD2	MHSD3	CR-UIB1
SigA	1	1	1	1	1	1	1	1
SigB	1	1	1	1	1	1	1	1
SigC	1	1	1	1	1	1	1	1
SigD	1	1	1	1	1	1	1	1
SigE	2	2	2	2	2	2	2	2
SigF	1	0	1	1	1	1	1	1
SigG	1	1	1	1	1	1	1	1
SigH	1	2	1	1	1	1	1	1
SigI	1	1	1	2	2	1	1	1
SigJ	3	3	4	3	3	4	3	4
SigK	1	1	0	1	1	0	1	0
SigL	1	1	1	1	1	1	1	1
SigM	1	2	1	1	1	1	1	1
SigX	0	0	1	1	1	1	0	1
Sin clasificar	3	2	2	1	1	2	2	3

Tabla 5.9. Reguladores negativos (Factores antisigma) identificados en la prospección de los proteomas. Se indica la presencia (+) o ausencia (-) de cada factor en los distintos genomas.

	CCUG 47445^T	CCUG 47286^T	CCUG 50184^T	MG2	MG8	MHSD2	MHSD3	CR-UIB1
Anti-SigH(RshA)	+	+	+	+	+	+	+	+
Anti-SigF(RsbW)	-	+	+	+	+	+	+	+
Anti-SigE(RseA)	+	+	+	+	+	+	+	+
Anti-SigM(RsmA)	+	+	+	+	+	+	+	+
Anti-SigD(RsdA)	+	+	+	+	-	+	+	+
Anti-SigL(RslA)	+	+	+	+	+	+	+	+
Anti-SigK(RskA)	+	+	-	+	+	-	+	-

Sistemas de dos componentes

Todos los RR identificados anteriormente se hallaron asociados a histidina quinasas (HK), formando sistemas de dos componentes (SDC) (Tabla 5.10).

Tabla 5.10. Número de elementos relacionados con SDC encontrados en los genomas analizados. Se indica el número de histidina quinasas, reguladores de respuesta y el número de SDC completos formados entre ellos. Se indica también el número de elementos para los cuales no se ha hallado el elemento relacionado que completaría el SDC.

	CCUG 47445^T	CCUG 47286^T	CCUG 50184^T	MG2	MG8	MHSD2	MHSD3	CR-UIB1
HK	17	19	18	17	17	17	16	17
RR	18	18	18	17	17	18	18	18
TCS	16	17	17	15	15	16	15	15
HK solitarias	1	2	1	2	2	1	1	2
RR solitarios	2	1	1	2	2	2	3	3
Prot. fosfotransferasa	0	0	1	1	1	1	0	1

Además, en la Tabla 5.10 se incluyen dos tipos de reguladores no asociados a una HK: los de tipo Amir_NasR (presentes en todas las cepas) y los de tipo CheY (presentes en todas las cepas excepto en CCUG 47286^T, CCUG 50184^T y MHSD2). En las cepas

MHSD2, MHSD3 y CR-UIB1 también apareció un RR del tipo OmpR sin una HK aparentemente asociada. De igual manera, siempre según la anotación obtenida aparecieron algunas HK que no parecen estar asociadas a RR.

5.3.4. Elementos móviles

La búsqueda de potenciales islas genómicas dio como resultado una serie de regiones que son compatibles con estos elementos (Tabla 5.11), de las cuales la posición exacta en el genoma sólo se pudo concretar en los genomas completos de las cepas tipo de *M. chelonae* y *M. immunogenum* (Figura 5.7). Sin embargo, ninguna de ellas pudo ser asociada a una isla de patogenicidad. Consultada la base de datos PAI-DB, donde pueden encontrarse islas genómicas o de patogenicidad definidas para diferentes microorganismos, se comprobó que en el caso de la cepa tipo *M. tuberculosis* H37Rv tan solo presenta dos potenciales islas genómicas de pequeño tamaño. En este sentido, si un patógeno bien definido del género *Mycobacterium* no parece tener islas de patogenicidad bien definidas, o una gran abundancia de ellas, se podría extrapolar que en los genomas de micobacterias ambientales es posible que tampoco se encuentren.

Profundizando en la búsqueda de elementos móviles, se detectaron, a través de la anotación del genoma, solamente dos integrasas en la cepa de origen clínico *Mycobacterium* sp. CR-UIB1. En lo que respecta a la detección de regiones relacionadas con potenciales profagos, mediante PHAST fueron clasificadas en tres categorías: intactos, incompletos y profagos dudosos. Así, se detectaron profagos clasificados como intactos en todos los genomas, a excepción de *Mycobacterium* sp. MHSD2 y *M. abscessus* subsp. *bolletii* CCUG 50184^{T} (Tabla 5.11). Tampoco se detectaron en la cepa CR-UIB2 de *M. tuberculosis.* La posición relativa de los profagos en los genomas sólo pudo determinarse con exactitud en los genomas cerrados de las cepas tipo de *M. chelonae* y *M. immunogenum* (Figura 5.8).

En muchos casos los profagos aparentemente intactos según PHAST carecen de elementos de inserción del tipo integrasa o transposa. En ningún caso fue posible la identificación del profago concreto en el contexto de los micobacteriófagos conocidos, ya que los resultados del análisis hecho con PHAST reflejaba similitudes de las diferentes proteínas en diferentes micobacteriófagos.

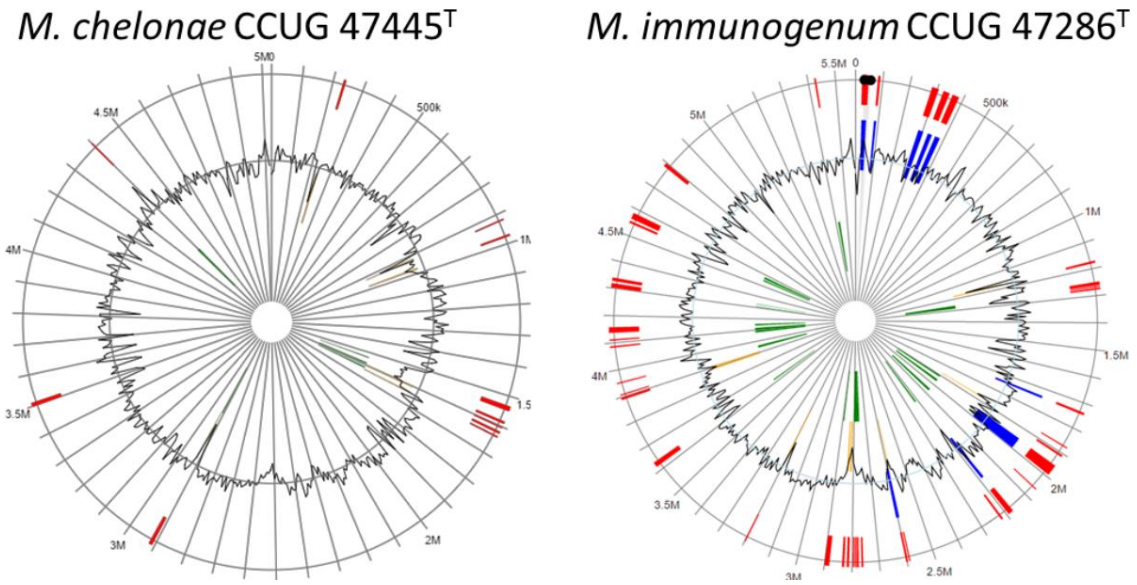


Figura 5.7. Hipotéticas islas genómicas encontradas en los distintos genomas estudiados. La posición relativa de las mismas sólo es concluyente en los genomas cerrados de las cepas tipo CCUG 47445^{T} y CCUG 47286^{T}**.**

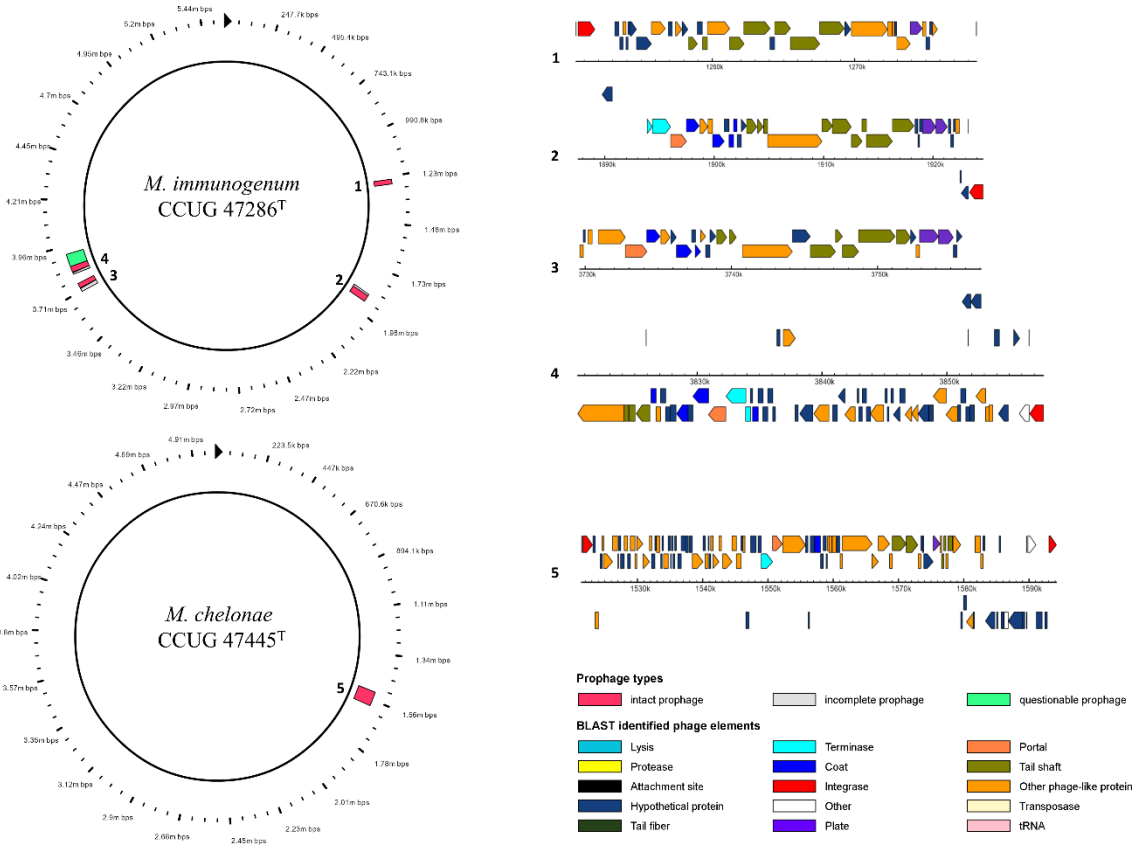


Figura 5.8. Posición relativa de los potenciales profagos encontrados en las cepas tipo de *M. chelonae* y *M. immunogenum*. Se indican también los componentes génicos de cada uno de los profagos indicados como "intactos" por PHAST.

Tabla 5.11. Número de integrasas y transposasas, Islas genómicas y fagos intactos encontrados en las cepas estudiadas.

	CCUG 47445^T	CCUG 47286^T	CCUG 50184^T	MG2	MG8	MHSD2	MHSD3	CR-UIB1
Integrasas	4	3	2	1	1	0	1	1
Transposasas	0	0	0	0	0	0	0	2
Islas genómicas	9	41	20	20	18	14	18	17
Fagos intactos	1	4	0	1	2	1	1	1

5.3.5. Percepción del Quórum

A través de la anotación en la base de datos KEGG, se identificaron en los genomas analizados entre 30 y 33 proteínas potencialmente implicadas en QS, de las cuales 25 correspondieron a elementos comunes presentes en todas las cepas (Tabla 5.12). Sólo en un caso se detectó un elemento exclusivo de una cepa, concretamente una proteína hipotética de la cepa tipo de *M. chelonae*. Aquellos elementos identificados que, según la bibliografía, pertenecían a sistemas más complejos fueron analizados conjuntamente con las proteínas adyacentes, tanto corriente arriba como corriente abajo, con la base de datos STRING para identificar elementos potencialmente relacionados que completasen el sistema. En todos los casos, la base de datos STRING reflejó en los resultados que el microorganismo con las proteínas más parecidas era *M. abscessus*. Los gráficos proporcionados por STRING constan de nodos conectados entre ellos por líneas de colores. Los nodos pequeños corresponden a proteínas de las que no se dispone de estructura 3D conocida, mientras que los nodos grandes corresponden a proteínas que sí disponen de esta información. Las conexiones entre nodos presentan diferentes colores en función de la información que proporcionan: verde (relación de vecindad entre nodos), rojo (fusión génica), azul (co-ocurrencia génica), amarillo (evidencias bibliográficas), negro (evidencias de coexpresión), fucsia (relación determinada experimentalmente entre nodos), entre otras. STRING se nutre de toda esta información a partir de otras bases de datos, las cuales utiliza para determinar las posibles interacciones entre las proteínas analizadas con el fin de proporcionar una predicción de relaciones entre ellas, basadas en información contenida en dichas bases de datos y contrastada con proteínas homólogas presentes en otros microorganismos.

Tabla 5.12. Conjunto de elementos agrupados en la categoría funcional "Percepción del quorum" por la base de datos KEGG. Se indica la presencia (+) o ausencia (-) de cada elemento en cada caso.

	CCUG 47445^T	CCUG 47286^T	CCUG 50184^T	MG2	MG8	MHSD2	MHSD3	CR-UIB1
Antranilato sintasa	+	+	+	-	+	+	+	+
Antranilato sintasa, piocianina específica	+	+	+	+	+	+	+	+
Proteina de union a soluto extracelular	-	-	+	+	+	-	+	+
permeasa DppB	+	+	+	+	+	+	+	+
Glutamato decarboxilasa	+	+	+	+	+	+	+	+
GsiACD	+	+	+	+	+	+	+	+
LivHF	+	+	+	+	+	+	+	+
Proteína hipotética	+	-	-	-	-	-	-	-
KdpE	+	+	+	+	+	+	+	+
Proteína de unión a L, I, V, T	+	+	+	+	+	+	+	+
Permeasa de L, I, V	+	+	+	+	+	+	+	+
LptB	+	+	+	+	+	+	+	+
Proteasa Lon	+	+	+	+	+	+	+	+
FadD15	+	+	+	+	+	+	+	+
YidC	+	+	+	+	+	+	+	+
fosfolipasa C	-	+	+	-	-	+	-	-
OppAC	+	+	+	+	+	+	+	+
Aldolasa	+	+	-	-	-	+	-	-
Aldolasa AroG	-	-	+	+	+	-	+	+
SecAEYG	+	+	+	+	+	+	+	+
YajC	+	+	+	+	+	+	+	+
Transportador de glutamato/GABA	-	+	+	-	-	+	-	+
isocorismatasa	+	+	+	+	+	+	+	+
RibD	-	+	+	+	+	+	+	+

Continuación de la Tabla 5.12. Conjunto de elementos agrupados en la categoría funcional "Percepción del quorum" por la base de datos KEGG. Se indica la presencia (+) o ausencia (-) de cada elemento en cada caso.

	CCUG 47445^T	CCUG 47286^T	CCUG 50184^T	MG2	MG8	MHSD2	MHSD3	CR-UIB1
Peptidasa señal I	+	+	+	+	+	+	-	+
Proteina de PRS	+	+	+	+	+	+	+	+
FtsY	+	+	+	+	+	+	+	+
DdpA	+	+	-	-	-	+	-	-

Todas las proteínas detectadas, así como la indicación de los genomas en los que están presentes, se recogen en la Tabla 5.12. En líneas generales se encontraron proteínas que justifican su participación en el QS sintetizando un compuesto, proteínas asociadas a la membrana, proteínas relacionadas con sistemas de dos componentes, entre otras.

Entre las proteínas cuya función en el QS se basa en la síntesis de algún tipo de compuesto, se encontraron el componente 1 de la antranilato sintasa (piocianina específica) y la enzima isocorismatasa. Estas proteínas corresponden a las enzimas PhzE y PhzD que, junto con PhzC, ya fueron explicadas en el apartado de factores de virulencia. Atendiendo a los resultados obtenidos por STRING, aparecieron codificados y en este orden una carboxilesterasa tipo B, los tres elementos implicados en la síntesis de fenazinas (PhzD, PhzE y PhzC), una deshidrogenasa y un gen codificante para una amidohidrolasa. (Tabla 5.13). Según los resultados obtenidos a través de STRING, y para *M. abscessus*, existieron evidencias de coexpresión (conexiones negras entre nodos) y vecindad (conexiones verdes) entre PhzD, PhzE, PhzC, la deshidrogenasa y la amidohidrolasa (Figura 5.9). Este pequeño bloque de genes venía precedido por una proteína hipotética que, al ser analizada por BLAST en UniProt y en Pfam, correspondería a una transposasa, la cual también estuvo representada en *M. abscessus.* Además, los resultados de BLAST sólo mostraron proteínas altamente similares en las especies *M. abscessus*, *M. abscessus* subsp. *bolletii*, *M. chelonae*, M. *saopaulense* y *M. immunogenum.*

Otra proteína implicada de alguna manera en la biosíntesis de compuestos relacionados con el QS, y detectada por la base de datos KEGG, es una proteína relacionada con la biosíntesis de riboflavina (RibD) y aparentemente implicada en la síntesis de la toxoflavina.

El segundo gran grupo de factores implicados en la QS correspondió a toda una serie de proteínas de membrana. Entre ellas destacaron dos grandes familias de permeasas: las permeasas del tipo Opp (del inglés *OligoPeptide Permease*) y las de tipo Dpp (del inglés *DiPeptide Permease*). Las permeasas Opp están relacionadas con la captación de péptidos señal del ambiente. Dentro de este grupo, se determinó un conjunto de genes con sintenia conservada con respecto a *M. abscessus*, y que estaría constituido por los genes correspondientes a dos permeasas: oppA y oppC. Entre estos dos genes se halló el gen de una tercera permeasa; y tras el bloque formado por estos tres genes se encontró un transportador de ATP, todos ellos implicados aparentemente en la importación de glutationina según la anotación realizada con Prokka (GsiC y GsiA respectivamente, Tabla 5.14). Estos cuatro elementos corresponderían a cuatro subunidades de un mismo sistema de transporte para el cual STRING refleja evidencias de coexpresión entre ellos (Figura 5.10). Nuevamente el BLAST del bloque de proteínas sólo reflejó homólogos claros en el género *Mycobacterium*, concretamente en las especies *M. abscessus*, *M. abscessus* subsp. *bolletii*, *M. chelonae*, *M. immunogenum* y *M. saopaulense*, aparte de toda una serie de genomas del género bajo el nombre de *Mycobacterium* sp.

Por su parte, dentro del grupo de permeasas tipo Dpp se encontraron cuatro proteínas codificadas de forma consecutiva (DdpA, DppB, GsiD y la subunidad de unión al ATP) (Tabla 5.15), entre las que STRING determinó evidencias de coexpresión según el modelo de *M. abscessus* (Figura 5.11), por lo que podría tratarse nuevamente de cuatro subunidades de un trasportador de oligopéptidos (de forma similar al caso descrito con las permeasas tipo Opp) potencialmente relacionado con el QS.

Siguiendo con las proteínas asociadas a membrana, KEGG identificó cinco proteínas pertenecientes a un sistema de transporte de aminoácidos de cadena ramificada, cuyos genes además conservaban la misma disposición en *M. abscessus* (Tabla 5.16) y entre las que se determinó evidencias de coexpresión según STRING (Figura 5.12). Al ampliar el número de proteínas analizadas, tanto corriente arriba como corriente abajo de la anotación, se halló una sexta que parece conservar también la sintenia con respecto a las cinco anteriores y que correspondería a una proteína reguladora de dos componentes (Tabla 5.16). De acuerdo con la anotación con Prokka, la comparación con *M. abscessus* a través de STRING y las búsquedas con BLAST en UniProt, estos genes constituirían el

bloque conocido como LivFGHM, el cual se asocia a una proteína periplasmática que puede ser LivK o bien LivJ en bacterias Gram negativas como *E. coli*. Algunas dudas quedarían en este caso para determinar cuál de estas dos proteínas sería la asociada a este bloque génico, según KEGG sería LivJ. Los resultados de las búsquedas realizadas con BLAST en UniProt de esta proteína evidenciaron que podrían no ser tan exclusivos, apareciendo también en otras micobacterias como *M. phlei, M. iranicum* y *M. llatzerense*. Los resultados de anotación, la comparación con *M. abscessus* y los resultados de BLAST en UniProt confirmarían la posibilidad de que en este caso es un homólogo de LivJ la proteína que completa el sistema. Cabe destacar que, como se puede apreciar en la tabla de factores de virulencia descritos en apartados anteriores (Tabla 5.5), el análisis realizado en KEGG también identificó el gen Mchelo_02579 como el gen *ctr*D, aunque posteriormente se puntualizó en ese mismo apartado que pertenecía a un sistema completo de transporte de aminoácidos de cadena ramificada, hecho que se ha confirmado aquí ya que en ambos casos se trata del mismo conjunto de genes.

Finalizando con las proteínas de membrana, en los genomas de MCR estudiados se encontró la proteína YidC, implicada al parecer en la inserción de proteínas en la membrana celular, tanto dependientes como independientes de los transportadores Sec (Tabla 5.17 y Figura 5.13).

En cuanto a los sistemas de dos componentes, en la anotación inicial por KEGG se detectó el regulador transcripcional KdpE. El estudio de las proteínas relacionadas con KdpE permitió identificar dos sistemas de dos componentes consecutivos en la anotación (KdpD/E, KdpB/C). Al hacer BLAST en GenBank de estos componentes sólo se obtuvieron resultados de genomas correspondientes al grupo *abscessus-chelonae-immunogenum*. Analizando los genes corriente abajo del SDC KdpD/E se consiguió encontrar el operón completo kdpFABC (para la Kdp ATPasa) entre las que STRING determinó la existencia de relaciones de coexpresión entre ellas (Figura 5.14). Por su parte, el gen corriente abajo de la subunidad KdpA no dio ningún resultado en STRING (Tabla 5.18); no obstante, en la anotación obtenida con Prokka se denominó como KdpF.

Para concluir este apartado, se hallaron otras proteínas relacionadas con QS que merecen como mínimo ser enumeradas. Entre ellas, la proteína de la partícula de reconocimiento

de señal o Ffh. Ésta, junto con un ARN ribosomal 4.5S, proteína constituye la denominada partícula de reconocimiento de señal o SRP (del inglés *Signal Recognition Particle*); clave en el transporte de proteínas que deben ser trasladadas a la membrana plasmática. Relacionada con dicha proteína se detectó también la proteína FtsY, un receptor de membrana que reconoce el complejo formado por la SRP y el ribosoma, cuyo péptido naciente estaría destinado a ser transportado a través de la membrana. La proteína anotada como peptidasa señal I presentó una homóloga en *M. abscessus* denominada peptidasa señal LepB (MAB_3223c), la cual consistiría en una endopeptidasa de membrana perteneciente a las peptidasas señal de tipo I (SPase I). Estos datos se recogen en la Tabla 5.19, junto con la información obtenida de las proteínas codificadas entre los genes de estos tres elementos. El análisis en STRING mostró la existencia de relaciones tanto de proximidad como de co-expresión entre estas tres proteínas (Figura 5.15). De hecho, la propia SRP es esencial para su propia inserción en la membrana en un proceso dependiente de translocadores Sec (UniProt).

Tabla 5.13. Número identificativo de las proteínas en el genoma de *M. chelonae* CCUG 47445^T (anotación con Prokka v1.10 y su homólogo en el genoma de *M. abscessus* para el conjunto de proteínas implicadas en la síntesis de fenazinas). Se indica el nombre de la anotación según STRING, así como los valores de identidad y bit-score en cada caso.

Proteína	STRING	Anotación	Identidad	Bit-score
Mchelo_0283	MAB_0292c	ISxo8 transposasa putativa	92 %	822
Mchelo_0284	MAB_0294	carboxilesterasa tipo B putativa	88 %	913
Mchelo_0285	MAB_0295	putativa PhzC	93 %	743
Mchelo_0286	MAB_0296	2,3-dihidro-2,3dihidroxibenzoato deshidrogenasa	94 %	427
Mchelo_0287	MAB_0297	Isochorismatasa/PhzD	95 %	410
Mchelo_0288	MAB_0298	putativa PhzE	92 %	1146
Mchelo_0289	MAB_0299	putativa amidohidrolasa	88 %	577

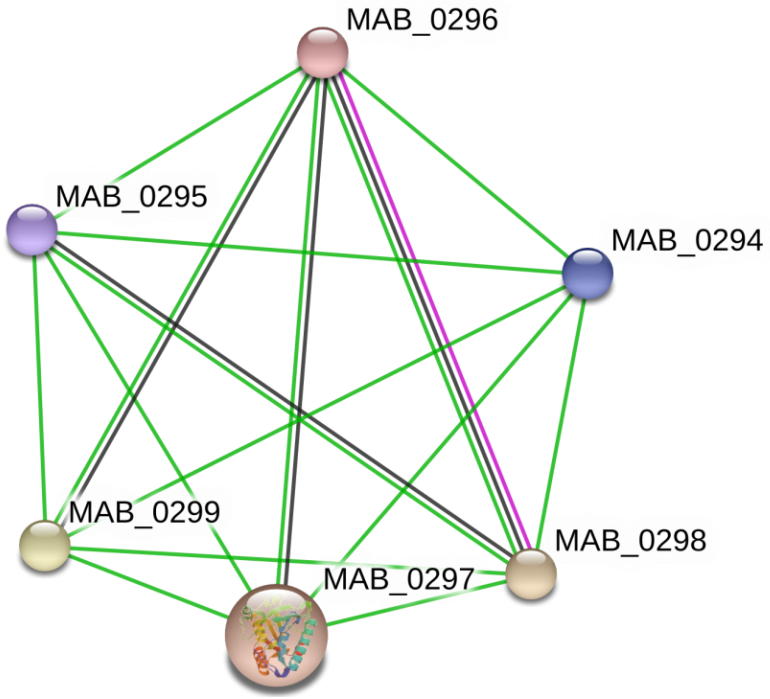


Figura 5.9. Red de relaciones predichas entre las proteínas homólogas en el genoma de *M. abscessus*. Se encontraron relaciones de proximidad (verde), coexpresión (negro) y evidencias experimentales (rosa).

Tabla 5.14. Número identificativo de las proteínas en el genoma de *M. chelonae* CCUG 47445^T (anotación Prokka v1.10) y su homólogo en el genoma de *M. abscessus* para el conjunto de proteínas relacionadas con permeasas Opp. Se indica el nombre de la anotación según STRING, así como los valores de identidad y bit-score en cada caso.

Proteína	STRING	Anotación	Identidad	Bit-score
Mchelo_00420	acsA	acetil-CoA proteasa	94 %	1264
Mchelo_00421	MAB_0426	Proteína de unión a oligopeptido, OppA	94 %	1065
Mchelo_00422	MAB_0427	Permeasa del sistema de transporte de glutationina, GsiC	95 %	594
Mchelo_00423	MAB_0428	Proteína del sistema de transporte de oligopeptidos, OppC	96 %	585
Mchelo_00424	MAB_0429	Proteina de union al ATP (importe de glutationina), GsiA	92 %	1008
Mchelo_00425	MAB_0430	proteína hipotética	94 %	434

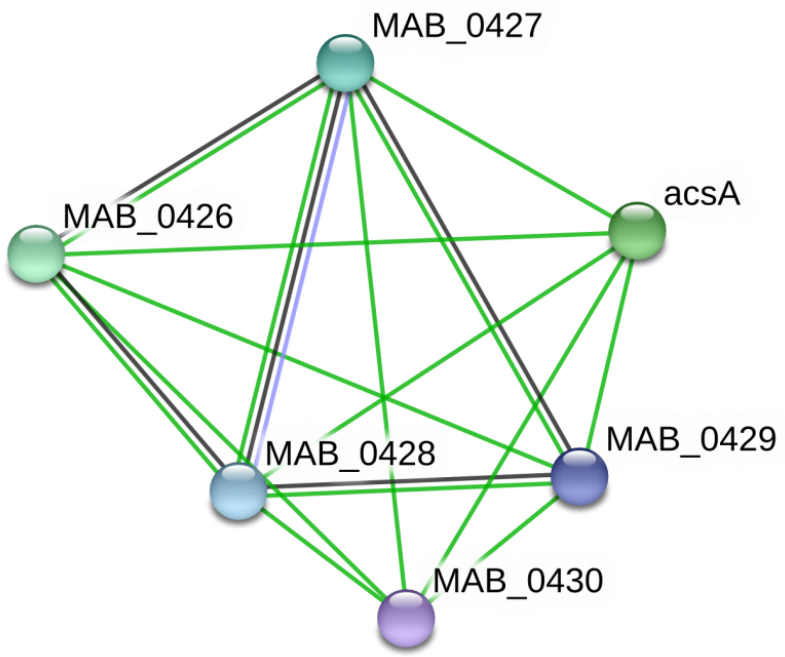


Figura 5.10. Red de relaciones predichas entre las proteínas identificadas como permeasas Opp homólogas en el genoma de *M. abscessus*. Se encontraron relaciones de proximidad (verde), co-expresión (negro) y existencia de datos procedentes de bases de datos curadas (azul).

Tabla 5.15. Número identificativo de las proteínas en el genoma de *M. chelonae* CCUG 47445^T (anotación Prokka v1.10) y su homólogo en el genoma de *M. abscessus* para el conjunto de proteinas relacionadas con permeasas Dpp. Se indica el nombre de la anotación según STRING, así como los valores de identidad y bit-score en cada caso.

Proteína	STRING	Anotación	Identidad	Bit-score
Mchelo_00682	MAB_0718c	Epimerasa/dehidratasa dependiente de NAD	92 %	602
Mchelo_00683	MAB_0719	Proteina periplásmica de union a D,D-Dipeptidos, DdpA	90 %	1002
Mchelo_00684	MAB_0720	Proteína permeasa del Sistema de transporte de dipeptidos, DppB	95 %	637
Mchelo_00685	MAB_0721	Proteína permeasa del Sistema de transporte de glutationina, GsiD	92 %	518
Mchelo_00686	MAB_0722	putativo transpotador ABC de logopéptidos, proteína de union al ATP	91 %	989

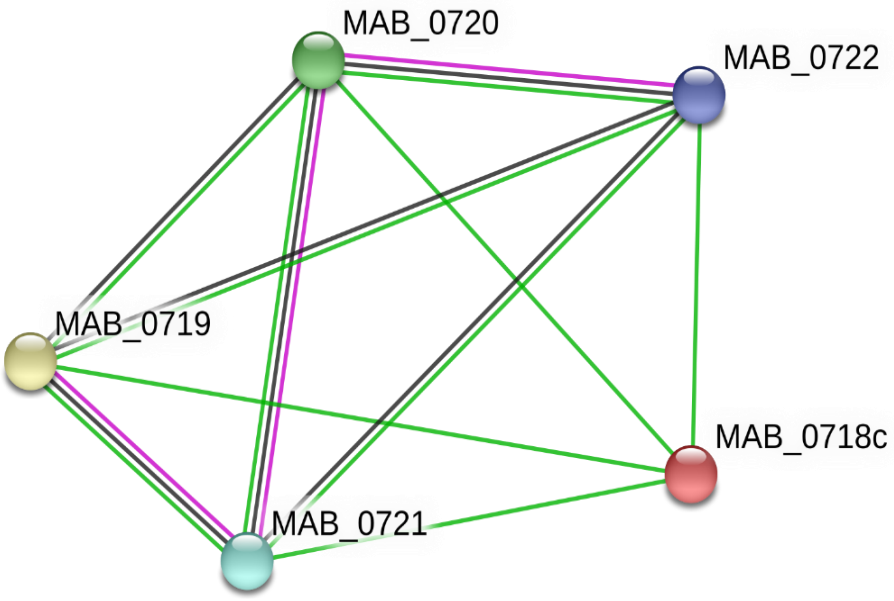


Figura 5.11. Red de relaciones predichas entre las proteínas homólogas en el genoma de *M. abscessus* para el conjunto de proteínas relacionadas con permeasas Dpp. Se encontraron relaciones de proximidad (verde), coexpresión (negro) y existencia de datos procedentes de bases de datos curadas (azul).

Tabla 5.16. Número identificativo de las proteínas en el genoma de *M. chelonae* CCUG 47445^{T} (anotación Prokka v1.10) y su homólogo en el genoma de *M. abscessus* para el conjunto de proteinas que conforman las hipotéticas subunidades del transportador LivFGHM. Se indica el nombre de la anotacíon según STRING, así como los valores de identidad y bit-score en cada caso.

Proteína	STRING	Anotación	Identidad	Bit-score
Mchelo_02579	MAB_2622c	transportador tipo ABC de ACR	98 %	479
Mchelo_02580	MAB_2623c	transportador tipo ABC de ACR	85 %	538
Mchelo_02581	MAB_2624c	transportador de alta afinidad tipo ABC de ACR, LivM	92 %	729
Mchelo_02582	MAB_2625c	transportador tipo ABC de ACR	96 %	571
Mchelo_02583	MAB_2626c	transportador ABC, proteína de union al ligando	90 %	738
Mchelo_02584	MAB_2627c	proteína reguladora de dos componentes	99 %	401

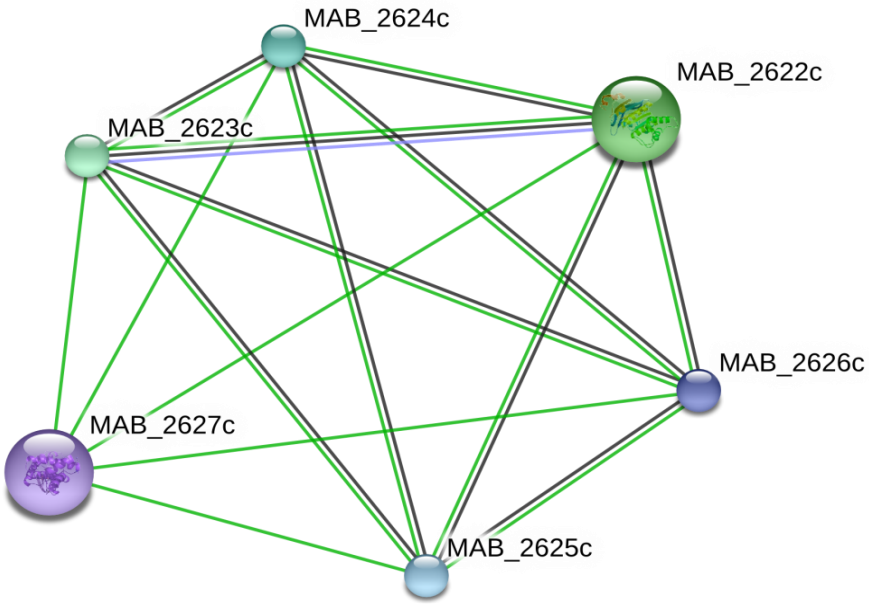


Figura 5.12. Red de relaciones predichas entre las proteínas homólogas en el genoma de *M. abscessus*, pertenecientes al trasnportador de cadenas de aminoácidos ramificadas LivFGHM Se encontraron relaciones de proximidad (verde), coexpresión (negro) y existencia de datos procedentes de bases de datos curadas (azul).

Tabla 5.17. Número identificativo de las proteínas en el genoma de *M. chelonae* CCUG 47445^{T} (anotación Prokka v1.10) y su homólogo en el genoma de *M. abscessus* para las proteínas YajC, SecF, SecD y la proteína de unión a solutos. Se indica el nombre de la anotación según STRING, así como los valores de identidad y bit-score en cada caso.

Proteína	STRING	Anotación	Identidad	Bit-score
Mchelo_02852	apt	Adenina fosforibosiltransferasa	85 %	291
Mchelo_02853	MAB_2878c	Proteína bacteriana de union a solute extracelular	95 %	1066
Mchelo_02854	secF	SecF exportador de proteínas, proteína de membrana	89 %	758
Mchelo_02855	secD	Subunidad SecD, translocador de proteínas	91 %	972
Mchelo_02856	MAB_2881c	subunidad YajC, translocador de pre-proteínas	88 %	194

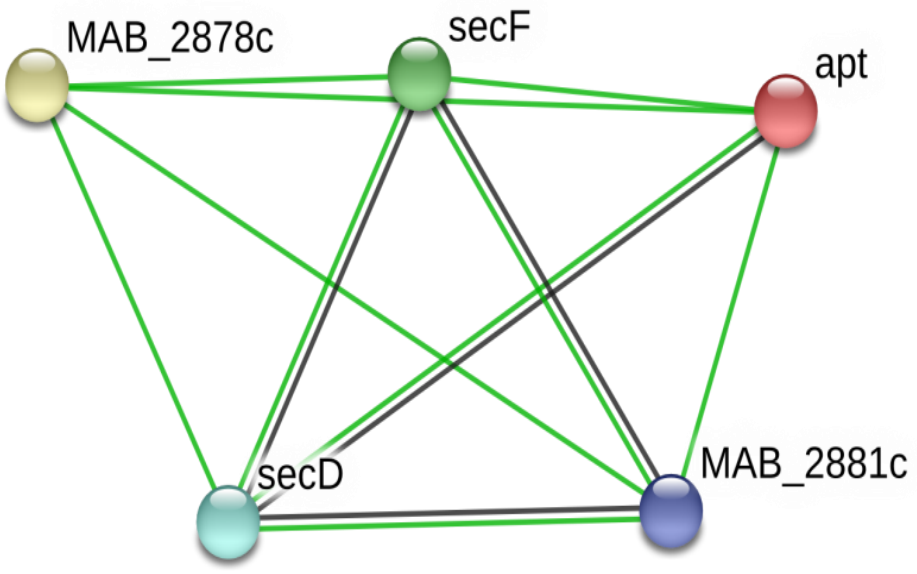


Figura 5.13. Red de relaciones predichas entre las proteínas homólogas en el genoma de *M. abscessus* para las proteínas YajC, SecF, SecD y la proteína de unión a solutos Se encontraron relaciones de proximidad (verde), coexpresión (negro).

Tabla 5.18. Número identificativo de las proteínas en el genoma de *M. chelonae* CCUG 47445^T (anotación Prokka v1.10) y su homólogo en el genoma de *M. abscessus* para el conjunto de proteínas que conforman las hipotéticas subunidades del SDC KdpD/E y la Kdp-ATPasa (KdpFABC). Se indica el nombre de la anotación según STRING, así como los valores de identidad y bit-score en cada caso.

Proteína	STRING	Anotación	Identidad	Bit-score
Mchelo_03247	MAB_3249	Proteína hipotética	95 %	272
Mchelo_03248	MAB_3250c	Regulador transcripcional KdpE	99 %	445
Mchelo_03249	MAB_3251	Proteína sensora KdpD	95 %	1492
Mchelo_03250	MAB_3252c	Cadena C de la ATPasa de transporte de Kx (KDP)	88 %	325
Mchelo_03251	kdpB	Cadena B de la ATPasa de transporte de Kx (KDP)	92 %	1241
Mchelo_03252	kdpA	Cadena A de la ATPasa de transporte de Kx (KDP)	93 %	1063
Mchelo_03253	---	---	---	---

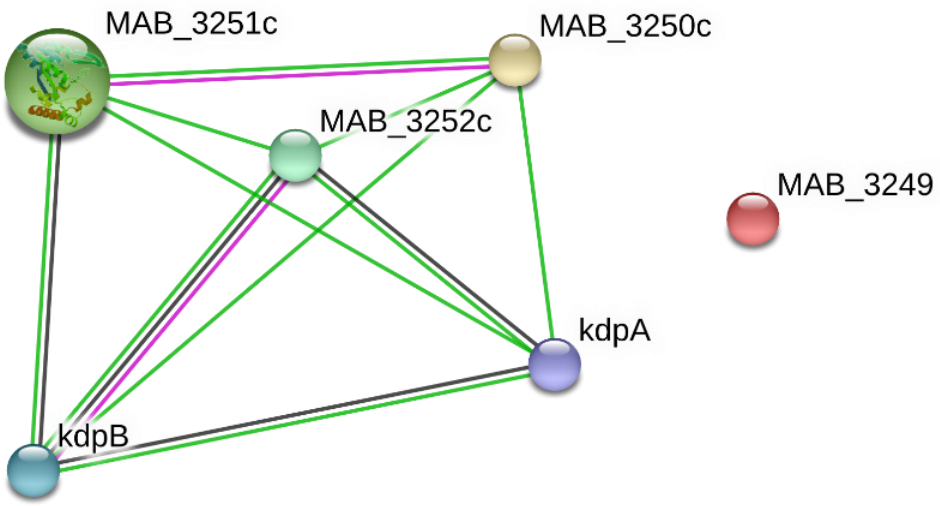


Figura 5.14. Red de relaciones predichas entre las proteínas homólogas en el genoma de *M. abscessus*, pertenecientes a las hipotéticas subunidades del SDC KdpD/E y la Kdp-ATPasa (KdpFABC). Se encontraron relaciones de proximidad (verde), coexpresión (negro) y evidencias experimentales (rosa).

Tabla 5.19. Número identificativo de las proteínas en el genoma de *M. chelonae* CCUG 47445^T (anotación Prokka) y su homólogo en el genoma de *M. abscessus* para el conjunto de la peptidasa señal I, *ffh* y *fts*Y, así como las proteínas codificadas entre ellas. Se indica el nombre de la anotación según STRING, así como los valores de identidad y bit-score en cada caso.

Proteína	STRING	Anotación	Identidad	Bit-score
Mchelo_03218	MAB_3223c	peptidasa señal I, LepB	85 %	440
Mchelo_03219	rplS	proteína ribosomal L19	96 %	223
Mchelo_03220	MAB_3225	lipoproteína putativa LppW	94 %	559
Mchelo_03221	trmD	trnA	93 %	455
Mchelo_03222	rimM	Proteina del procesamiento del 16S rRNA, RimM	85 %	288
Mchelo_03223	MAB_3228c	proteína hipotética	99 %	138
Mchelo_03225	rpsP	Proteina ribosomal S16	98 %	251
Mchelo_03230	MAB_3230c	proteína hipotética	91 %	246
Mchelo_03231	MAB_3231	proteína hipotética	89 %	209
Mchelo_03232	MAB_3232c	putativa oxidoreductasa tipo luciferasa	93 %	639
Mchelo_03233	MAB_3234	D-alanyl-D-alanina carboxipeptidasa DacB	85 %	473
Mchelo_03234	MAB_3235c	proteína hipotética	75 %	119
Mchelo_03235	MAB_3236c	amidohidrolasa	90 %	615
Mchelo_03236	ffh	Proteina de reconocimiento SRP, Ffh	97 %	972
Mchelo_03237	MAB_3238c	PII uridylyl-transferase	95 %	1462
Mchelo_03238	MAB_3239c	nitrogen regulatory protein P-II	100 %	218
Mchelo_03239	MAB_3240c	Transportador de amonio	97 %	840
Mchelo_03240	ftsY	Receptor del complejo SRP y cadena naciente del ribosoma (RNC)	89 %	680
Mchelo_03241	MAB_3242	Isopentenil pirofosfato isomerasa	94 %	606

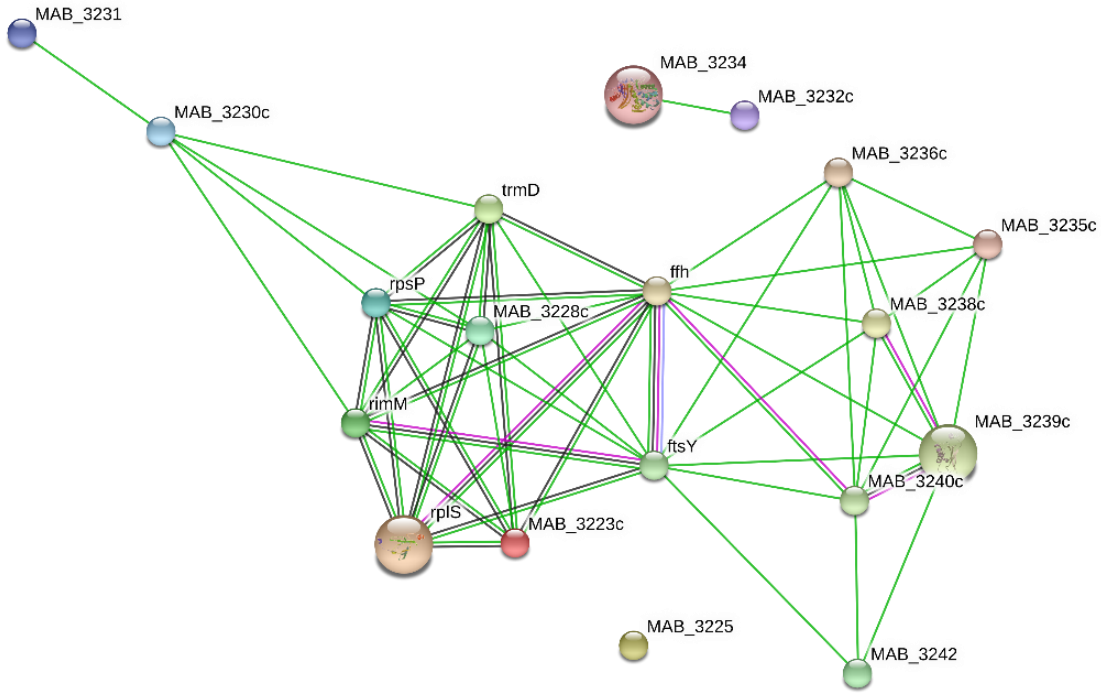


Figura 5.15. Red de relaciones predichas entre las proteínas homólogas en el genoma de *M. abscessus*, pertenecientes a la peptidasa señal I, ffh y ftsY, así como las proteínas codificadas entre ellas Se encontraron relaciones de proximidad (verde), coexpresión (negro), evidencias experimentales (rosa), y existencia de datos procedentes de bases de datos curadas (azul).

5.4. Discusión

5.4.1. Perfil de resistencias

Los estudios genómicos pueden aportar una información considerable en cantidad y de gran utilidad sobre lo que una determinada bacteria es o no capaz de hacer, en base a la simple identificación de los elementos genéticos implicados en la realización de una determinada función. En el ámbito de los perfiles de resistencias, la indagación realizada en los genomas de las cepas de MCR secuenciadas reveló un considerable abanico de potenciales resistencias a antibióticos así como a otros factores externos; hecho que incrementa su relevancia clínica, no solamente por la capacidad de alcanzar intrahospitalariamente al paciente sino también por la dificultad de combatir la infección una vez se haya producido, especialmente en pacientes con problemas de salud de base [11].

En los genomas estudiados se hallaron genes y mutaciones que teóricamente podrían ser responsables de desarrollar resistencia a 18 tipos de antibióticos, además de entre 10 y 14

bombas de expulsión en cada uno los genomas, y que podrían estar implicados en la expulsión de antibióticos específicos como cloranfenicol, estreptogramina, tetraciclina, fluoroquinolonas; o bombas inespecíficas que pueden ocuparse de la expulsión de varios tipos. En estas cepas también se encontraron mutaciones y genes implicados en la resistencia a cuatro importantes agentes antituberculosos como son la isoniacida, etambutol, rifampicina y piracinamida. Es de destacar el hecho de que estos antibióticos suelen combinarse para aprovechar las diferentes capacidades bactericidas o bacteriostáticas en el tratamiento de infecciones provocadas por *M. tuberculosis* con el fin de combatirlas. Además, dichos tratamientos suelen prolongarse varios meses para asegurar la completa eliminación del agente infeccioso [129,130]. Todas las cepas cuyos genomas se han analizado en el transcurso de este proyecto presentaron potenciales resistencias a por lo menos dos de estos fármacos, y en el caso concreto de la cepa tipo *M. abscessus* subsp. *bolletii* CCUG 50184^{T} presentaría las cuatro. En este sentido, la posibilidad de acceder a la información genómica podría ser determinante a la hora de administrar un tratamiento de forma más eficaz con aquellos antibióticos a los que, desde el punto de vista genómico, no son resistentes. Otros casos más particulares fueron la presencia de potenciales resistencias a la bleomicina y daunorubicina, ambos fármacos utilizados principalmente en el tratamiento de determinados tipos de cáncer. En el caso de la bleomicina, este compuesto puede tener un uso antimicrobiano, aunque debido a su alta toxicidad no se permite su aplicación en este tipo de tratamientos (The American Society of Health-System Pharmacists, http://www.drugs.com). Otro caso destacable es la presencia de resistencias a la biciclomicina, un antibiótico utilizado principalmente para el tratamiento de infecciones con bacterias Gram negativas.

La presencia de MBL en todos los aislamientos de MCR estudiados, además de demostrar experimentalmente que son productoras de este tipo de enzimas, incrementa la importancia clínica de estas cepas. El análisis comparativo realizado con respecto a los diferentes tipos de MBL existentes, reveló la agrupación de las MBL detectadas en los aislamientos de MCR en ramas independientes muy estables, con valores de bootstrap superiores a 80, y que reflejan la gran diferencia existente entre algunas MBL de las nuevas cepas de MCR analizadas con respecto al resto de MBLs conocidas y presentes en la base de datos CARD. Este hecho apunta a la posibilidad de que pertenezcan a un

nuevo tipo. La presencia de MBL también se ha descrito en otras dos especies de MCR, concretamente en *M. orubense* y *M. arupense* [131,132].

En cuanto a la resistencia a otros agentes biocidas, en general todas las cepas estudiadas presentarían cierto potencial genético para hacer frente a la presencia de metales pesados, condiciones ácidas y, muy importante, a elementos desinfectantes. La gran importancia de la presencia de estos últimos elementos se debe a que las hace capaces de sobrevivir a la acción de agentes utilizados para la desinfección de aguas o superficies, lo que les proporciona una herramienta más para superar este tipo de medidas preventivas que se adoptan para evitar la propagación de microorganismos. La combinación de todos estos factores puede ser clave en su capacidad demostrada de diseminación a través de sistemas de distribución de aguas naturales hacia instalaciones como hospitales, ya que les permitiría superar las medidas preventivas y de seguridad relativas a higiene y desinfección adoptadas en este tipo de dependencias, para evitar la entrada de agentes microbianos que pudieran agravar la situación de los pacientes o interferir en su recuperación. El hecho de presentar esta potencial capacidad, junto con su presencia en el agua corriente, puede facilitarles alcanzar y contaminar todo tipo de material hospitalario, convirtiéndose en elementos clave para proporcionar el acceso a los pacientes. En líneas generales, el perfil de resistencia genómico de las micobacterias estudiadas les provee de un gran inventario con posibilidades y recursos para el acceso al paciente y de resistencia a los posibles tratamientos.

5.4.2. Elementos asociados a la virulencia de las cepas

Como factores de virulencia se consideran todas aquellas herramientas o elementos de los que un microorganismo patógeno dispone para penetrar las barreras de defensa de un hospedador, sobrevivir e incluso desarrollarse dentro de él. Entendiendo esto, en un microorganismo capaz de desencadenar una infección es importante determinar los recursos de que dispone para favorecer su desarrollo. Esta información puede ser crucial a la hora de desarrollar estrategias que ayuden a restringir su desarrollo a la vez que a eliminar el agente causal de la infección.

El sistema de transporte tipo VII Esx-3 es muy común en las especies que conforman el género *Mycobacterium*, al igual que en otras especies de bacterias Gram positivas [133].

Utilizando como modelo el sistema definido en *M. smegmatis* [134] se observó que la organización génica está prácticamente conservada. En el caso de algunas de las cepas ambientales, al disponer de un genoma a nivel de draft, hay algunos genes que no fueron ensamblados, encontrándose fragmentos de este sistema de secreción al final o inicio de *contigs* diferentes, información que por otra parte podría utilizarse para el diseño de cebadores de PCR con el fin de intentar cerrar esas zonas del genoma. En cualquier caso, la organización parece conservarse en todos ellos. Este sistema está especialmente implicado en la captación de hierro a través de micobactinas [135], cuya síntesis esta codificada por lo genes *mbt*. Estos genes también fueron localizados en genomas de *M. smegmatis*, ubicados relativamente cerca de la región donde se encuentra el sistema de secreción. Este conjunto de genes contribuye a favorecer el desarrollo del microorganismo en condiciones de bajas concentraciones de hierro, tal sería el caso de algunos ambientes extracelulares dentro del hospedador [136]; lo cual les otorgaría una clara ventaja durante la infección. En el caso de los genes *mbt*, la sintenia observada no fue tan conservada como en el ejemplo anterior del sistema de transporte. Estos componentes, unidos a la presencia de los transportadores regulados por hierro irtA e irtB así como del regulador ideR, parecen indicar que es una funcionalidad real en estas cepas, aunque esta necesitaría evidentemente de su confirmación experimental, así como determinar la importancia del papel que desempeñaría está en la capacidad infecciosa del microorganismo.

El conjunto de genes que conforman el complejo antígeno 85, *fbp*ABC codifican para proteínas de secreción de gran importancia para la célula ya que, aparte de su capacidad para unirse a fibronectina (hecho que les proporciona el nombre), su actividad micoliltransferasa es clave para la biogénesis de la pared celular [137]; por lo que contribuyen a la integridad estructural de esta barrera responsable en gran medida de la protección frente a condiciones de estrés ambiental o incluso de resistencia a determinados antibióticos, por ejemplo, en *M. tuberculosis* [138]. Por lo tanto, su papel como factor de virulencia resultaría evidente, al menos en esta especie, ya que contribuye a su supervivencia, permitiendo la progresión de la infección. Los ácidos micólicos son un componente clave en la pared celular de las micobacterias [139], por lo que no es extraño encontrar enzimas relacionados con su metabolismo en todas ellas. Es de esperar que el efecto beneficioso que ejerce sobre la capacidad patogènica de *M. tuberculosis* se

pueda extrapolar a un patógeno oportunista como *M. abscessus* o *M. chelonae*; ya que en ambos casos contribuyen al mantenimiento de una estructura que, por sus características en este género, puede complicar y prolongar en gran medida el tratamiento de un paciente.

El bloque génico anotado como *ure*ABCFGD estaría formado por las subunidades estructurales *ure*ABC, que constituirían la enzima ureasa propiamente dicha, y los genes accesorios *ure*FGD, cuya función estaría implicada en la activación de la enzima [140]. La enzima ureasa ha sido descrita como un factor de virulencia clave, por ejemplo, en la colonización de la mucosa gástrica por parte del patógeno *Helicobacter pylori* [141], mecanismo sin el cual este microorganismo es completamente incapaz de desarrollar la infección [142,143]. La ureasa proporciona en gran medida capacidad de supervivencia en ambientes ácidos. Además, fruto de su actividad se libera amonio al medio, el cual resulta tóxico para las células epiteliales, llegando a ocasionar daños en la mucosa gástrica en el caso de *H. pilory* [144]. Como ejemplos adicionales podrían destacarse los daños en la capa de glusosaminoglicanos del tracto urinario debido a esta misma acción en especies ureasa positivas como algunas incluidas en géneros como *Proteus* o *Klebsiella* [141,145,146]. Teniendo en cuenta que la urea puede encontrarse en muchas partes del cuerpo, esta puede ser objetivo de muchos patógenos, o patógenos oportunistas, para usarla en su beneficio desde diferentes flancos. Por ejemplo, esto podría relacionarse con los importantes daños en la piel ocasionados en infecciones cutáneas por parte de *M. chelonae* o *M. immunogenum*, o incluso favorecer la supervivencia de las micobacterias en ambientes ácidos durante el proceso de infección. En cualquier caso, los efectos provocados por la acción de la ureasa son considerados necesarios para la progresión de determinadas infecciones, como el caso de *H. pylori* mencionado previamente, y en consecuencia se considera un importante factor de virulencia. Otro elemento implicado en la supervivencia en condiciones ácidas es LipF. El promotor inducible que controla su síntesis sólo responde a condiciones de bajos niveles de pH [124] y la mutación del mismo da lugar a una disminución significativa del desarrollo de *M. tuberculosis* en el tejido pulmonar [147]. Este hecho justificaría la importancia que este elemento tiene para la patogenicidad manifestada por esta misma especie.

Las proteínas codificadas por los genes *phd*E1, *phd*D1 y *phd*C1 estarían implicadas en la síntesis de fenazinas, metabolitos secundarios con importante función antimicrobiana. La

presencia de estos genes en micobacterias ya se ha observado en *M. abscessus* [148], pero su función y producción se describen fundamentalmente en especies como *Pseudomonas aeruginosa* (piocianina) [149]. En cualquier caso, la presencia de estos genes en MCR también apunta a que estas podrían ser potenciales productores, lo cual les beneficiaría en la competencia en el ambiente, pero también les otorgaría un mecanismo más para contribuir a su potencial virulencia. En algún caso se ha apuntado a que estos compuestos también podrían contribuir a la persistencia celular [150].

Por su parte, el factor de virulencia HbhA es una proteína de membrana en *M. tuberculosis* que adquirió gran importancia al ser descrita como un elemento necesario para la diseminación extra pulmonar de la infección por este microorganismo [151], siendo clave en su capacidad de interactuar con células epiteliales y para la supervivencia en el hospedador [22]. Su presencia tiene lugar predominantemente en especies virulentas, pero no está presente en especies como *M. smegmatis* [102], la cual raramente causa infección [152]. Esta última especie ha sido utilizada en numerosos intentos para la confección de una vacuna contra *M. tuberculosis*, creando un recombinante que expresara precisamente el gen *hbh*A [153].

Por su parte, la proteasa Clp, junto con las Clp-ATPasas asociadas, forman un complejo degradador de proteínas que en *M. tuberculosis* es esencial para el crecimiento y está estrechamente relacionado con la virulencia de este patógeno [154]; siendo al parecer muy importante su actividad en condiciones de estrés. Nuevamente se trata de un elemento que ha despertado gran interés en los últimos años como diana para el desarrollo de antibióticos o tratamientos alternativos que, mediante la inhibición de su función, contribuyan a eliminar de una forma más eficiente la infección [155]. La identificación de estos elementos en las cepas y especies estudiadas no sólo incorpora un nuevo recurso para favorecer su potencial virulencia en determinadas situaciones, sino que también es una diana alternativa en el mismo sentido que el apuntado para *M. tuberculosis.*

Es un hecho bien conocido que la catalasa-peroxidada KatA favorece la supervivencia intracelular de un importante patógeno respiratorio como es *Legionella pneumophila*, gracias a su acción protectora contra el peróxido de hidrógeno que se genera en el interior de los macrófagos [125]. En el caso de micobacterias como *M. avium* o *M. tuberculosis*,

también se pueden encontrar este tipo de enzimas, como *kat*G, que parecen ejercer el mismo efecto protector contra el peróxido de hidrógeno exógeno generado durante la internalización en los macrófagos [156,157], contribuyendo a favorecer la supervivencia intracelular del microorganismo.

La proteína CtrD forma parte del sistema de transporte de polisacáridos hacia el exterior celular para la formación de la cápsula en microorganismos como *Neisseria meningitidis*, un importante agente causal de las meningitis bacterianas [158]. La cápsula es un importante factor de virulencia para todos aquellos patógenos capaces de producirla, ya que sirve como mecanismo de evasión de diferentes sistemas de defensa pertenecientes a la inmunidad humoral del hospedador [159,160], contribuyendo decisivamente a la infección. La presencia de estructuras similares en *M. tuberculosis* se conoce desde hace tiempo [161]. Durante el proceso de infección de este patógeno se observó la presencia de una envoltura capsular en los bacilos [162,163]. Dicha estructura, formada fundamentalmente por polisacáridos y proteínas [164], actuaría de forma similar a una cápsula, es decir, como barrera defensiva para permitir sortear los sistemas de defensa interpuestos por el hospedador. Sin embargo, a pesar de encontrarse en algunos de los genomas estudiados una proteína del mismo tipo que CtrD, la base de datos KEGG no la catalogó como transportador de polisacáridos, sino como un transportador de proteínas, juntamente con las proteínas adyacentes, tal y como se reflejó en el apartado de resultados. En concreto el análisis en KEGG la asoció con posibles funciones de QS. En este punto se podría apuntar a diversas posibilidades. Una de ellas se refiere a que simplemente se trata de una proteína similar pero cuya función esté relacionada con la comunicación con el ambiente. Otra posibilidad es que se trate de un transportador de proteínas y que realmente forme parte de la supuesta cápsula en estas especies. Finalmente, no hay que descartar el hecho de que sean procesos de QS los que induzcan a la célula a activar estos transportadores con el fin de protegerse desarrollando la estructura capsular. En cualquier caso, la posibilidad de generar una cápsula supone un salto cualitativo importante en la capacidad patogénica, por ejemplo de *M. chelonae*, ya que sólo en este caso se ha encontrado codificada esta capacidad; tanto en la cepa tipo como en cepas próximas a ella.

La proteína RelA, junto con SpoT, inducen la respuesta ante condiciones de falta de nutrientes, en un proceso que forma parte de lo que se conoce como "respuesta estricta" (*the stringent response*). Estas dos proteínas son las encargadas de producir nucleótidos hiperfosforilados de guanosina (ppGpp y pppGpp). Estos nucleótidos actúan como alarmonas**,** moléculas de señalización intracelular en bacterias que son producidas frente a factores y situaciones ambientales desapacibles y que contribuyen a regular la expresión génica a nivel de transcripción [165]. En definitiva, servirían de señal para desencadenar distintos procesos intracelulares con la finalidad de adaptarse a la nueva situación ambiental planteada. El equivalente de la función de estas dos proteínas en bacterias Gram positivas lo realiza únicamente la proteína Rel [166–171]. En *M. tuberculosis*, Rel_{mtb} parece jugar un importante papel en este tipo de situaciones de estrés durante la infección, favoreciendo la aparición de células persistentes que pueden sobrevivir a las condiciones adversas durante un largo periodo de tiempo, para volver a reproducir la infección cuando las condiciones son más favorables [172]. Este tipo de situaciones también puede darse en el caso de infecciones por micobacterias no tuberculosas, como las presentadas en esta tesis, clínicamente difíciles de tratar y que pueden prolongarse en el tiempo probablemente gracias a la existencia de este tipo de recursos en su genoma.

La enzima isocitrato liasa (*icl*) participa en el ciclo del glioxilato, escindiendo el isocitrato para dar lugar a succinato y glioxilato. Forma parte de la derivación metabólica en el ciclo de los ácidos tricarboxílicos. Esto permite la utilización de los ácidos grasos como fuente de carbono cuando otras fuentes escasean, favoreciendo la supervivencia del microorganismos en ese tipo de situaciones [173]. En *M. tuberculosis* tanto los niveles de expresión como la actividad de esta enzima se incrementan de forma significativa durante el proceso de infección [174]. Esta importante función llevó a la descripción de esta proteína como un elemento de gran importancia en la persistencia de este patógeno [127].

Algunas proteínas reguladoras son también consideradas factores de virulencia por el efecto modulador que ejercen determinados grupos de genes, los cuales responden a determinadas situaciones como la falta de nutrientes, hipoxia, presencia de agentes externos agresivos u otro tipo de situaciones de estrés; favoreciendo la virulencia del microorganismo. En este sentido, la importancia de los SDC en especies patógenas ya ha sido descrita en numerosas ocasiones y en diferentes microorganimos, incluyendo

Bordetella pertussis [175], *Neisseria meningitidis* [176] o *Salmonella typhimurium.* En este último caso, se habla concretamente de *pho*P. Se trata de la parte sensora del SDC *pho*P - *pho*Q, la cual responde a los niveles de Mg^{2+} del medio. En situaciones de privación de este elemento, provoca la activación del regulador transcripcional asociado para que se desencadene la respuesta adecuada. Estas respuestas pueden consistir en la expresión de genes que favorecen la supervivencia en macrófagos, resistencia a condiciones ácidas o la modulación de su propia capacidad invasiva [122], por lo que resulta evidente su importancia para este tipo de patógenos. En el caso de *M. tuberculosis*, más próximo a los genomas estudiados en profundidad en la presente tesis, *pho*P, perteneciente al SDC *pho*P - *pho*R, presenta evidencias experimentales que apuntan también al desempeño de un papel de cierta relevancia en la regulación de la expresión de genes implicados en la virulencia, además de la alteración de la capacidad del microorganismo de multiplicarse en el ambiente intracelular al ser mutado *pho*P [122]. Por extensión, su función podría tener su equivalente en la patogenicidad oportunista de las cepas objeto de estudio

Por último, pero también de gran importancia es la presencia de operones MCE en todos los genomas estudiados. En lo que respecta a *M. tuberculosis,* no sólo es capaz de penetrar en macrófagos, sino también en células no fagocíticas como las epiteliales. Esto llevó a postular en su momento que el proceso estaba directamente modulado por la propia bacteria debido a las propias características de las células invadidas. Estudios más detallados del genoma de la cepa tipo *M. tuberculosis* H37Rv permitieron la identificación de un gen, que al ser clonado en *Escherichia coli* le otorgaba la capacidad de penetrar y sobrevivir en el interior de células epiteliales [177]. Por este motivo se le llamó *mammalian cell entry protein* (proteína de entrada a células de mamíferos, *mce*). Además, se determinó que pertenecía a un operón cuya organización constaba de unos seis genes precedidos de dos transportadores de membrana similares a los transportadores ABC [178,179], identificándose hasta cuatro operones de este tipo en dicha especie. El análisis detallado de otros genomas permitió describir su presencia en otras especies, especialmente del orden Actinomicetales; además del hecho que no eran exclusivas de patógenas, al encontrarse en especies ambientales como *M. vanbaalenii* [180] que presenta un gran potencial en biorremediación, o *M. abascessus* [148]. La presencia de operones *mce* en micobacterias puramente ambientales puede estar relacionada con el

hecho de que en el ambiente podrían quizás desempeñar otra función, aunque su potencial implicación en la patogenicidad de microorganismos como *M. tuberculosis* induce a pensar que pueden tener también su importancia en patógenos oportunistas del género *Mycobacterium* en general; tales como las cepas y especies estudiadas en este trabajo. Por lo tanto, deberían tenerse en consideración al mismo nivel que el factor de virulencia *mce*A1 del agente causal de la tuberculosis, en la medida de que no hay estudios suficientes que puedan descartar su implicación en el proceso de infección de estas especies. En este sentido, podría entrar en la lista de nuevas dianas alternativas a considerar para el desarrollo de tratamientos que favorezcan la erradicación de su infección.

Recapitulando, los genomas estudiados de cepas pertenecientes o cercanos a las especies *M. chelonae, M. immunogenum, M. abscessus* y *M. abscessus* subsp. *bolletii* disponen de toda una serie de recursos en forma de potenciales factores de virulencia que les otorgan toda una serie de capacidades que pueden ser de crucial importancia, tanto en el inicio como durante el desarrollo de una infección oportunista. La resistencia a condiciones ácidas, la capacidad de reaccionar ante una situación de estrés oxidativo, adaptarse a una situación de falta de nutrientes utilizando fuentes de carbono alternativas, la disposición de elementos que pueden facilitar su penetración en células de mamíferos o incluso la posibilidad de formar algún tipo de cápsula protectora; son habilidades propias de un patógeno estricto, como es el caso de *M. tuberculosis*. La presencia de elementos equivalentes en los genomas de las MCR estudiados las convierten en potenciales usuarios de esos mismos recursos. Así, la detección de los mismos abre un interesante abanico de dianas alternativas para el futuro desarrollo de fármacos que hagan más eficaz el tratamiento de las infecciones en las que estas micobacterias estén implicadas y, además, pone de manifiesto que muchos de los estudios realizados con el fin de desarrollar tratamientos sobre especies como *M. tuberculosis,* podrían ser extrapolables a especies de MCR, ajustándose a las características de las mismas, pero siguiendo una misma estrategia.

5.4.3. Capacidad reguladora relacionada con la patogenicidad

Para una bacteria con capacidad infectiva no sólo son importantes aquellos elementos génicos de los que pueda disponer para tal fin, sino también de la capacidad reguladora de la expresión de los mismos según lo requiera la situación. Los genomas analizados en este estudio presentan la información genética necesaria para una gran variedad de resistencias, tanto a antibióticos como a otro tipo de agentes biocidas; además de toda una serie de factores de virulencia de los que pueden valerse en el momento de desencadenar una posible infección. Al mismo tiempo, disponen de un amplio repertorio de proteínas reguladoras clasificadas en diferentes familias.

En el conjunto de familias de proteínas reveladas en estos genomas se encuentran representantes con implicaciones en la modulación de la respuesta frente a antibióticos, metales pesados, situaciones de estrés como los provocados por "choques térmicos", proteínas reguladoras de los procesos de reparación del ADN o incluso la ya comentada formación de células persistentes. A todas ellas se pueden añadir familias implicadas en los procesos de QS y de formación de biopelículas, estructuras que pueden contribuir a incrementar todavía más su resistencia, al quedar embebidos en estructuras organizadas y que les aportan protección [181]. Concretamente, y si atendemos a los resultados reflejados en la Tabla 5.6, aparte de la regulación de procesos comunes del metabolismo bacteriano, se encontraron familias en las cuales se pueden encontrar representantes relacionados con elementos de virulencia (AraC, AsnC, Crp, IclR, LuxR, MarR y Mga) o patogenicidad (Crp y TetR). En otras familias como DtxR, FeoC, Fur y Rrf2 se encuentran representantes que responden al hierro, implicados, por ejemplo, en la homeostasis de este elemento (DtxR y Fur). También se hallaron FT implicados en la resistencia a factores externos como metales pesados (ArsR, CsoR), antibióticos (IclR, MarR, PadR y TetR), elementos que regulan la respuesta ante un choque térmico (HrcA, MarR y PadR) u otras situaciones de estrés (AraC, Fur, LuxR, MarR, MerR y PadR). Destacan también las familias IclR, LuxR, LysR y PadR, que, entre otras funciones, pueden estar implicadas en la regulación de los procesos de comunicación intercelular. Además, LysR incluye representantes relacionados con la formación de biopelículas. Otras familias halladas podrían estar implicadas en la regulación de procesos de

reparación del ADN (AsnC y LexA), persistencia bacteriana (AsnC) e incluso en la síntesis de antibióticos (GntR y TetR)).

De estos datos parece derivarse que la capacidad reguladora sobre diferentes aspectos que pueden estar implicados en su patogenicidad es extensa, además de poder dar respuesta a un amplio espectro de situaciones, las cuales pueden darse tanto en el ambiente exterior como en el interior de un hospedador. En el conjunto de familias de proteínas reguladoras la más representada con diferencia es la familia TetR (más de 100 en todos los casos analizados), lo cual es normal teniendo en cuenta que es especialmente importante en aquellos microorganismos adaptados a vivir en entornos cambiantes, como es el caso de las bacterias de origen ambiental en general y de las micobacterias en particular. En este contexto, la capacidad de responder en el momento adecuado representaría una ventaja al microrganismo en ambientes donde los continuos cambios le provocan situaciones de estrés. En el caso de patógenos estrictos, especialmente en aquellos especializados en un tipo de hospedador, aunque las condiciones pueden llegar a ser altamente estresantes, no dejan de ser comparativamente más estables, o los cambios no son de índole tan diversa, por lo que la dependencia de este tipo de reguladores parece no ser tan elevada [182]. El ejemplo es claro si se compara el número de representantes de la familia TetR presentes en genomas de especies de índole ambiental como *M. chelonae* con el número encontrado en *M. tuberculosis*. En el caso de *M. chelonae* se superan los cien, mientras que en *M. tuberculosis* el número oscila en torno a los cincuenta según los resultados obtenidos al haber realizado la misma determinación de proteínas reguladoras sobre el genoma de la cepa tipo *M. tuberculosis* H37Rv y la cepa *M. tuberculosis* CR-UIB2 en el presente trabajo (Tabla suplementaria 2, anexo 2), dato que se corresponde con lo encontrado en la literatura [182].

Los SDC juegan igualmente un importante papel en la capacidad de respuesta del microorganismo a un determinado estímulo. Se trata de sistemas biológicos de señalización en los que existe en la membrana un componente especializado, habitualmente una proteína histidina quinasa (HK*)*, destinado a detectar la señal adecuada. La HK, una vez recibido el estímulo adecuado, se autofosforila para después transferir la señal química a un regulador de respuesta (RR) intracelular, que a su vez se encargara de desencadenar la respuesta modulando la expresión de determinados genes

en cada caso [2]. En los genomas estudiados se detectaron entre 15 y 17 SDC completos, con sus respectivas HK y RR. El RR hallado más abundante fue el de tipo OmpR. Un ejemplo de las funciones que pueden desempeñar este tipo de RR es el del sistema EnvZ/OmpR, presente en microorganismos como *E. coli*, cuya función es la de regular los niveles de expresión de las porinas OmpF y OmpC en respuesta a cambios en la osmolaridad ambiental [183] para hacer frente a esta clara situación de estrés. Pero las funciones de este tipo de SDC con RR tipo OmpR pueden ir más allá y participar en la modulación de la expresión de genes que influyen en la patogenicidad del microorganismo, como son el operón flagelar *flh*DC o el locus *ssr*A/B-*spi*C alojado en una isla de patogenicidad de *Salmonella* [184]; así como la regulación de genes de virulencia de otros patógenos como *Yersinia enterocolitica*, *Salmonella typhi* o *Shigella flexneri* [185]. Estos precedentes sitúan a la familia de RR OmpR como serios candidatos a participar en la modulación de la patogenicidad de las cepas estudiadas.

Dentro de las proteínas reguladoras, los factores sigma desempeñan un papel transcendental ya que, mediante la unión de diferentes factores sigma a la RNA polimerasa van a modificar el grado de afinidad de esta enzima por diferentes promotores. De esta forma, cada factor sigma influye de forma diferente sobre la expresión de determinados grupos de genes, cuyas funciones pueden ser necesarias en un momento determinado [2].

En base a estudios previos sobre especies patógenas tanto de *M. tuberculosis* o *M. bovis*, como de MCR, por ejemplo, *M. smegmatis* [186], se pueden hacer hipótesis sobre las posibles funciones que desempeñan los diferentes FS en los genomas de las cepas estudiadas. Para empezar, SigA es el principal FS, el único para el cual hay evidencias de que es esencial para el crecimiento celular. SigB estaría implicado en la respuesta a una situación de baja aireación, presencia de peróxido de hidrógeno o un choque térmico, ya que ante estos estímulos se ve incrementada su expresión [187,188]. SigC, por su parte, parece estar presente sólo en especies patógenas y estaría implicado en la regulación de la expresión de factores de virulencia, como por ejemplo el gen *fbp*C. En el caso de SigD estaría implicado en la regulación de la expresión de genes participantes en la respuesta frente a la falta de nutrientes o la llamada respuesta rigurosa (*stringent response*), permitiendo la adaptación a este tipo de situaciones. La expresión del propio SigD podría

estar regulada por el tetrafosfato de guanosina producido por *rel*A, un factor de virulencia identificado previamente también en estos genomas. SigE participaría en la regulación de la respuesta en presencia de antibióticos como la isoniazida y vancomicina, a la presencia de SDS o ante un choque térmico. En patógenos como *M. bovis*, SigF es capaz de responder a antibióticos como la rifampicina, etambutol, estreptomicina y cicloserina; así como también modular la expresión de determinados genes ante una situación de falta de nutrientes, estrés oxidativo o ante un descenso de temperatura. También se apunta hacia su posible participación en el metabolismo anaeróbico. SigG es de los FS que más se inducen cuando los bacilos se encuentran dentro de los macrófagos [189,190] y las evidencias apuntan hacia su implicación en la expresión de los sistemas de reparación del ADN (respuesta SOS). SigH parece tener un papel central en la regulación de la respuesta ante una situación de incremento de temperaturas o de estrés oxidativo. Por su parte, SigI y SigJ son dos FS cuyas funciones no están del todo claras pero que apuntan a una regulación de la respuesta frente a la presencia de peroxido de hidrógeno via KatG (SigJ) y a un descenso de temperaturas (SigI). Los factores SigL y SigM son otros ejemplos cuya función no está clara pero que podría estar relacionada bien con la virulencia del microorganismo en el primer caso, o con la adaptación a determinados ambientes en el hospedador para permitir una adaptación a largo plazo en el segundo. En el caso de SigK parece estar bien conservado en el género, pero con funciones diferentes si atendemos a lo visto en *M. tuberculosis y M. bovis*. Por último, SigX ha demostrado su capacidad para modular la formación de biopelículas o su implicación en aspectos relacionados con la virulencia, por ejemplo, en *Pseudomonas aeruginosa* [191,192]. Se han encontrado homólogos de Sigx en todos los genomas menos las cepas CCUG 47445^{T}, CCUG 47286^{T} y MHSD3.

Como se puede comprobar, las MCR analizadas disponen de la información genómica para factores sigma que les permitirían iniciar la expresión tanto de factores de virulencia como de genes implicados en dar respuesta a diferentes situaciones de estrés; desde la presencia de antibióticos o de estrés oxidativo, hasta situaciones de escasez de nutrientes o choques térmicos. En conjunto, las MCR disponen codificado en sus genomas de un amplio abanico de potenciales resistencias y factores de virulencia que merecen ser tenidas en cuenta desde el punto de vista patogénico; de proteínas reguladoras que, atendiendo a los resultados obtenidos, pueden regular la expresión de esos mismos

elementos y hacer frente a un amplio rango de situaciones adversas, además de los factores sigma precisos para, en definitiva, poder iniciar su transcripción acorde a los requerimientos ambientales de cada situación. Desde un punto de vista genómico se puede justificar la capacidad de estos microrganismos de desencadenar infecciones oportunistas en determinadas situaciones, ya que disponen de las herramientas genéticas para hacerlo. Sin embargo, nuevamente es necesario ir más allá y realizar estudios más específicos desde el punto de vista experimental para demostrar el alcance de los resultados genómicos meramente descriptivos obtenidos, además de establecer las posibles conexiones existentes entre los distintos elementos destacados en este apartado. Solo así se puede determinar cómo funciona la maquinaria patogénica de estas cepas, a qué responden realmente y en qué situaciones es necesario cada uno de los elementos del engranaje detectados.

5.4.4. Mobiloma

El conjunto de elementos móviles de un microorganismo puede ser una buena indicación de su plasticidad genómica, lo cual puede ser importante a la hora de tener en cuenta su capacidad para captar nuevas funciones, que a la larga podrían favorecer su patogenicidad o la capacidad de reorganización genómica de la que disponen.

Uno de estos elementos son las islas genómicas, elementos que se transfieren de forma horizontal entre microorganismos y que pueden desempeñar un papel determinante si se trata de islas de patogenicidad (IP), o fracciones del ADN genómico de un microorganismo patógeno que le faculta como virulento [2]. En el caso de las MCR estudiadas, a pesar de que se detectaron potenciales islas genómicas en varios genomas, ninguna pareció corresponder a una IP propiamente dicha (datos no mostrados). Además, tampoco disponen de un extenso repertorio de integrasas y transposasas, lo que induce a pensar que no son muy dadas a movilizar elementos genéticos a través de estos componentes. Tampoco se detectaron plásmidos, pero sí diversos profagos insertados en sus genomas, de los cuales solo se tuvieron en cuenta aquellos a los que el análisis en PHAST consideraba como intactos. Estos profagos pueden ser importantes por las posibles funciones que pueden proporcionar a la bacteria hospedadora.

En resumen, el mobiloma de los genomas estudiados no aporta mucha información relativa a la potencial virulencia o patogenicidad de los mismos.

5.4.5. Implicaciones de los elementos de percepción del Quórum detectados

Los fenómenos de QS permiten la modulación de la expresión génica en función de la densidad poblacional, y esa modulación puede conducir al desarrollo de respuestas; incluidas la expresión de factores de virulencia, resistencias o incluso la formación de biopelículas [2]. Por este motivo se consideró que era un aspecto importante a tener en cuenta en la prospección de los genomas estudiados para entender mejor su comportamiento durante el desarrollo de una infección. Así, se detectaron un gran número de elementos potencialmente implicados en el QS. A partir de algunos de los elementos detectados se pudo encontrar todos los componentes relacionados al mismo y que conformaban un determinado sistema, afianzando la posibilidad de que pudieran llevar a cabo dicha función al disponer del sistema completo.

Siguiendo el mismo esquema que en el apartado de resultados, se empezará a discutir las conexiones entre los elementos encontrados y el QS por las proteínas implicadas en la síntesis de compuestos, comenzando por las proteínas implicadas en la síntesis de fenazinas. En estudios previos ya se describió la presencia en *M. abscessus* de estos elementos "no micobacterianos", importantes para el proceso infeccioso en otras especies como *P. aeruginosa*, [148], por lo que existen precedentes que refuerzan la posibilidad de que realmente puedan estar capacitados para la síntesis de estos compuestos. Además, el hecho de encontrar este bloque de genes cerca de una proteína hipotética con dominios de transposasa, abre la posibilidad a que sea un elemento que haya podido ser recibido por transferencia lateral. El posible beneficio de estos componentes en la patogenicidad ya ha sido descrito en la discusión de los factores de virulencia encontrados.

Como se observa en la Tabla 5.12, también se identificó una proteína potencialmente implicada en la síntesis de antranilato. En *P. aeruginosa*, el antranilato puede ser utilizado para la síntesis de la denominada quinolona señal PQS (del Inglés *Pseudomonas Quinolone Signal*), implicada en la regulación de genes de virulencia [193]. La síntesis de antranilato se engloba en la ruta de síntesis del triptófano, esencial para la virulencia

de *M. tuberculosis* [194]. La síntesis de este compuesto se desarrolla en dos pasos llevados a cabo por las dos subunidades de la antranilato sintasa (TrpG y TrpE). En ausencia de la subunidad TrpG, TrpE es capaz de sintetizar antranilato a partir de corismato y altas concentraciones de amonio [195–197].

Siguiendo con las proteínas de síntesis, la supuesta proteína RibD pertenecería según lo reflejado en resultados a la ruta de síntesis de la toxoflavina, una toxina que puede tener actividad antibiótica. La toxoflavina está producida por microorganismos como *Burkholderia glumae,* en la que representa un importante factor de virulencia. La síntesis en este microorganismo está regulada por QS, dependiente en este caso de N-octanoil homoserina lactona [198].

Otro gran grupo de proteínas implicadas en el QS son las proteínas de membrana, las cuales son clave tanto en la comunicación directa con el ambiente como en los procesos de intercambio de sustancias entre el entorno y el medio intracelular. Uno de los elementos destacados fueron las permeasas Opp, destinadas al transporte de oligopéptidos. En Gram positivos existe la capacidad de modular la composición de la superficie celular en respuesta a la detección de señales en forma de oligopéptidos transportados por estas permeasas, hecho que ya ha sido descrito en patógenos como *M. tuberculosis* [199], por lo que suponen una herramienta efectiva para la comunicación del microorganismo con el ambiente que lo rodea. De hecho, la mutación del gen *opp*A en *M. avium*, altamente expresado en ratones durante una infección pulmonar, de hígado o bazo, conduce a la reducción de su viabilidad frente a macrófagos [200], por lo que se apunta como un elemento implicado en la virulencia de la bacteria basada en la captación de péptidos señal que inducirían cambios en la expresión génica. En este mismo contexto podríamos incluir los transportadores Dpp, destinados en este caso al transporte de dipéptidos.

A parte de estos transportadores de oligopéptidos, también se detectaron proteínas aparentemente pertenecientes a un sistema de transporte de aminoácidos de cadena ramificada. Existen tres aminoácidos de cadena ramificada (ACR) importantes en bacterias (leucina, isoleucina y valina). Estos ACR pueden desempeñar un importante papel ante una falta de nutrientes, actuando como moléculas señal que provocarían un

cambio en la regulación de la expresión génica, por ejemplo, de factores de virulencia que permitirían la supervivencia. Podemos encontrar sistemas de captación de baja afinidad (BrnQ) o de alta afinidad (LivFGHMJ y LivFGHMK). En este caso estaríamos ante un sistema de alta afinidad actuando LivJ aparentemente como proteína de unión al sustrato, la cual por otra parte tiene la capacidad de unirse a los tres tipos de aminoácidos mencionados [201,202]. LivJ (y LivK), son proteínas que actúan en el periplasma, tal y como se ha indicado en apartados anteriores. Esta estructura es característica de la envoltura celular de bacterias Gram negativas y, por ende, no está presente en micobacterias. Por lo tanto, el enigma que restaría por resolver es dónde se produce dicha unión al sustrato o cómo es el mecanismo de acción de dicha proteínas en el género *Mycobacterium*.

Otra proteína de membrana destacada en el apartado de resultados fue la ATPasa SecA. En micobacterias, así como en algunos Gram positivos patógenos, se pueden encontrar dos copias de esta proteína: SecA y un parálogo (SecA1 y SecA2), donde SecA2 estaría implicado en la exportación de proteínas más pequeñas y, en algunos casos, relacionadas con la virulencia. De nuevo volvemos a encontrarnos con el mismo caso que el demostrado para *M. tuberculosis*, es decir, las dos copias no redundantes de dicha ATPasa. En esta especie patógena se ha demostrado cómo SecA1 era esencial para el crecimiento [203].

Para terminar con el inventario de las proteínas de membrana, trabajos realizados sobre *M. tuberculosis* demuestran la importancia de la proteína YidC en este patógeno, la cual no lo olvidemos también fue detectada en los genomas estudiados en este proyecto. Al mutarla en *M. tuberculosis* se observan cambios considerables en los patrones de expresión de genes elementales para la respiración celular, ATP-sintasas y genes implicados en general en la supervivencia en condiciones de hipoxia. Además, YidC ha demostrado ser esencial para el crecimiento in vitro y para la supervivencia en el interior de macrófagos [204].

En el contexto del QS, un grupo diferente de proteínas hallado corresponde a las constituyentes de SDC. En este caso se consiguió describir el sistema KdpD/E, un interesante SDC encargado de la homeostasis del K^+. La función de este SDC podría estar

conectada con la virulencia y el QS. El sistema KdpD/E respondería a los niveles de K^+ del medio de tal forma que niveles elevados del mismo inducirían la fosforilación de KdpE, quien induciría la expresión de la Kdp ATPasa para la captación del mismo. Esto puede ser de gran relevancia durante el proceso de infección como ya se ha comprobado en otras especies como *Staphylococcus aureus* [205,206], donde su actuación entraría en competencia directa con el transporte activo de K^+ a través de la membrana del fagosoma en los neutrófilos. El K^+ es clave para la liberación de péptidos antimicrobianos que iniciarían la cascada de ataque hacia la bacteria, de manera que incrementa su supervivencia al dificultar dicha respuesta [207]. En experimentos sobre las especies del complejo *M. tuberculosis* o *M. avium* se vio que la acción de estos genes incrementa la virulencia en el primer caso y que por algún motivo se transcribe activamente durante la fagocitosis en el segundo [207,208]. Este SDC es capaz, según las evidencias aportadas por estos estudios, de responder, aparte de a la concentración de potasio, a estímulos relacionados con la osmolaridad, turgencia celular o señales procedentes del hospedador que indican algún tipo de respuesta contra la bacteria. Además, sería capaz de completar respuestas relacionadas con la densidad poblacional (QS) integrando respuestas procedentes de los sistemas de QS. Una posible explicación a este hecho sería, por ejemplo, que la detección del potasio ambiental le indicaría a la bacteria si se encuentra en un ambiente extra o intracelular y si, además, existe una densidad poblacional suficiente para desencadenar una respuesta favorable a la infección [207].

Aparte de estos grupos de proteínas, se detectaron otros elementos cuya función puede ser muy relevante desde el punto de vista de la comunicación de la célula con el medio que la rodea. Entre estos elementos destacó la proteína de reconocimiento de señal, clave para la formación de la previamente comentada SRP. Esta proteína se une al complejo péptido naciente-ribosoma cuando aparece una secuencia señal correcta y lo desplaza hasta la membrana. Esta interacción permite la transferencia del complejo ribosomal al translocador, de tal forma que el péptido pueda ser transportado al otro lado de la bicapa lipídica o mediar la inserción de proteínas de membrana a medida que se sintetiza [209]. Por tanto, estos elementos son parte de un sistema clave de la secreción de proteínas que, entre otras funciones, pueden estar relacionadas tanto con la virulencia del microorganismo como con el QS. Este complejo puede asociarse a la peptidasa señal encontrada tal como se ha indicado en resultados. Estas peptidasas señal de tipo I son

claves para la liberación de las proteínas que se están transportando a través de la membrana [210]. Estudios en *M. tuberculosis* han demostrado la presencia de un homólogo a esta proteína, en copia única, y que además parecía esencial para el crecimiento del microorganismo [9], por lo que se postula como una posible diana terapéutica.

Otro elemento recogido en la Tabla 5.12, y que debe destacarse, es GabB. En *Escherichia coli*, GabB pertenece al sistema de resistencia a condiciones ácidas *gad*, cuya acción contribuye a la supervivencia de las bacterias ante situaciones de pH extremadamente ácidas [211]. La descarboxilación del glutamato en el interior celular se realiza a costa del uso de protones del mismo, dando lugar a la formación de ϒ-aminobutirato (GABA), el cual aporta unas características más alcalinas. GABA es transportado al exterior por el transportador GABA, el cual expulsa GABA a la vez que capta más glutamato del medio. En conjunto se produce un incremento del pH tanto dentro como fuera de la célula debido a la eliminación de protones intracelulares y a la sustitución del glutamato extracelular por GABA [212]. En microorganismos como *M. leprae* se ha indicado la posible conexión de este transportador y la presencia de este patógeno en el tejido nervioso, donde hay que recordar que GABA es un importante neurotransmisor con función inhibitoria [213].

Por último, debe destacarse el hecho de que las proteasas Lon y ClpxP son las responsables de la mayor parte de los procesos proteolíticos en bacterias. Estas proteínas, desempeñan un importante papel en la eliminación de proteínas mal plegadas o desnaturalizadas, pero también en la regulación, por ejemplo, controlando la disponibilidad de determinadas proteínas reguladoras [214,215]. Algunos trabajos demuestran la implicación de esta proteasa en la regulación negativa de procesos de QS en *Pseudomonas aeruginosa* [216], a la vez que presenta relación con sistemas toxina-antitoxina, encargándose de la degradación de la antitoxina en determinadas situaciónes [217].

6. Capítulo 4: Sistemas toxina-antitoxina

6.1. Introducción

Los sistemas toxina antitoxina (STA) son pequeños elementos genéticos, que codifican para la toxina y su antitoxina asociada. Ambos elementos se expresan e interactúan entre ellos, estableciéndose una situación en la que la antitoxina bloquea el efecto de la toxina. En muchas ocasiones, bajo condiciones de estrés, la menor estabilidad de la antitoxina promueve su degradación a una mayor velocidad, liberando a la toxina y permitiendo ejercer su función [218]. Los STA se encuentran ampliamente distribuidos tanto en bacterias como arqueas [219] y se clasifican en diferentes tipos en función de sus características. Los sistemas tipo I se definen como aquellos en los que la toxina es inhibida por un ARN antisentido a nivel de ARNm [220]; en los sistemas tipo II ambos elementos se traducen a proteínas, produciéndose en este caso la inhibición de la toxina por interacción proteína-proteína [221]; en los sistemas tipo III la antitoxina es un RNA de pequeño tamaño que inhibe la acción de la toxina interactuando de forma directa con la proteína [222], en los sistemas tipo IV, designados por el STA *yee*U/*yee*V de *Escherichia coli*, la regulación implica la estabilización de otras proteínas relacionadas [223]; y finalmente los sistemas tipo V, representados por el STA *gho*S/*gho*T recientemente descrito en *E. coli*, la antitoxina actúa como una ribonucleasa que degrada el ARNm de la toxina, inhibiendo así su traducción [224]. La función de los STA se puede relacionar con la protección de elementos genéticos móviles, formación de células persistentes y supervivencia en condiciones de estrés, entre otras hipotéticas funciones [225–228].

Los STA han ido adquiriendo gran importancia debido al papel que podrían desempeñar mediante su posible implicación frente a factores adversos externos; por ejemplo, en el transcurso de una infección, debido a las funciones con las que se les relaciona. En este sentido estos sistemas se han llegado a postular como posibles dianas para el desarrollo de fármacos destinados al tratamiento de infecciones [221]. Sin embargo, y por extensión, estos sistemas no solo podrían tener un papel importante en las especies patógenas por excelencia como *M. tuberculosis*, sino que también podrían ser clave en el caso de infecciones desarrolladas por patógenos oportunistas del grupo de micobacterias de crecimiento rápido (MCR).

En base a los antecedentes apuntados, el objetivo del presente capítulo es la identificación de este tipo de elementos genéticos en los genomas de las cepas del grupo MCR secuenciadas, utilizando como modelo comparativo la cepa *M. tuberculosis* CR-UIB2, con el fin de contrastar lo que se puede encontrar en el genoma de un patógeno primario con respecto a los genomas de potenciales patógenos oportunistas. Sin embargo, y tal como se apuntaba sobre la funcionalidad de determinados genes en el análisis del pangenoma en el Capítulo 3, encontrar operones o elementos genéticos de este tipo en la anotación de un genoma no es suficiente, se debe demostrar experimentalmente que es funcional. Por este motivo, partiendo de las secuencias proteicas de potenciales toxinas y antitoxinas se realizó una primera búsqueda en las anotaciones de los genomas de dominios proteicos o unidades modulares de ambas proteínas que pudieran llevar a cabo una función determinada en clara referencia a la presencia de componentes estables de sus estructuras. La finalidad última era la de determinar si, en principio, tenían los dominios o combinaciones de dominios necesarios para llevar a cabo la función que se les presupone como STA. Los esfuerzos se centraron posteriormente en el desarrollo y adaptación de un protocolo de ensayo experimental para determinar su funcionalidad. Finalmente, se realizó un estudio más profundo de la secuencia y estructura primaria, secundaria y terciara de los componentes de aquellos STA que mostraron ser funcionales in vitro. De esta forma se pretendió dejar patente que estas proteínas tenían la estructura y los elementos característicos del grupo de proteínas al que pertenecen, así como demostrar experimentalmente su funcionalidad.

6.2. Materiales y Métodos

6.2.1. Identificación de sistemas toxina-antitoxina

La catalogación de los STAs se realizó, como primera aproximación, mediante la búsqueda de las palabras clave "toxin" y "antitoxin" en los archivos gbk de los genomas anotados con Prokka v1.10 [75]. La búsqueda se realizó utilizando el programa UGENE v1.16.1 [110]. Además, a partir de las secuencias nucleotídicas de los genomas en formato FASTA, se realizó una búsqueda en bases de datos especializadas como RASTA-bacteria [229], una herramienta web diseñada para la identificación de loci toxina-antitoxina en procariotas. Finalmente, se efectuaron análisis con BLAST y la secuencia completa de los genomas

contra la base de datos TADB (del inglés *Toxin-Antitoxin DataBase*), especializada en secuencias de STA de tipo II [230].

6.2.2. Clonación de los sistemas toxina antitoxina

6.2.2.1. Vectores de expresión

La elección de los vectores de expresión para la clonación de los STA se basó en los trabajos de Sberro *et al.* (2013) [231]. Brevemente, se seleccionaron los vectores pBAD/HA (Invitrogen™, Carlsbad, California, EEUU) para la clonación de la antitoxina y el vector pRSF-Duet™ (Merck Millipore, Billerica, Massachusetts, EEUU) para la clonación de la toxina. Este segundo vector presenta dos regiones de clonación múltiple o *Multiple Clonning Site* (MCS) (Figura 6.1).

Vector	Tamaño (bp)	Inductor	Casa comercial
pBAD/HA	4102	Arabinosa	InvitrogenTM (V43001)
pRSFDuetTM-1	3829	IPTG	Novagen® (71341-3)

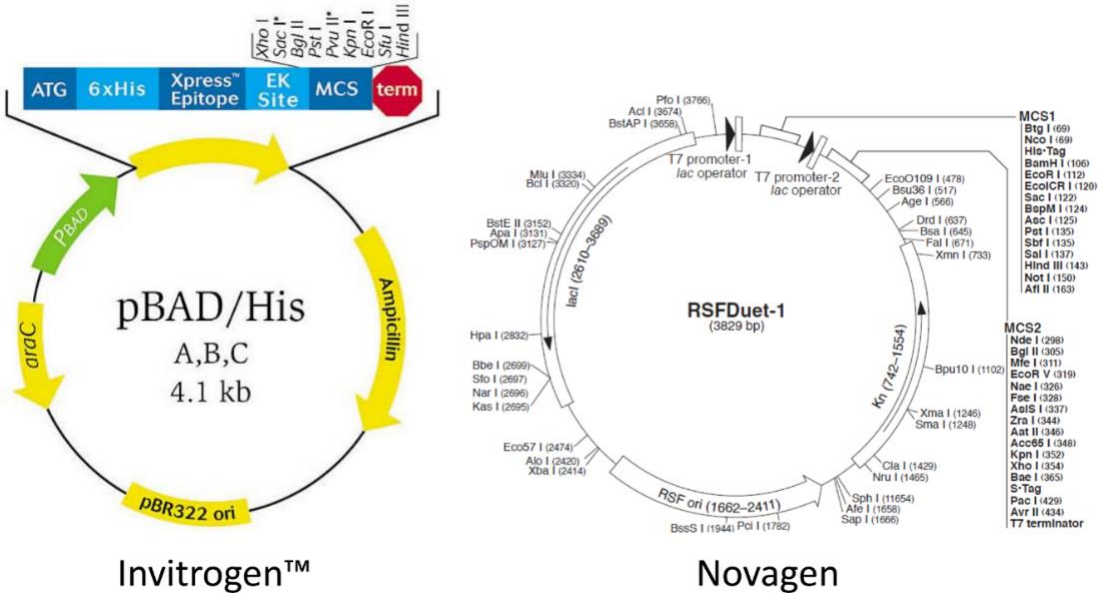


Figura 6.1. Características generales y mapas de los vectores de expresión seleccionados. Se indica el tamaño total en pares de bases (pb) y el inductor al que responden los respectivos promotores (imágenes de los mapas obtenidas a partir de los manuales de las respectivas casas comerciales).

La clonación de antitoxinas en el vector pBAD/HA se realizó en los sitios *Nco*I/*Hind*III. Para la clonación en el vector pRSF-DuetTM-1 de las toxinas y potenciales terceros

elementos asociados al sistema se escogieron los sitios de restricción *Nco*I/*Hind*III en la región MCS1 o *Nde*I/*Xho*I en la región MCS2 (Figura 6.1). La elección de clonar en MCS1 o MCS2 dependió del análisis de dianas de restricción presentes en los genes de interés. Las dianas *Nco*I y *Nde*I regeneran el codón de inicio ATG entre el vector y el inserto. De esa manera se asegura la clonación sin elementos anteriores a la secuencia de interés.

6.2.2.2. Diseño de cebadores

El diseño se realizó de forma manual sobre los codones de inicio y fin de cada gen en cuestión, incluyendo en los extremos 5' de cada cebador las dianas de restricción pertinentes en cada caso. El análisis de la compatibilidad de cebadores se realizó con el programa AmplifX v1.5.4 (http://crn2m.univ-mrs.fr/pub/amplifx-dist). El diseño y simulaciones virtuales de las construcciones resultantes se realizó con PlasmaDNA [232] y la conservación de los marcos de lectura se confirmaron a nivel de secuencia nucleotídica con UGENE. Al mismo tiempo se realizó una PCR virtual sobre los genomas y las construcciones con la ayuda del programa FastPCR [233], utilizando en el caso de estas últimas un cebador *forward* del vector (Tabla 6.1) y un cebador *reverse* del inserto como control del tamaño estimado del amplicón que se debía obtener en los clones potencialmente correctos.

Tabla 6.1. Secuencia de los cebadores comerciales utilizados. Se indica la diana para la que están diseñados y la referencia donde se encuentran sus secuencias.

Cebador	Secuencia (5'->3')	Diana	Referencia
pBADHA-SF	ATGCCATAGCATTTTTATCC	pBAD/HA	Invitrogen
ACYCDuetUP1-F	GGATCTCGACGCTCTCCCT	pRSF-Duet (MCS1)	Merck Millipore
DuetUP2	TTGTACACGGCCGCATAATC	pRSF-Duet (MCS2)	Merck Millipore

6.2.2.3. Amplificación por PCR

Se aplicó un programa de PCR general de 35 ciclos para todos los STAs (desnaturalización a 95 °C, 1 minuto; hibridación a 55 °C, 1 minuto; elongación a 72 °C, 1 minuto 30 segundos) incluyendo un paso inicial de desnaturalización a 95 ºC durante 10 minutos y una elongación final de 72 ºC durante 10 minutos en un termociclador Mastercycler Personal (Eppendorf, Madrid, España). Este programa de PCR se modificó únicamente para amplicones de tamaño superior a 1 kb, en cuyos casos se incrementó el tiempo de

elongación en 30 segundos La amplificación se realizó con la polimerasa DFS-Taq (BIORON, Ludwigshafen, Alemania). Las reacciones de PCR se prepararon partiendo de 10-20 ng de ADN, una concentración final de tampón de PCR 1x, 0,2 mM de dNTPs (0,05 mM cada uno) y 2,5 U DFS-Taq polimerasa. Las reacciones se ajustaron con agua mili-Q a un volumen final de 50 µl.

6.2.2.4. Electroforesis en geles de agarosa y secuenciación Sanger

Los productos de PCR obtenidos se comprobaron mediante electroforesis en geles de agarosa 1,5 % (p/v) (100 V, 35 minutos) y se purificaron con el kit comercial *Illustra GFX PCR DNA and Gel Band Purification* (GE Healthcare, Chicago, Illinois, EEUU). Una vez purificados, los productos de PCR obtenidos se secuenciaron por el método de Sanger con el kit *BigDye® Terminator v3.1 Cycle Sequencing* (Applied Biosystems™, Foster City, California, EEUU), tal y como se explicó en el Capítulo 1.

6.2.2.5. Clonación

La digestión de los amplicones y vectores se llevó a cabo con las enzimas de restricción pertinentes *Nco*I, *Hind*III, *Nde*I, *Xho*I (TaKaRa, Otsu, Shiga, Japón) durante 3 h a 37 ºC en un volumen de reacción de 20 µl, siguiendo las especificaciones del fabricante. La inactivación de las enzimas de restricción después del proceso de digestión se llevó a cabo a 65 ºC durante 10 minutos. No se realizaron dobles digestiones simultáneas para asegurar la completa acción de cada enzima durante el proceso. En su lugar, los productos de digestión resultantes fueron sucesivamente purificados después de cada proceso de digestión individual, obteniendo un producto limpio para ser utilizado en los sucesivos pasos.

La ligación de los productos digeridos se realizó con la enzima ligasa T4 (TaKaRa, Otsu, Shiga, Japón), en un volumen final de reacción de 20 µl y siguiendo las recomendaciones del fabricante. Para ello se calcularon las cantidades equimolares de vector/inserto y se aplicó una relación 1/3, partiendo de 50 ng de vector (Figura 6.2). La reacción de ligación se realizó a 16 ºC durante toda la noche y finalmente se inactivó a 65 ºC durante 10 minutos. A modo de control del proceso se realizaron PCRs a partir de las reacciones de ligación,

así como la secuenciación del producto resultante siguiendo las indicaciones descritas en el apartado anterior.

$$inserto(ng) = 3\ x\ \left(\frac{Longitud\ del\ inserto\ (pb)}{Longitud\ del\ vector\ (pb)}\right)\ x\ 50\ ng\ de\ vector$$

Figura 6.2. Fórmula utilizada para el cálculo de las cantidades equimolares con un ratio 1/3 (inserto/vector) a partir de 50 ng de vector.

6.2.2.6. Transformación

El producto de ligación obtenido se utilizó para la transformación de células competentes de alta eficiencia *Escherichia coli* BL21(DE3) pLysS (Invitrogen™, Carlsbad, California, EEUU) siguiendo el protocolo de transformación por choque térmico y las condiciones recomendadas para esta cepa por la propia compañía. Después de la transformación, las células se sembraron sobre placas de agar Luria-Bertani (LB) suplementadas con 30 µg/ml de kanamicina (Km) para pRSF-DUET, 50 µg/ml de ampicilina (Ap) para pBAD-HA y 34 µg/ml de cloranfenicol (Cm). Este tercer antibiótico se añade debido a que la cepa comercial contiene por defecto el plásmido pLysS, el cual es necesario asegurar la inhibición de la expresión basal de T7 RNA polimerasa con el fin de evitar que el inserto sometido al control del promotor T7-lac se exprese hasta que en el medio no aparezca el inductor adecuado. Este es el caso de los insertos que se clonaron en el vector pRSF-Duet™. Las placas se incubaron a 37 ºC durante 18 h.

A partir de los clones obtenidos se seleccionaron un mínimo de 10 para su replicación sobre placas nuevas de LB, inoculándolos simultaneamente en tubos con caldo LB; suplementados en ambos casos siempre con los correspondientes antibióticos. Las placas se incubaron a 37 ºC durante 18 horas, mientras que los tubos se incubaron a 30 ºC en agitación orbital (180 rpm) durante 18 horas. A partir de los cultivos líquidos se realizó una extracción de plásmido con el kit de extracción *UltraClean® 6 Minute Mini PlasmidPrep (*MO BIO Laboratories, Carlsbad, California, EEUU).

6.2.2.7. Análisis de los clones

A partir de 1 µl de cada extracción de plásmido se realizó una PCR control con un cebador del vector y un segundo del inserto, siguiendo el mismo protocolo de PCR y confirmación

por electroforesis descrito anteriormente. A partir de los clones que presentaban el tamaño esperado para cada inserto, se purificaron los productos de PCR obtenidos y se secuenciaron mediante la metodología de Sanger para confirmar la secuencia y el patrón de lectura. Los clones de interés una vez confirmados fueron guardados en glicerol 20 % (v/v) a -80 ºC.

6.2.3. Ensayo de la funcionalidad de los sistemas toxina-antitoxina

A partir de un clon confirmado por secuencia y de la cepa utilizada como hospedadora, BL21 (DE3) pLysS (sin transformar) se prepararon sendos preinóculos en tubos con 5 ml de LB. Ambos cultivos se suplementaron con Km (30 µg/ml), Ap (50 µg/ml) y Cm (34 µg/ml) en el primer caso y Cm (34 µg/ml) en el segundo, incubándose a 30 ºC en agitación (180 rpm) 18 h. Paralelamente se prepararon cuatro matraces Erlenmeyer con 50 ml de LB esterilizados previamente, de los cuales uno se suplementó con Cm y los tres restantes con los tres antibióticos y concentraciones antes apuntadas. A partir de los preinóculos, se utilizaron 100 µl de la cepa sin transformar y se inocularon en el Erlenmeyer con Cm; mientras que los otros tres se inocularon cada uno con 100 µl del preinóculo del clon seleccionado. Una vez realizados los inóculos, se incubaron a 37 ºC y en agitación a 180 rpm. El seguimiento del crecimiento se realizó mediante lecturas de densidad óptica (DO) a una longitud de onda de 600 nm, cada 30 minutos con 1 ml de cultivo utilizando un espectrofotómetro *Ultrospec 100 Pro* (GE Healthcare, Chicago, Illinois, EEUU). A las dos horas y media del inicio del experimento, con los cultivos en plena fase exponencial, se indujeron los diferentes elementos para establecer las siguientes condiciones experimentales:

- Control (inoculado con la cepa no transformada y sin inductor),
- Arabinosa 0,3 % (p/v) concentración final, para la inducción de la antitoxina,
- IPTG 0,2 mM concentración final, para la inducción de la toxina,
- Arabinosa + IPTG utilizando las concentraciones anteriores para la inducción de ambos elementos.

El experimento se continuó en las mismas condiciones hasta completar 7 h desde el inicio del mismo. Este protocolo se repitió por triplicado. Las DO registradas se utilizaron para la obtención de las curvas de crecimiento correspondientes. Los valores de DO se tradujeron a unidades formadoras de colonias por ml (UFC/ml) mediante el recuento de colonias. De forma resumida, a partir del momento de inducción (tiempo 0) se realizaron bancos de diluciones seriadas decimales a partir de los cultivos en medio líquido, sembrando por duplicado 100 µl de diluciones continuas en placas de LB suplementadas con los correspondientes antibióticos y arabinosa 0,3 % (p/v). Las placas se incubaron 18 h a 37 ºC, tras lo cual se realizaron los pertinentes recuentos. Los valores de UFC/ml calculados se utilizaron para la representación de las curvas de variación del número de UFC/ml a lo largo del tiempo.

6.2.4. Análisis de la expresión proteica

6.2.4.1. Extracción de proteínas

Partiendo de cultivos de 7 h sometidos a las mismas condiciones descritas en el apartado 6.2.3, se centrifugaron los 50 ml de cultivo a 6.000 g durante 5 minutos. El sedimento celular obtenido se resuspendió con 1 ml de Tris 50 mM pH 8,0 y se volvió a centrifugar. Finalmente, se resuspendió con 800 µl del mismo tampón. Las células se rompieron por sonicación, manteniendo la muestra en hielo para evitar la desnaturalización de las proteínas debido al calor generado, alternando pulsos de 5 segundos con descansos de 10 segundos, en un tiempo máximo de 2 minutos. La suspensión resultante se centrifugó 2 minutos a 3.000 g y el sobrenadante obtenido se volvió a centrifugar durante 30 minutos a 16.000 g. A partir del último sobrenadante obtenido, se tomaron 500 µl para los sucesivos análisis.

6.2.4.2. Electroforesis en geles de poliacrilamida

Los extractos proteicos obtenidos de las muestras correspondientes a las distintas condiciones experimentales de inducción se cargaron en geles de poliacrilamida en gradiente 4-20 % (p/v) *Amersham ECL Gel* (GE Healthcare, Chicago, Illinois, EEUU) para su separación por electroforesis horizontal en el sistema ECL *Gel Box* (GE Healthcare, Chicago, Illinois, EEUU), durante 1 hora y 30 minutos a 160 V y siguiendo las

recomendaciones del fabricante. Como marcador de peso molecular se utilizó el *Precision Plus Protein™ Dual Xtra* (Hercules, California, EEUU), el cual abarca un rango de tamaños de entre 2 y 250 kDa. Los geles se tiñeron con azul de Coomassie durante 18 horas y se destiñeron con una solución de etanol y ácido acético en agua al 5 y 7,5 % (v/v) respectivamente, a 55 ºC y agitación durante 2 horas, incluyendo dos cambios de la solución.

En aquellos casos de bandas diferenciales no bien separadas, se repitió la electroforesis en el sistema vertical *Mini-PROTEAN® Tetra Cell* (Hercules, California, EEUU), preparando geles homogéneos y con porcentajes de acrilamida del 10, 15 ó 20 % (p/v) en función del tamaño de la banda y mejor resolución obtenida para la zona en cuestión.

6.2.4.3. Espectrometría de masas MALDI-TOF (MALDI-TOF MS)

Las bandas extraídas directamente del gel se introdujeron en tubos Eppendorf de 1,5 ml. Se añadieron 100 µl de NH_4HCO_3 400 mM y 100 µl de CH_3CN. A continuación, se agitó el tubo con un agitador de tipo vórtex durante 20 minutos tras lo que se eliminó el sobrenadante y se volvió a repetir el mismo procedimiento. A continuación, se realizó un lavado con 100 µl de CH_3CN, eliminado el sobrenadante y volviendo a añadir 100 µl de la misma solución, dejando incubando durante un periodo de 10 minutos. Pasado este tiempo se eliminó el sobrenadante y se dejó secar la banda de gel. Posteriormente se enfriaron en hielo las bandas de gel y una solución de tripsina (20 µg/ml en NH_4HCO 350 mM) para minimizar la autodigestión de la tripsina durante el tiempo que tarda en penetrar el gel. El gel se rehidrató añadiendo entre 20 y 50 µl de solución de tripsina y se incubó a 4 °C durante 30 minutos para, posteriormente seguir incubando a 37 °C durante toda la noche. Se añadieron 5 µl de ácido fórmico al 5 % (v/v) y se mezcló con un agitador tipo vórtex durante 5 minutos. Finalmente, se añadieron 30 µl de solución de extracción (250 µl de CH_3CN, 25 µl de ácido trifluoracético y 225 µl de H_2O mili-Q) dejando incubar durante 10 minutos y repitiendo posteriormente este último paso una vez más.

Se depositó 1 µl del producto de digestión obtenido sobre una placa MSP 96 de acero pulido (Bruker Daltonics, Billerica, Massachusetts, EEUU) de MALDI-TOF MS, dejando secar a temperatura ambiente. Posteriormente se añadió sobre la muestra seca 1 µl de matriz (ácido 2,5-hidroxibenzoico o bien ácido β-ciano-4-hidroxicinámico en 70/30

acetonitrilo/agua con un 0,1 % v/v de ácido trifluoroacético), dejando secar nuevamente a temperatura ambiente. La placa fue analizada en un espectrómetro de masas Autoflex III MALDI-TOF/TOF (Bruker Daltonics, Billerica, Massachusetts, EEUU) utilizando el programa *Compass Flex Series* v1.4 (*flex Control* v3.4 y *flex Analyser* v3.4). Los espectros fueron calibrados utilizando el *Peptide Calibration Standard II* (Bruker Daltonics, Billerica, Massachusetts, EEUU).

Los espectros de péptidos obtenido se utilizaron para realizar una búsqueda en la base de datos UniProtKB/Swiss-Prot y en una base de datos primaria propia creada a partir de las secuencias proteicas de los componentes de los STAs funcionales, obtenidas por anotación de los genomas con Prokka v1.10. El proceso de análisis de los datos de espectrometría de masas para la identificación de proteínas se llevó a cabo a través del algoritmo Mascot (Matrix Sciences Ltd., www.matrixscience.com).

6.2.5. Caracterización estructural

Las secuencias proteicas de los STA funcionales se utilizaron para la búsquedas de secuencias homólogas con BLAST tanto en GenBank [76] (base de datos nr, secuencias no redundantes, de proteínas) como en UniProt [234], así como la búsqueda de dominios funcionales en la base de datos Pfam [112]. Partiendo del resultado obtenido con UniProt, se construyó un dendrograma con las 50 proteínas más parecidas en cada caso, utilizando el criterio de cobertura/similitud C50/S50. Las secuencias fueron alineadas con ClustalO [91]. El programa PhyML [82] se utilizó para elaborar el dendrograma de agrupaciones de secuencias, aplicando el algoritmo de estimación por máxima verosimilitud implementado en el programa y basando el soporte estadístico de las ramas en un análisis jerárquico de grupos mediante 100 *bootstraps* o soporte de agrupación.

La predicción de la estructura terciaria de las proteínas se realizó a través del servidor I-TASSER [235], una plataforma integrada para la simulación automatizada de la estructura de proteínas y predicción de su función partiendo de la secuencia de aminoácidos (estructura primaria), sobre la cual se determina la estructura secundaria de la proteína problema y, reclutando análogos estructurales de las bases de datos, proceder finalmente a realizar el cálculo de las potenciales estructuras terciarias. A través de I-TASSER, además, se obtuvo información sobre posibles funciones, centros activos y ligandos que podrían

acompañarla. La calidad del modelo predicho se determinó a través del C-Score, el valor del cual se estima entre -5 y 2, rango en el que cuanto más elevado es dicho valor más confiable es el modelo. La similitud estructural del modelo predicho con otras proteínas se mide con el TM-Score [236], según el cual un valor superior a 0,5 correspondería a una topología correcta del modelo predicho, mientras que valores inferiores a 0,17 indicarían similitudes debidas al azar. Finalmente, el alineamiento de estructuras para la identificación detallada de los posibles centros activos se realizó con el servidor *Partial Order Structure Alignment* (POSA) [237].

6.3. Resultados

6.3.1. Identificación de los sistemas toxina-antitoxina

Micobacterias del grupo MCR

En total se detectaron 23 potenciales STA a partir de los genomas analizados (Tabla 6.2). Entre ellos, tres sistemas tipo Vap (VapBC27 y VapBC28 en el genoma de la cepa tipo de *M. llatzerense* MG13^T y VapBC5 en el genoma de la cepa *M. abscessus* subsp. *bolletii* CR-UIB1). Se encontró una toxina zeta con un CDS que codificaría para una proteína hipotética (HP, del inglés *Hypothetical Protein*) corriente arriba de la misma, en una configuración que encaja con la organización propia de un STA tipo II destacada en la introducción. Así mismo, se halló una potencial toxina Doc (del inglés, *Death-on-curing*) acompañada, como en el caso anterior, de una proteína hipotética corriente arriba. Estos dos últimos posibles STA se hallaron en *Mycobacterium* sp. MHSD3. En todos los genomas a excepción de la cepa tipo de *M. llatzerense*, se encontró el STA denominado MT0933-MT0934 adoptando una organización atípica, con un gen que figura anotado como lipasa separando ambos componentes del sistema. En la cepa *M. Llatzerense* MG13^T, por su parte, se encontró un sistema MT0933-MT0934 en la organización esperada, es decir, con el CDS de la antitoxina superpuesto por su extremo 3' sobre el extremo 5' del CDS correspondiente a la toxina y sin una lipasa codificada entre ambos genes.[238].

Tabla 6.2. Potenciales STA identificados en los genomas de las cepas secuenciadas. Se incluyen así mismo aquellos potenciales sistemas cuyas toxinas han sido anotadas, pero para los cuales no se ha hallado la antitoxina correspondiente.

Cepa	Potencial STA
***M. llatzerense* MG13^T**	VapBC28 VapBC27 MT0933-34
***M. chelonae* CCUG 47445^T**	MT0933-Lipasa-Mt0934 Toxina MT0934 (sin antitoxina) x2
***M. immunogenum* CCUG 47286^T**	MT0933-Lipasa-MT0934
MG2	MT0933-Lipasa-MT0934 Toxina zeta (sin antitoxina)
MG8	MT0933-Lipasa-MT0934 Toxina zeta (sin antitoxina)
MHSD2	MT0933-Lipasa-MT0934 MT0934S Toxina zeta (sin antitoxina)
MHSD3	MT0933-Lipasa-MT0934 MT0934S Toxina zeta (sin antitoxina) Hipotética antitoxina-Toxina zeta Phd-Doc
CR-UIB1	MT0933-Lipasa-MT0934 MT0934S Toxina zeta (sin antitoxina) VapBC5

Por otra parte, también se encontraron toxinas MT0934 y toxinas zeta para las cuales no se localizaron las antitoxinas asociadas en su proximidad después de analizar las anotaciones en un margen de hasta 5 kb tanto corriente arriba como abajo del gen en cuestión, buscando nombres clave relacionados con antitoxinas y haciendo un análisis BLAST de las secuencias proteicas y nucleotídicas, especialmente de aquellos CDS anotados como proteínas hipotéticas. El margen de acotación aplicado se puede considerar suficiente debido a que los genes implicados de los sistemas STA conocidos se suelen encontrar a una distancia comprendida entre -20 y +30 pb [229] y las secuencias nucleotídicas vecinas no parecen codificar para potenciales genes que tengan relación alguna con antitoxinas.

Las secuencias proteicas de cada componente se analizaron con el objetivo de detectar dominios potencialmente relacionados con la actividad de los STA (base de datos Pfam) y para la búsqueda de proteínas homólogas en otros genomas (UniProt). En este segundo caso el interés principal fue buscar secuencias similares en otras especies o patógenos de referencia; en muchos de los cuales la función de estos sistemas o similares ya ha sido demostrada experimentalmente.

Para la presentación de estos resultados los sistemas se clasificaron en tres grupos:

- Grupo 1: Sistemas de dos componentes bien definidos (Tabla 6.3).
- Grupo 2: Sistemas MT0933-MT0934 con lipasa (Tabla 6.4).
- Grupo 3: Hipotéticas toxinas huérfanas, sin antitoxina asociada (Tabla 6.5).

Cabe hacer hincapié en el hecho de que en caso del sistema VapBC27 de la cepa tipo de *M. llatzerense* (Tabla 6.3), al hacer un análisis BLAST con la secuencia aminoacídica de VapC27 y VapB27 en UniProt se observó una secuencia más corta con respecto a las proteínas homólogas presentes en otros genomas. Además, en *M. llatzerense* MG13^{T}, el gen de la proteína VapC27 se halló separado por unas 130 pb de la antitoxina, hecho que no es habitual en sistemas del tipo VapBC27 de genomas de otras especies, como *M. tuberculosis*, donde incluso aparecen solapados [238]. Estos resultados sugirieron que la anotación podría ser incompleta. Para corroborar esta hipótesis se decidió ampliar, codón a codón, la secuencia corriente arriba e ir probando si, al añadir nuevos aminoácidos con cada triplete, la cobertura del alineamiento con respecto a las secuencias encontradas en UniProt mejoraba. Se observó que a medida que se añadían aminoácidos a la secuencia, éstos en su mayoría coincidían en los alineamientos con los aminoácidos presentes en esas mismas posiciones de las secuencias más parecidas halladas en UniProt (Figura 6.3). La secuencia final así obtenida, teóricamente completa al compararla mediante el alineamiento con sus homólogas más próximas, fue la finalmente utilizada para la realización de los sucesivos análisis.

Mediante el acceso a la colección de alineamientos múltiples de secuencias y modelos ocultos de Markov (HMM, del inglés *Hidden Markov Model*) que cubre buena parte de dominios proteicos y familias guardados en la base de datos Pfam, se buscaron los dominios funcionales de este y del resto de STA del Grupo 1 (Tabla 6.3). Resultado de ello

se detectó en las toxinas del tipo Vap el denominado dominio PIN, de unos 130 aminoácidos en longitud y que se caracterizaron por la presencia de tres residuos ácidos estrictamente conservados.

Tabla 6.3. Caracterización de la secuencia de toxinas y antitoxinas de los operones bien definidos (Grupo 1). Se indica el nombre de la familia asignada por Pfam para cada proteína, así como el nombre de la proteína más similar según los resultados de BLAST en UniProt.

Cepa	Componente TA	Familia (Pfam)	UniProt
MG13^T	VapB27	MazE_antitox	Antitoxina VapbB27
MG13^T	VapC27	PIN	Ribonucleasa VapC27
MG13^T	VapB28	PSK_trans_fac	Antitoxina VapB28
MG13^T	VapC28	PIN	Ribonucleasa VapC28
CR-UIB1	VapB5	PhdYeFM_antitox	Antitoxina putativa VapB5
CR-UIB1	VapC5	PIN	Ribonucleasa VapC5
MHSD3	Proteína hipotética	///	///
MHSD3	Doc	Fic/Doc	Toxina Doc
MHSD3	HP	ParD-like	HP
MHSD3	Toxina zeta	AAA_33	Proteína no caracterizada HI_1395
MG13^T	MT0933	MT0933_antitox	Antitoxina Rv0909
MG13^T	MT0934	Poliquetido Cyc2	Toxina Rv0910

Figura 6.3. Variación de la secuencia aminoacídica de las proteínas VapB27 y VapC27 antes y después de completar la secuencia. La secuencia añadida manualmente se resalta en morado, mientras que la secuencia original se encuentra resaltada en azul.

Por lo que respecta a las antitoxinas, la proteína VapB27 se incluyó dentro de la familia MazE_antitox, en la cual el modelo de referencia es la antitoxina MazE del STA MazEF. La antitoxina VapB28 se incluyó dentro de un grupo de factores de transcripción relacionados con operones PSK (del inglés *Post-Segregational-Killing*), otro de los STA tipo II en los que las proteínas de esta familia actúan como antídoto contra el gen de una toxina asociada. Por último, la antitoxina VapB5 presentó dominios propios de la familia PhdYeFM_antitox, otro grupo de antitoxinas pertenecientes a STA de tipo II.

La toxina Doc fue perfectamente catalogada como perteneciente a la familia Fic/Doc. Los genes que codifican para las toxinas Doc van asociados a una antitoxina Phd (del inglés *Protection-host-death*). En este caso, corriente arriba del gen *doc*, se encontró una proteína hipotética que encajaba en los parámetros de tamaño y posición en el genoma que definirían su antitoxina asociada, pero en la que no se encontró ningún tipo de dominio que permitiera su catalogación.

La toxina zeta se clasificó dentro de la familia AAA_33, constituida por proteínas que presentan el dominio AAA (del inglés *ATPases Associated with diverse cellular Activities*). No obstante, en esta familia se incluye un conjunto de proteínas que difieren mucho entre ellas en cuanto a actividad, estabilidad y mecanismo de acción. Por su parte, la HP codificada corriente arriba del gen de la toxina zeta, se clasificó dentro de la familia de proteínas tipo ParD, antitoxina del sistema ParDE, por lo que podría tratarse de la antitoxina asociada a dicha toxina zeta.

Por último, la pareja MT0933-MT0934 de *M. llatzerense* MG13^{T} incluye una antitoxina clasificada dentro de la familia denominada MT0933, mientras que la toxina MT0934 se engloba en una familia de enzimas denominadas poliquétido ciclasas, implicadas en la síntesis de poliquétidos. Como se destacó en el apartado anterior, los genes que codifican para las proteínas MT0933 y MT0934 también se encontraron en una configuración diferente en el resto de genomas secuenciados, constituida por tres genes, que se repitió dentro del grupo *abscessus-chelonae-immunogenum*. Se trata de los genes de una hipotética toxina MT0934 y una hipotética antitoxina MT0933, entre los cuales se encontró intercalado el gen de una lipasa secretora. Los resultados obtenidos en Pfam y UniProt para este último grupo fueron prácticamente idénticos en todos los casos (Tabla 6.4), hecho que

concuerda con la similitud en secuencia existente entre las proteínas del mismo tipo (Tabla suplementaria 2, Anexo 3). Los elementos más similares según UniProt, pertenecen a la cepa *M. tuberculosis* H37Rv, aunque con porcentajes de identidad inferiores al 50 % en el 50 % de la secuencia.

Tabla 6.4. Caracterización de la secuencia de los genes que conforman el potencial sistema de tres componentes. Se indica la cepa de procedencia, el componente TA en cuestión, así como la familia a la que se ha asignado según Pfam y el nombre de la proteína más similar según UniProt.

Cepa	Componente TA	Pfam	UniProt
CCUG 47445^T	MT0933	MT0933_antitox	Antitoxina Rv0909
CCUG 47286^T	MT0933	MT0933_antitox	Antitoxina Rv0909
CR-UIB1	MT0933	MT0933_antitox	Antitoxina Rv0909
MG2	MT0933	MT0933_antitox	Antitoxina Rv0909
MG8	MT0933	MT0933_antitox	Antitoxina Rv0909
MHSD2	MT0933	MT0933_antitox	Antitoxina Rv0909
MHSD3	MT0933	MT0933_antitox	Antitoxina Rv0909
CCUG 47445^T	Lipasa	Lipasa secrectora	Probable lipasa inactiva Rv1529c
CCUG 47286^T	Lipasa	Lipasa secrectora	Probable lipasa inactivaRv1529c
CR-UIB1	Lipasa	Lipasa secrectora	Probable lipasa inactivaRv1529c
MG2	Lipasa	Lipasa secrectora	Probable lipasa inactivaRv1529c
MG8	Lipasa	Lipasa secrectora	Probable lipasa inactivaRv1529c
MHSD2	Lipasa	Lipasa secrectora	Probable lipasa inactivaRv1529c
MHSD3	Lipasa	Lipasa secrectora	Probable lipasa inactivaRv1529c
CCUG 47286^T	MT0934	Poliquetido Cyc2	Toxina Rv0910
CR-UIB1	MT0934	Poliquetido Cyc2	Toxina MT0934
MG2	MT0934	Poliquetido Cyc2	Toxina Rv0910
MG8	MT0934	Poliquetido Cyc2	Toxina Rv0910
MHSD2	MT0934	Poliquetido Cyc2	Toxina Rv0910
MHSD3	MT0934	Poliquetido Cyc2	Toxina Rv0910

Por último, se analizó con Pfam y UniProt el conjunto de proteínas anotadas como MT0934 y toxinas de tipo zeta para las cuales no se encontró antitoxina asociada (Tabla 6.5). En estas bases de datos ambos tipos volvieron a ser clasificadas en las mismas familias que en el caso de MT0934 y toxinas zeta anteriores. Sin embargo, en este caso las toxinas zeta fueron notablemente diferentes en el sentido de que mostraron una secuencia de más de 1 kb, bastante más larga de lo que cabría esperar en un STA. Este hecho implicaba la no concordancia con las características básicas de los STA conocidos, donde los genes nunca llegan a alcanzar tamaños tan grandes.

Tabla 6.5. Caracterización de la secuencia de los genes anotados como toxinas para los que no se ha encontrado antitoxina asociada. Se indica la cepa de procedencia, el componente en cuestión, así como la familia a la que se ha asignado según Pfam y el nombre de la proteína más similar según UniProt.

Cepa	Componente TA	Familia	UniProt
CCUG 47445^{T}	MT0934S	Poliquetido Cyc2	Proteína no caracterizada MB1573
CCUG 47445^{T}	MT0934S	Poliquetido Cyc2	Toxina Rv0910
CR-UIB1	MT0934S	Poliquetido Cyc2	Toxina Rv0910
MHSD3	MT0934S	Poliquetido Cyc2	Toxina Rv0910
CR-UIB1	Toxina zeta	AAA_3	Proteína no caracterizada Mb2027c
MG2	Toxina zeta	AAA_3	Proteína no caracterizada Mb2027c
MG8	Toxina zeta	AAA_3	Proteína no caracterizada Mb2027c
MHSD2	Toxina zeta	AAA_3	Proteína no caracterizada Mb2027c
MHSD3	Toxina zeta	AAA_3	Proteína no caracterizada Mb2027c

***Mycobacterium tuberculosis* CR-UIB2**

En el genoma de la cepa CR-UIB2, aislada en la Clínica Rotger de un caso clínico de tuberculosis extrapulmonar, o no respiratoria, se encontraron hasta un total de 68 STA distribuidos en distintos tipos (Figura 6.4), lo que supone un número muy superior a los encontrados tanto individualmente como en el conjunto de todos los otros genomas de MCR incluidos en este estudio. Destaca el elevado número de STAs del grupo VapBC (se contabilizaron hasta 46) con respecto a los demás tipos que se encontraron, suponiendo el 67,6 % del total de sistemas determinados en este genoma. Además, se encontró un sistema de tres componentes y que se enmarcaría dentro de los conocidos como TAC (del inglés *Toxin-Antitoxin-Chaperone*) y que están caracterizados por presentar una proteína chaperona que resultaría indispensable para el correcto funcionamiento del STA, facilitando el plegamiento y protegiendo a la antitoxina de la degradación [239].

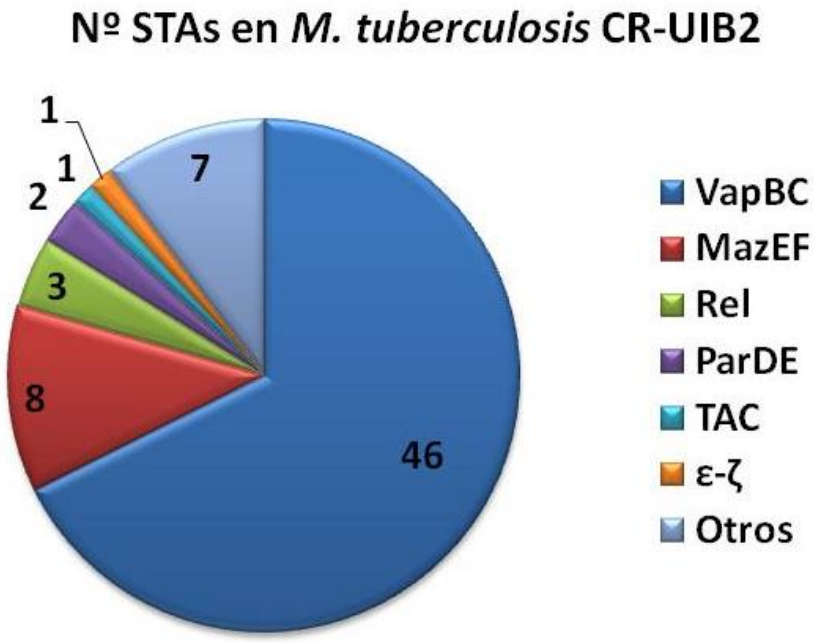


Figura 6.4. Distribución de los 68 STAs de CR-UIB2 de acuerdo con los distintos tipos encontrados en su genoma.

6.3.2. Clonacion de los sistemas toxina-antitoxina

Los cebadores diseñados y utilizados para amplificar por PCR los genes de los sistemas STA detectados se recogen en la Tabla suplementaria 1 del anexo 3.

Mediante las simulaciones *in silico* de las PCRs con el programa FastPCR, se estimaron los tamaños de cada producto, así como las condiciones de PCR a aplicar. Los datos obtenidos *in silico* se pudieron confirmar con los resultados obtenidos al visualizar la movilidad electroforética de los productos de PCR en los geles de agarosa. La estimación de los tamaños para cada amplicón, tanto el del amplificado inicial a partir de ADN cromosómico como el del obtenido en la amplificación control realizada a partir de los clones, se utilizó para confirmar los tamaños obtenidos en los productos de PCR (Figura 6.5). Los insertos amplificados a partir del ADN plasmídico de los clones y con el tamaño esperado fueron confirmados por secuenciación, no detectándose cambios con respecto a la secuencia original previamente obtenida en el genoma, ni alteraciones del patrón de lectura. Estos clones fueron los seleccionados para evaluar su funcionalidad en el ensayo experimental.

6.3.3. Ensayo de la funcionalidad de los sistemas toxina-antitoxina

A partir de los clones confirmados, se realizaron los respectivos ensayos de comprobación para la supuesta toxicidad de la toxina, la inocuidad de la antitoxina por sí sola, así como de su posible capacidad para neutralizar los efectos de la toxina.

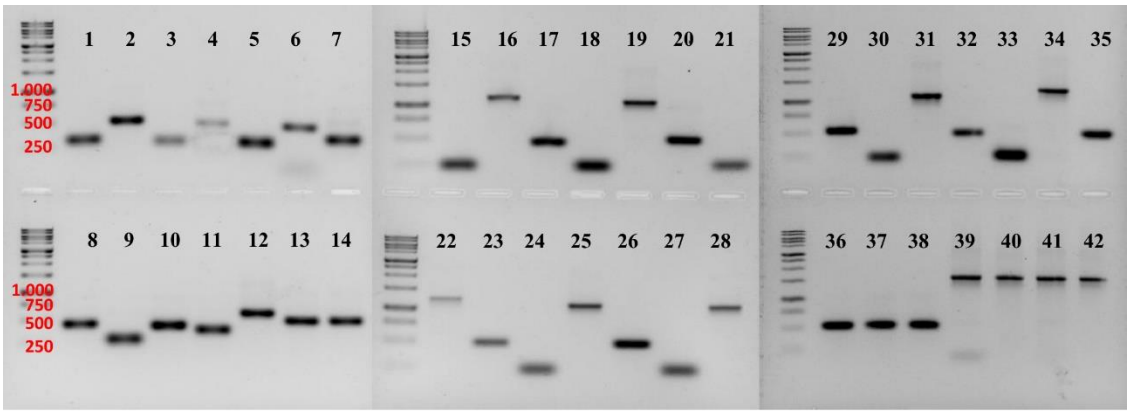
Aunque se utilizaron las mismas condiciones y se aplicó el mismo procedimiento experimental en cada caso, tal como se ha descrito en la sección de materiales y métodos, no todos los componentes de toxina de los STA analizados mostraron tener el efecto tóxico esperado durante el crecimiento (Tabla 6.6). Sólo tres de los 23 potenciales sistemas identificados mostraron un comportamiento compatible con la expresión de un STA y en las distintas condiciones experimentales aplicadas. En todos los casos se realizaron tres réplicas para confirmar la reproducibilidad de los efectos observados. En aquellos sistemas aparentemente funcionales se realizaron recuentos en placa.

En el caso del STA VapBC27, las curvas pertenecientes al control, a la expresión de la antitoxina y a la expresión de ambos elementos evolucionaron de forma muy similar hasta llegar a su fase estacionaria, aunque la evolución de la curva del control terminó por encima de las dos anteriores. Por su parte, la curva donde se refleja la evolución del cultivo donde se indujo la expresión de la toxina, su crecimiento quedó estancado en valores de DO por debajo de 1 al cabo de dos horas y media después del inicio de la inducción. Cuando se tradujeron estos valores de DO a UFC por ml, basados en el recuento de colonias, se observó un incremento progresivo del número de UFC/ml en el control, la expresión de la antitoxina y de ambos elementos del STA, aunque con un ligero repunte seguido de un descenso en lo que se refiere a los dos últimos contajes de los cultivos en los que se utilizaron ambos inductores. En cualquier caso, estos tres cultivos transcurridas 7 horas, al final del experimento, presentaron recuentos del orden de 10^8 UFC/ml. Por su parte, en el cultivo donde se expresó la toxina, a pesar del segundo punto (4 horas), donde se vio una evolución similar a los otros 3 casos, el recuento descendió y se mantuvo en el orden de 10^6 UFC/ml durante el resto del experimento (Figura 6.6A).

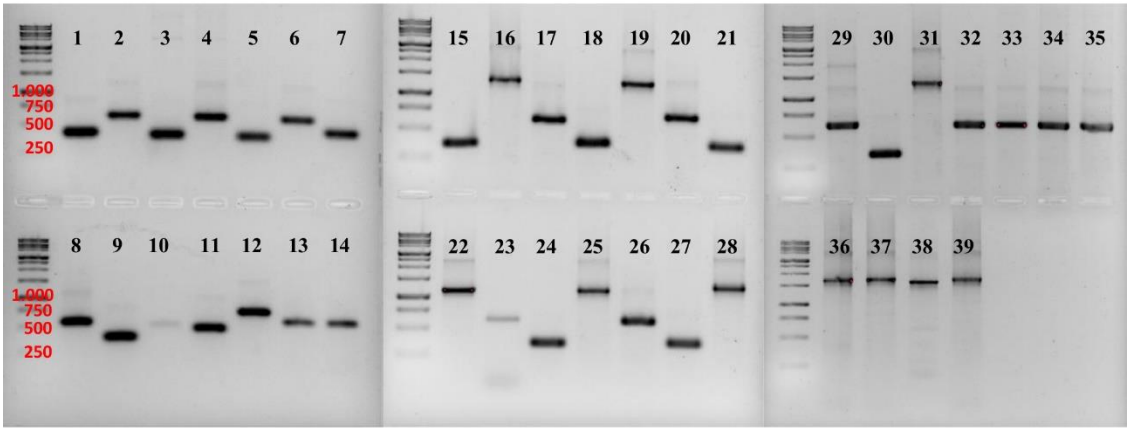
En el caso del STA VapBC28 quedaron patentes las diferencias entre la evolución de los cuatro cultivos en paralelo. Los cultivos control y de inducción de antitoxina evolucionaron de forma casi idéntica a lo largo de las 7 horas de duración del experimento. El cultivo donde se indujeron ambos elementos mostró un ligero retraso con respecto a los dos

anteriores. Sin embargo, en el cultivo donde se indujo la toxina volvió a detenerse su incremento trascurridas 2 horas después de la inducción y se mantuvo a una DO entre 0,8 y 0,9. Como se pudo comprobar en el recuento de UFC/ml, los experimentos basados en los cultivos control y de inducción de antitoxina no parecieron tener dificultades para incrementar la población, llegando a las 7 horas con dos órdenes de magnitud por encima del contaje inicial, alcanzando las 10^8 UFC/ml. Por su parte la expresión de la toxina desencadenó una serie de altibajos en la evolución del número de UFC/ml, alcanzando el mismo nivel inicial después del descenso del primer contaje, es decir, en torno a las 10^6 UFC/ml. En el caso del experimento del cultivo induciendo los dos elementos del STA VapBC28, a pesar de que en el segundo contaje (1 hora después de la inducción) se observó una tendencia ascendente en el número de UFC/ml, esta tendencia no se mantuvo y cayó a niveles similares a los del cultivo donde se expresó la toxina, prosiguiendo con una evolución similar en ambos casos (Figura 6.6B).

Por último, en el caso del STA proteína hipotética-toxina de tipo zeta, las curvas obtenidas al representar las densidades ópticas determinadas a lo largo de las 7 horas confirmaron la existencia de diferencias en la evolución de los cultivos al someterlas a distintas condiciones. Así, tanto en el control como en el cultivo donde se indujo la expresión de la antitoxina se observó un crecimiento aparentemente sin problemas, aunque la inducción de la antitoxina provocó que se llegase antes a la fase estacionaria. Por otro lado, el cultivo donde se indujo la expresión de ambos elementos del STA pareció tener más dificultades que los dos anteriores para crecer, pero llegó a alcanzar densidades ópticas similares a las del caso de la inducción de la antitoxina una vez alcanzado el final del experimento. Por último, el cultivo donde se expresó la toxina vio prácticamente anulado su crecimiento inmediatamente después de añadir el inductor (IPTG), llegándose a densidades ópticas por debajo de 0,2 al final del experimento. El cultivo con la antitoxina inducida siguió una evolución similar a la del control en cuanto a recuentos. En el caso de la expresión de ambos elementos del STA, aunque la evolución no fue exactamente igual con respecto al control, si hubo un incremento en el número de UFC en dos órdenes de magnitud con respecto al contaje inicial. Estos tres cultivos terminaron con recuentos que alcanzaron las 10^8 UFC/ml. Por su parte, en el experimento donde se indujo únicamente la toxina el cultivo evolucionó de forma contraria a los otros tres casos, disminuyendo progresivamente hasta alcanzar las 10^5 UFC/ml, un orden de magnitud por debajo del contaje inicial (Figura 6.6C).



1. VapB27 (Mll)
2. VapC27 (Mll)
3. VapB28 (Mll)
4. VapC28 (Mll)
5. MT0934 (Mll)
6. MT0933 (Mll)
7. VapB5 (UIB1)
8. VapC5 (UIB1)
9. HP-phd (D3)
10. doc (D3)
11. HP-antiZ(D3)
12. Toxina Z (D3)
13. MT0934-S (Mch)
14. MT0934-S2(Mch)
15. MT0933(Mch)
16. Lipasa(Mch)
17. MT0934(Mch)
18. MT0933 (Mim)
19. Lipasa (Mim)
20. MT0934 (Mim)
21. MT0933 (MG2)
22. Lipasa (MG2)
23. MT0934(MG2)
24. MT0933(MG8)
25. Lipasa (MG8)
26. MT0934 (MG8)
27. MT0933 (D2)
28. Lipasa (D2)
29. MT0934(D2)
30. MT0933(D3)
31. Lipasa (D3)
32. MT0934 (D3)
33. MT0933 (UIB1)
34. Lipasa (UIB1)
35. MT0934 (UIB1)
36. Toxina Z*(MG2)
37. Toxina Z*(MG8)
38. Toxina Z* (D2)
39. Toxina Z* (D3)
40. MT0934 -S(D2)
41. MT0934 -S(D3)
42. MT0934 -S(CR)



1. VapB27 (Mll)
2. VapC27 (Mll)
3. VapB28 (Mll)
4. VapC28 (Mll)
5. MT0934 (Mll)
6. MT0933 (Mll)
7. VapB5 (UIB1)
8. VapC5 (UIB1)
9. HP-phd (D3)
10. doc (D3)
11. HP-antiZ(D3)
12. Toxina Z (D3)
13. MT0934-S (Mch)
14. MT0934-S2(Mch)
15. MT0933(Mch)
16. Lipasa(Mch)
17. MT0934(Mch)
18. MT0933 (Mim)
19. Lipasa (Mim)
20. MT0934 (Mim)
21. MT0933 (MG2)
22. Lipasa (MG2)
23. MT0934(MG2)
24. MT0933(MG8)
25. Lipasa (MG8)
26. MT0934 (MG8)
27. MT0933 (D2)
28. Lipasa (D2)
29. MT0934(D2)
30. MT0933(D3)
31. Lipasa (D3)
32. MT0934 (D3)
36. MT0934 -S(D2)
37. MT0934 -S(D3)
38. MT0934 -S(CR)
36. Toxina Z*(MG2)
37. Toxina Z*(MG8)
38. Toxina Z*(D2)
39. Toxina Z*(D3)

Figura 6.5. Amplicones obtenidos mediante PCR a partir de ADN genómico (géles superiores) y a partir del ADN plasmídico de los clones finales (geles inferiores). La estimación de los tamaños permite confirmar que, a priori, los tamaños obtenidos son los esperados. Las toxinas MT0934S y las toxinas zeta* corresponden a las potenciales toxinas sin antitoxina asociada.

Tabla 6.6. Análisis del efecto de los STA sobre el crecimiento bacteriano. Se indican las cepas de origen de cada sistema, así como la observación de efecto toxico por parte de la toxina de cada sistema.

Cepa	Potencial TAs	Toxicidad observada
***M. llatzerense* MG13^T**	VapBC28	Si
	VapBC27	Si
	MT0933-34	No
***M. chelonae* CCUG 47445^T**	MT0933-Lipasa-Mt0934	No
	Toxina MT0933 (sin antitoxina)	No
	Toxina MT0933(sin antitoxina)	No
***M. immunogenum* CCUG 47286^T**	MT0933-Lipasa-MT0934	No
MG2	MT0933-Lipasa-MT0934	No
	Toxina zeta (sin antitoxina)	No
MG8	MT0933-Lipasa-MT0934	No
	Toxina zeta (sin antitoxina)	No
MHSD2	MT0933-Lipasa-MT0934	No
	MT0934S	No
	Toxina zeta (sin antitoxina)	No
MHSD3	MT0933-Lipasa-MT0934	No
	MT0934S	No
	Toxina zeta (sin antitoxina)	No
	HP-Toxina zeta	Si
	Doc-phd	No
CR-UIB1	MT0933-Lipasa-MT0934	No
	MT0934S	No
	Toxina zeta (sin antitoxina)	No
	VapBC5	No

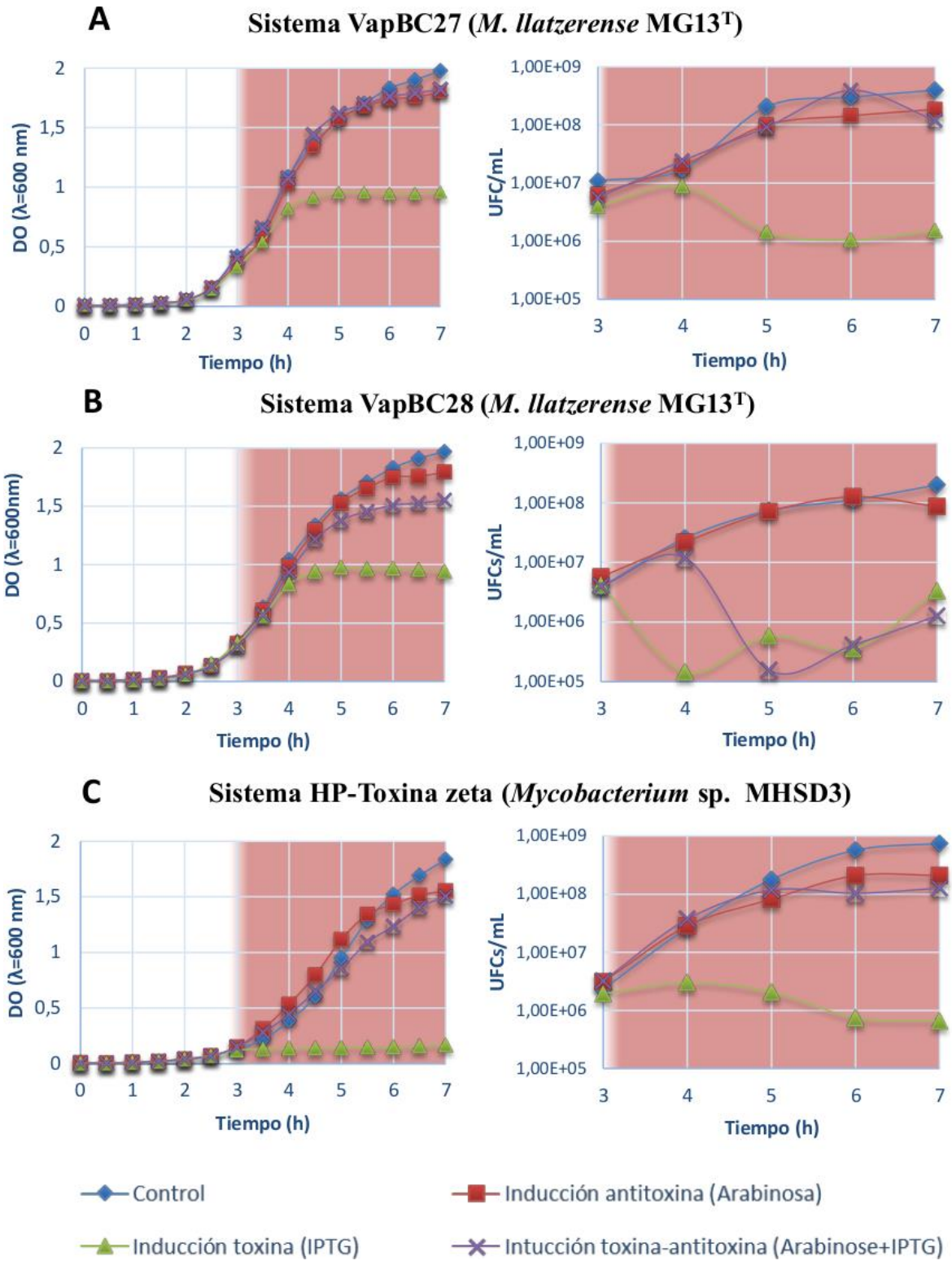


Figura 6.6. Curvas de crecimiento obtenidas a partir de las DO durante 7 horas y curvas de la evolución del número de UFC/ml a lo largo del tiempo a partir del momento de inducción (3 horas) de las distintas condiciones experimentales de los tres sistemas funcionales: A) VapBC27, B) VapBC28 y C) HP-Toxina zeta.

6.3.4. Análisis de la expresión proteica

6.3.4.1. Identificación de bandas

Los resultados obtenidos para el peso molecular teórico de las proteínas de interés (Tabla 6.7) permitieron delimitar la zona de los geles analíticos de poliacrilamida donde deberían aparecer las bandas diferenciales; concretamente entre las bandas correspondientes a los 5 y 25 kDa del marcador de peso molecular utilizado.

Tabla 6.7. Estimación de los pesos moleculares basada en secuencia de aminoácidos de las distintas proteínas que conforman los tres STA funcionales.

Cepa	Proteína	PM (kDa)
***M. llatzerense* MG13^T**	VapB27	8,34
	VapC27	14,93
	VapB28	8,9
	VapC28	14,32
MHSD3	HP	13,12
	Toxina zeta	21,64

En el caso de los ensayos de expresión de los STA VapBC27 y VapBC28 (Figura suplementaria 1, Anexo 3) se obtuvo un patrón de bandas prácticamente idéntico, mostrando hasta incluso las mismas bandas diferenciales. Por su parte, en aquellas situaciones experimentales en las que se indujo la expresión individual de un solo elemento del STA no se observaron bandas diferenciales en el rango de tamaños esperado. Sin embargo, en los experimentos donde se indujo la expresión de ambos componentes del operón VapBC28 apareció una banda especialmente intensa con un peso comprendido entre 15 y 10 kDa y que podría representar la toxina, hecho que no se observó con claridad en el caso del sistema VapBC27.

En algunos aspectos la situación es similar para el caso del STA HP-Toxina zeta (Figura suplementaria 1, Anexo 3). Así, no se observaron bandas diferenciales en aquellos carriles del gel en los que se analizaron situaciones experimentales donde se expresó solamente uno de los elementos que conforman este sistema. Sin embargo, sí se observaron diferencias en el carril donde se analizó la expresión de ambos elementos. Concretamente se observó una banda sobreexpresada comprendida entre los 20 y 25 kDa, además de una

segunda banda entre los 10 y 15 kDa, y que en base a los tamaños estimados se corresponderían con la toxina y la antitoxina respectivamente (Figura suplementaria 1, Anexo 3).

6.3.4.2. MALFI-TOFF MS

Sistema VapBC27 y VapBC28

Los resultados obtenidos mostraron la presencia de la toxina VapC28 al identificarse por MALDI-TOF MS 9 péptidos (Tabla 6.8), cubriendo un 50,4 % de la secuencia proteica (Figura 6.7). Además, estos fueron confirmados por espectrometría de masas. Para las proteínas VapB28, VapC27 y VapB27 no se consiguió identificar ningún péptido que validara su presencia.

Tabla 6.8. Secuencias peptídicas identificadas como pertenecientes a la toxina VapC28, relacionadas con las respectivas masas observadas experimentalmente, esperadas y teóricas. Se indica el % de error observado entre la masa teórica y experimental. "M" representa el número de lisinas o argininas no digeridas por la tripsina.

Inic.-Fin	Masa Observada	Masa Esperada	Masa Teórica	%	M	Péptido
81-84	537.3571	536.3498	536.2707	0.0148	0	R.QAYR.D
75-80	671.5249	670.5176	670.4126	0.0157	0	R.QALLAR.Q
24-31	754.5612	753.5540	753.4497	0.0138	0	R.AIVAGAPR.R
112-117	785.5736	784.5663	784.4483	0.0150	0	R.EPLLWK.G
111-117	941.6798	940.6726	940.5494	0.0131	1	R.REPLLWK.G
81-88	984.6490	983.6417	983.4825	0.0162	1	R.QAYRDFGK.G
118-127	1060.6020	1059.5947	1059.4734	0.0115	0	K.GDDFGHTGVR.S
14-23	1139.6493	1138.6420	1138.4891	0.0134	0	R.DENDAAVYSR.A
57-74	1925.1660	1924.1587	1924.0517	0.0056	0	K.LDELLGTAGVIIEPVTER.Q

Figura 6.7. Secuencia proteica de la toxina VapC28 cubierta con los péptidos identificados. La secuencia aminoacídica de los péptidos identificada corresponde a 71 de 133 aminoácidos, resaltados en rojo.

Sistema HP-Toxina zeta

El análisis por MALDI-TOF del patrón de picos obtenidos a partir de la muestra dio como resultado la identificación positiva de las proteínas en la base de datos propia creada a partir de las secuencias proteicas de los STA funcionales, mientras que no se obtuvo ningún resultado cuando se realizó la búsqueda con el programa MASCOT en Swiss-Prot.

En lo que respecta a la toxina se identificaron 13 péptidos (Tabla 6.9) que cubrieron hasta un 80 % del total de la secuencia de la proteína (Figura 6.8), mientras que en el caso de la antitoxina se identificaron nueve (Tabla 6.10) péptidos que en este caso abarcaron el 65 % (Figura 6.9). Todos los picos fueron confirmados posteriormente mediante espectrometría de masas.

Tabla 6.9. Secuencias peptídicas identificadas como pertenecientes a la toxina zeta relacionadas con las respectivas masas observadas experimentalmente, esperadas y teóricas. Se indica el % de error observado entre la masa teórica y experimental. "M" representa el número de lisinas o argininas no digeridas por la tripsina.

Inic-Fin	Masa observada	Masa esperada	Masa teórica	%	M	Péptido
85-89	629.4743	628.4670	628.3908	0,0121	0	K.LDLIR.S
52-57	680.4117	679.4044	679.3289	0,0111	0	R.AYEAAR.I
132-136	694.4388	693.4315	693.3558	0,0109	1	R.ARYER.L
58-63	717.4772	716.4699	716.3817	0,0123	0	R.IAEQTR.Q
90-99	1081.6030	1080.5958	1080.5200	0,007	0	R.SAQAADYTVR.L
43-51	1082.5730	1081.5657	1081.4829	0,0077	0	R.WPDEDPAPR.A
118-129	1222.7630	1221.7557	1221.6353	0,0099	0	R.VEAGGHSVPIEK.I
41-51	1366.7608	1365.7535	1365.6425	0,0081	1	K.QRWPDEDPAPR.A
100-113	1622.1286	1621.1213	1620.9814	0,0086	0	R.LLVLLVPEELTVQR.V
22-40	2113.3526	2112.3453	2112.1255	0,0104	0	K.FLAPLLHESVFVNADEIAK.Q
64-84	2313.4832	2312.4759	2312.2277	0,0107	0	R.QALISQGRPFIAETVFSHPSK.L
171-190	2321.4734	2320.4661	2320.2117	0,011	0	R.GQVVGTLTWPQWTPAPLWQR.W
137-163	2854.7691	2853.7618	2853.4549	0,0108	0	R.LWPLVVDAIALADSSVVFDNSSEPGPR.V

Tabla 6.10. Secuencias peptídicas identificadas como pertenecientes a la HP, relacionadas con las respectivas masas observadas experimentalmente, esperadas y teóricas. Se indica el % de error observado entre la masa teórica y experimental. "M" representa el número de lisinas o argininas no digeridas por la tripsina.

Inic-Fin	Masa observada	Masa esperada	Masa teórica	%	M	Péptido
2-10	982.6832	981.6759	981.5356	0,0143	0	M.AAPVDRPTR.V
32-39	1053.6712	1052.6639	1052.5152	0,0141	0	K.QQLDHWAR.L
55-66	1261.8439	1260.8366	1260.6496	0,0148	0	R.VEAALSGQLSMR.E
88-98	1266.8578	1265.8505	1265.6728	0,0140	0	R.IAATHLQDELR.A
54-66	1417.9621	1416.9548	1416.7507	0,0144	1	R.RVEAALSGQLSMR.E
87-98	1422.9743	1421.9670	1421.7739	0,0136	1	R.RIAATHLQDELR.A
11-25	1445.9502	1444.9430	1444.7158	0,0157	0	R.VASDLLDSAAAEGAR.Q
53-66	1574.0906	1573.0833	1572.8518	0,0147	2	R.RRVEAALSGQLSMR.E
67-86	2302.4850	2301.4777	2301.1376	0,0148	0	R.ELTPEEGVVFNAEIEVELER.R

Figura 6.8. Secuencia cubierta con los péptidos identificados. La secuencia aminoacidica de los péptidos identificados en el total de las proteínas analizadas (156 aminoácidos para la toxina y 78 aminoácidos para la proteína hipotética) se indican en rojo.

6.3.5. Caracterización estructural de los sistemas toxina-antitoxina

6.3.5.1. Sistemas tipo Vap

VapBC27

Atendiendo a los resultados de búsqueda que se obtuvieron mediante BLAST, tanto en el NCBI como en UniProt, de la secuencia proteica correspondiente a la antitoxina VapB27, se observaron coberturas superiores al 90 % en base al alineamiento con otras proteínas homólogas, con las que se alcanzaron identidades superiores al 80 % en otras especies de MCR e inferiores en especies de MCL, aunque siempre se mantuvieron por encima del 50 %. En *M. mucogenicum* se observó la presencia de dos proteínas en la misma disposición que fueron muy similares a las proteínas VAPB27 y VAPC27 de *M. llatezerense* MG13^{T},

con una cobertura del 99 % y una identidad del 87 % en el caso de la toxina y del 100 % tanto de cobertura como de identidad para la antitoxina. Manteniendo siempre el criterio adoptado de búsqueda con BLAST de un mínimo de 50 % de identidad en por lo menos el 50 % de la secuencia, se detectaron proteínas similares también en otros géneros como los de *Tetrasphaera*, *Microlunatus* o *Gordonia*.

En el momento de representar gráficamente las 50 proteínas más semejantes a la antitoxina detectadas en UniProt (Figura 6.9A, números de acceso en la Tabla suplementaria 3 del Anexo 3), principalmente reguladores transcripcionales (TR, del inglés *Transcriptional Regulator*) y proteínas no caracterizadas (UP, del inglés *Uncharacterized Protein*), se obtuvieron dos grandes ramas muy estables, cuyos valores de soporte de agrupación superaron el 93 %. Como se puede observar en la figura 6.9A, la rama superior se bifurca a su vez en dos grandes agrupaciones. Por una parte, un grupo constituido por proteínas muy similares pertenecientes a las especies patógenas de *M. tuberculosis*, *M. caprae*, *M. africanum*, *M. canetii* y *M. orygis*, con las que también se agrupó un TR putativo de *Blastococcus sasobxidens* DD2. Además, el conjunto de proteínas pertenecientes a distintas cepas del género *Frankia* spp., un género también incluido en el orden *Actinomicetales*, se agruparon de forma estable. La segunda gran rama del dendograma, agrupó proteínas de diferentes especies. A pesar de que los valores del análisis jerárquico de grupos mediante *bootstrap* que se obtuvieron para esta rama fueron altos, no se establecieron agrupaciones suficientemente robustas para la mayoría de las proteínas incluidas. Sin embargo, la antitoxina VapB27 de *M. llatzerense* MG13^{T}y el TR de *M. mucogenicum* se agruparon en el 95 % de los casos.

Aplicando el mismo procedimiento para la toxina VapC27 (Figura 6.9B, números de acceso en la Tabla suplementaria 4 del Anexo 3), el análisis jerárquico de grupos volvió a revelar una gran rama constituida por ribonucleasas del genero *Frankia* agrupadas con ribonucleasas de las especies patógenas comentadas anteriormente, con valores de soporte de agrupación elevados. Sin embargo, y al margen de esta, en el resto de nodos no se consiguieron agrupaciones de forma eficiente, a excepción de las ribonucleasas VapC de MCR, tal como se puede observar en la parte inferior del dendrograma de la figura 6.9B, en donde nuevamente la toxina VapC27 se agrupó con la proteína homóloga presente en *M. mucogenicum* con valores del soporte de agrupación del 85 % mediante 100 veces de iteración del proceso.

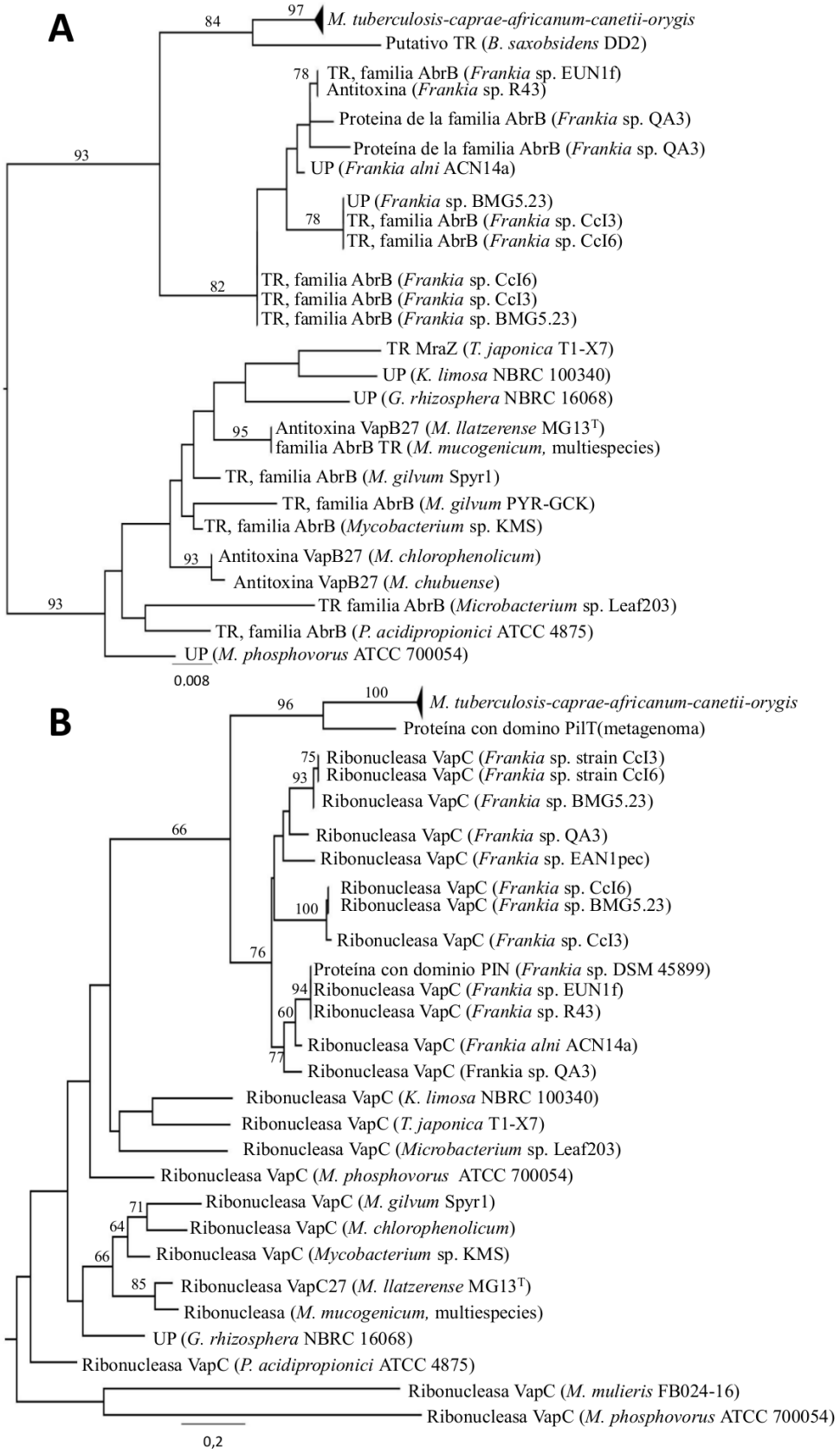


Figura 6.9. Dendrogramas obtenidos por MLE (100 iteraciones) con las proteínas que mediante BLAST mostraron una similitud de secuencia aminoacidica superior al 50 % (con un 50 % de cobertura) a VapB27 (A) y VapC27 (B). TR (del inglés *Trascriptional Regulator*), UP (del inglés *Uncharacterized Protein*).

La predicción estructural automatizada basada en la secuencia de aminoácidos de ambos elementos y que se llevó a cabo con la plataforma integrada I-TASSER (Figura 6.10) contempla una estructura de la toxina formada por siete hélices α y cuatro láminas β. El análogo estructural definido experimentalmente más próximo fue la VapC27 (TM-score 0,81) y que corresponde a la toxina del sistema VapBC de *Shigella flexneri*, con número de acceso 3TND en el PDB [240].

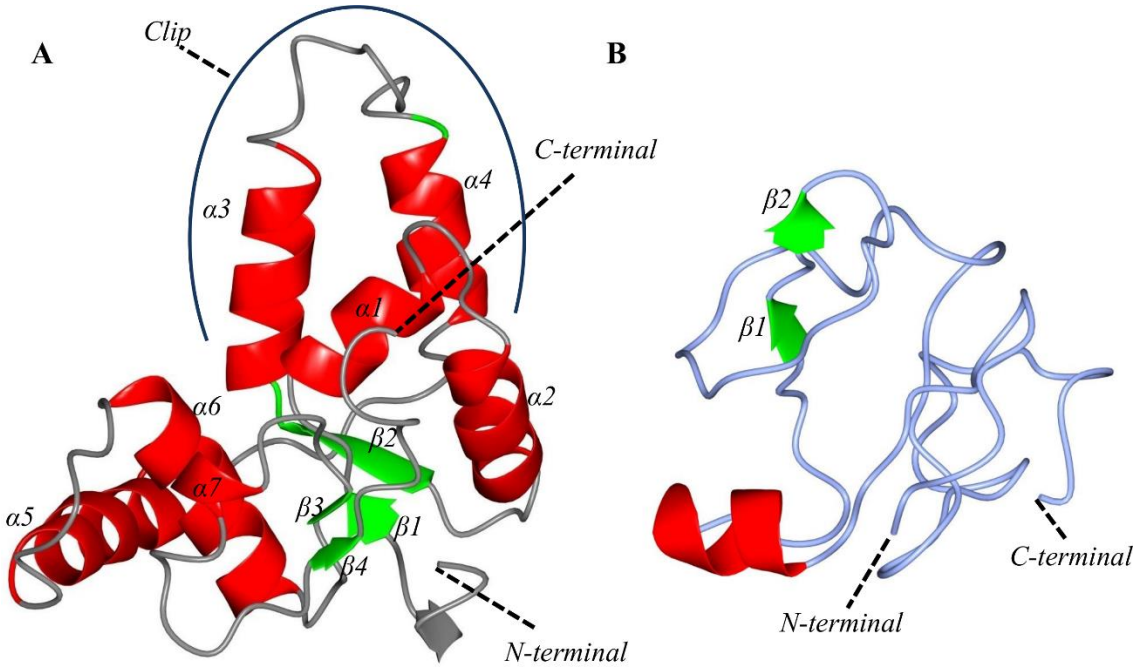


Figura 6.10. Predicción estructural de la potencial toxina VapC27 (A) y de la potencial antitoxina VapB27 (B) obtenida a través de la plataforma bioinformática I-TASSER. Se destacan en rojo las hélices α y en verde las láminas β. Dichas estructuras aparecen numeradas en orden de aparición desde el extremo amino-terminal al extremo carboxilo-terminal dentro de cada tipo.

Tal como se pudo observar al superponerlas (Figura 6.11), las estructuras del modelo calculado y el análogo estructural definido experimentalmente son prácticamente idénticos, con la salvedad de pequeñas diferencias existentes. En ambos casos se pudo observar la representación típica de la estructura correspondiente al dominio PIN. Por su parte, la predicción estructural de la antitoxina VapB27 puso de manifiesto la existencia de una hélice α y dos láminas β. El análogo estructural más próximo a la antitoxina fue la proteasa Lon de *Meiothermus taiwanensis* (con número de acceso 4FW9 en PDB), aunque con un TM-score de 0,46, reflejando poca concordancia estructural (Figura 6.12). Otros ejemplos a destacar incluirían una dipeptidasa de *Pyrococcus horikoshii* OT3 (con número de acceso 1WN1 en PDB) y otra dipeptidasa PepE de *M. ulcerans* (con número de acceso 4EGE en PDB).

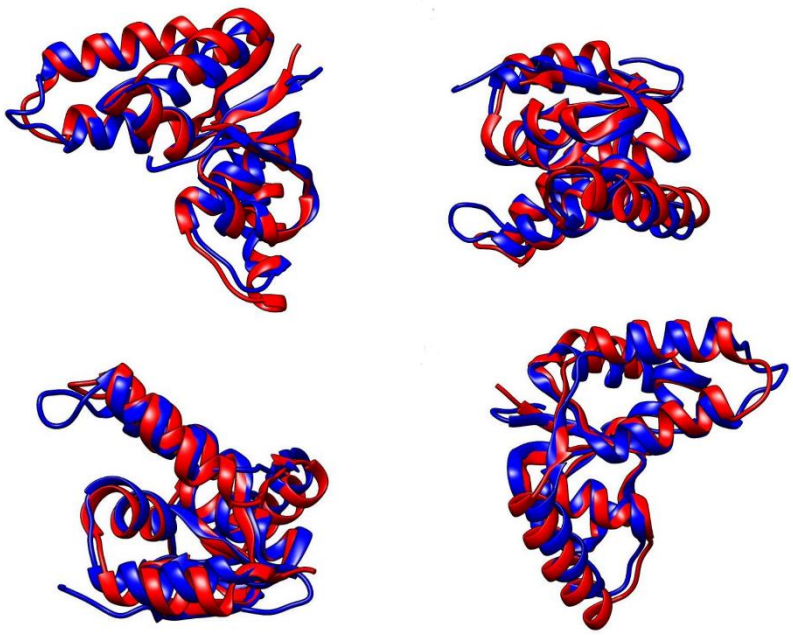


Figura 6.11. Representación de la superposición estructural entre la toxina VapC de *S. flexneri* (rojo) y *M. llatzerense* MG13^{T} (azul) donde se observa la coincidencia entre los distintos elementos estructurales de ambas proteínas.

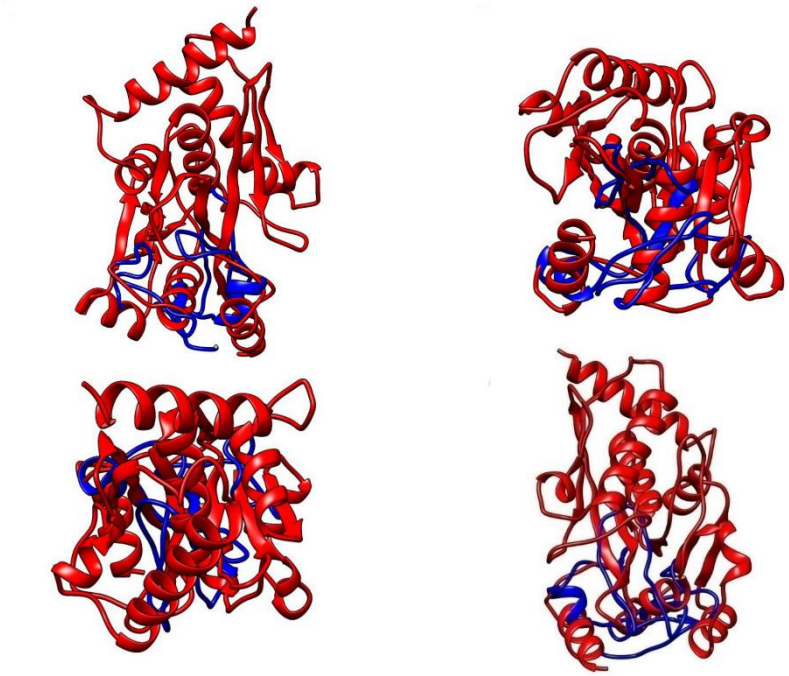


Figura 6.12. Comparativa estructural entre la toxina VapB27 (azul) de *M. llatzerense* MG13^{T} y la proteasa Lon de *Meiothermus taiwanensis* (rojo). En el modelo se pueden observar las notables diferencias estructurales entre ambas proteínas.

VapBC28

Los resultados de BLAST en el caso de la toxina, tanto en el NCBI como en UniProt, mostraron similitudes con ribonucleasas tipo VapC presentes en otras especies del género *Mycobacterium*, así como con ribonucleasas o proteínas no caracterizadas (UP) presentes en otros géneros como *Frankia, Athrobacter, Pseunonocaria* y *Gordonia*, todas ellas con coberturas por encima del 70-80 % e identidades superiores al 60 %. La misma situación se encontró cuando se hizo un análisis equivalente para la antitoxina VapB28. No obstante, en este caso a pesar de que los primeros resultados correspondían a proteínas con coberturas por encima del 80 % e identidades de más del 70 %, los resultados no fueron tan abundantes como en el caso anterior, apareciendo rápidamente en el listado proteínas que, a pesar de presentar buenas coberturas, las identidades no superaron el 50 %.

Las proteínas de las especies patógenas *M. tuberculosis, M. caprae, M. africanum, M. canetii, M. orygis* y *M. bovis* fueron muy similares y se agruparon perfectamente entre ellas, tanto en el caso de la toxina como de la antitoxina. Por su parte, la antitoxina VapB28 no pudo agruparse de forma estable con las antitoxinas de otras micobacterias ni con otros grupos de proteínas (Figura 6.13A, números de acceso en la Tabla suplementaria 5 del Anexo 3)

En el caso del dendrograma obtenido para la toxina VapC28, todas las ribonucleasas VapC de micobacterias se agruparon de forma homogénea entre ellas y con buenos valores de soporte de agrupación. Dentro de este grupo, la toxina VapC28 de *M. llatzerense* MG13^{T} formó claramente una rama independiente del resto. Una vez más las ribonucleasas del género *Frankia* formaron una agrupación propia, pero en este caso la relación con las correspondientes a micobacterias patógenas no fue tan evidente (Figura 6.13B, números de acceso en la Tabla suplementaria 6 del Anexo 3).

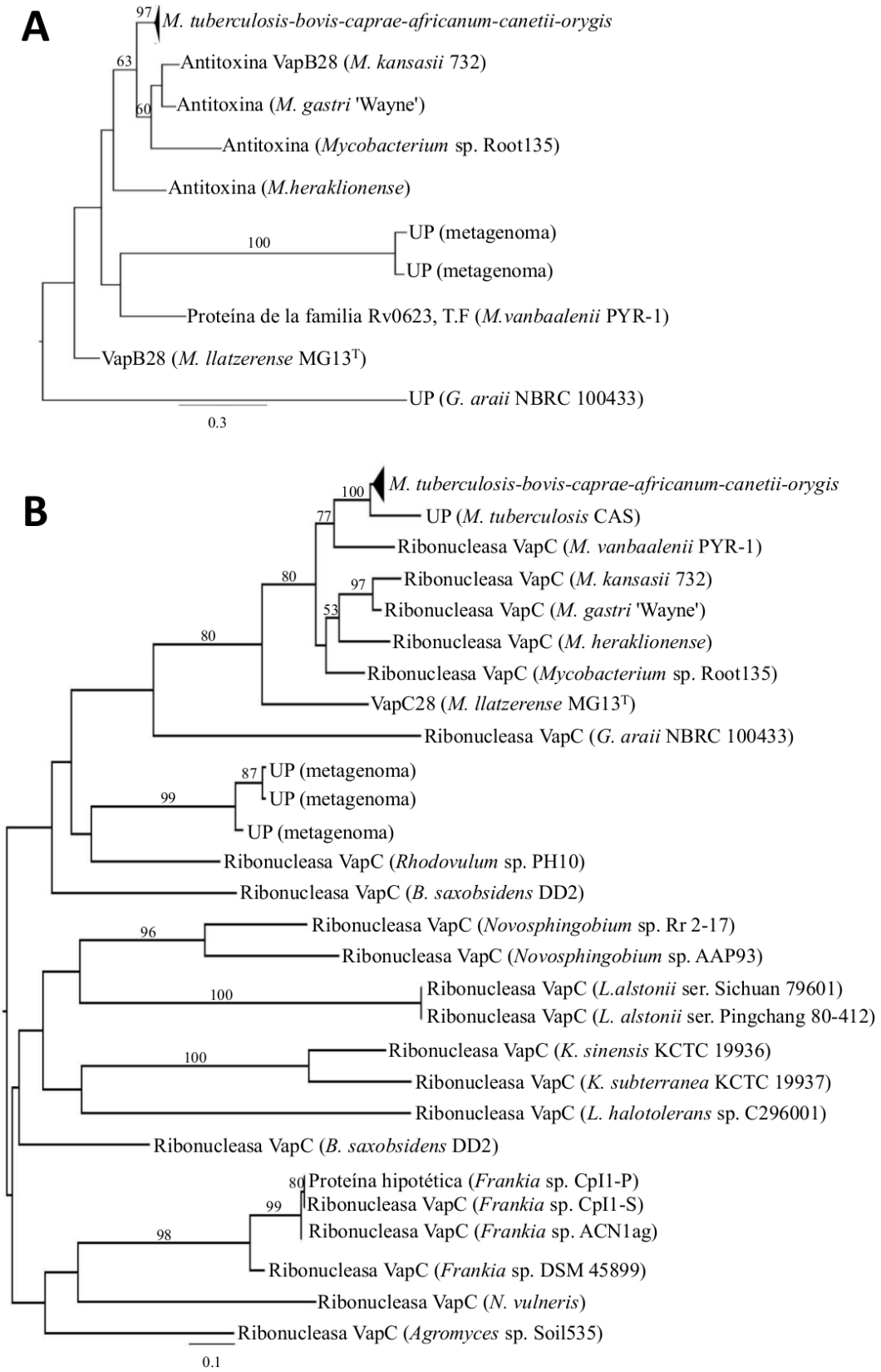


Figura 6.13. Dendogramas obtenidos mediante el algoritmo de MLE (100 iteraciones) a partir de las proteínas que mostraron más de un 50 % de similitud (y más de un 50 % de cobertura) con las proteínas VapB28 (A) y VapC28 (B). UP (Proteína no caracterizada).

Como proteína englobada dentro del grupo de proteínas con dominios PIN, VapC28 volvió a mostrar el motivo estructural que le caracteriza (Figura 6.14), en este caso constituida por seis hélices α y cuatro láminas β según la predicción realizada con I-TASSER.

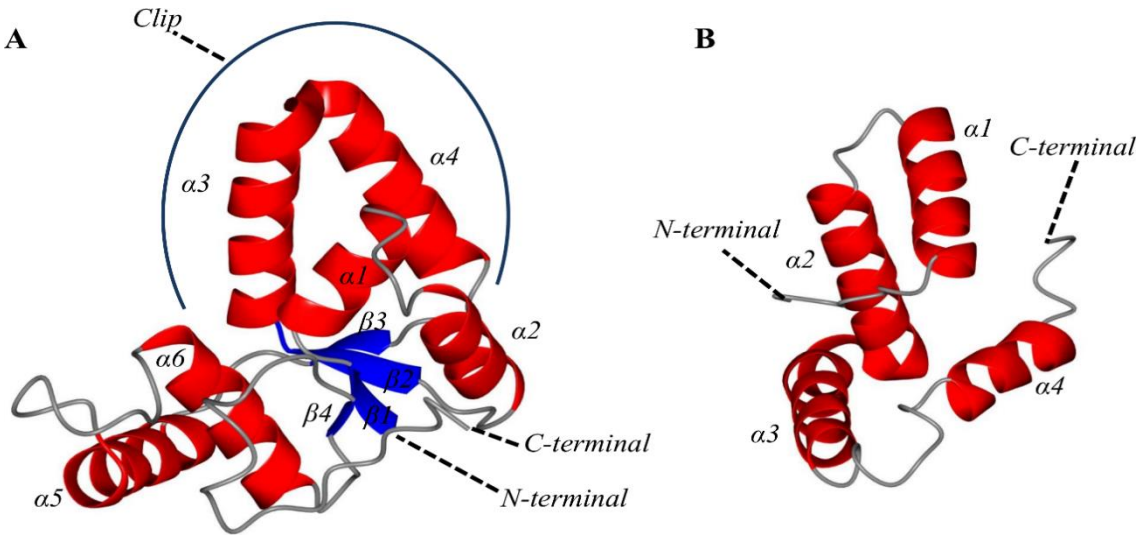


Figura 6.14. Predicción estructural de la toxina VapC28 (A) y de la antitoxina VapB28 (B) obtenida a través del programa I-TASSER. Se indica en rojo las hélices α y en azul las láminas β. Dichos elementos se enumeran por orden de aparición desde el extremo amino-terminal al extremo carboxilo-terminal dentro de cada tipo de estructura.

Su análogo estructural más próximo fue la toxina del STA VapBC30 de *M. tuberculosis* (con número de acceso 4XGQ en PDB), con un C-Score de 0,97. La estructura tridimensional para la toxina VapC30 fue previamente determinada experimentalmente [241]; y mediante los modelos de superposición de la estructura pronosticada en base a dicho homólogo, y salvo pequeñas diferencias resultaron ser prácticamente iguales (Figura 6.15). Por su parte, el modelo predictivo hecho para la estructura de la antitoxina VapB28 mostró una configuración estructural con cuatro hélices α. Esta proteína fue incluida por Pfam en la superfamilia "*PKS Transcriptional factor*", caracterizada por una estructura que contiene un motivo lazo-hélice-hélice de unión al ADN.

Su homólogo estructural más próximo resultó ser una proteína represora denominada Arc (1UP9 en PDB) con un C-Score de 0,57, la cual está incluida en Pfam dentro del mismo grupo en el que se clasificó la antitoxina VapB28 (clan CL0057 en Pfam), que incluye también otras familias de antitoxinas como ParD, ParG o CcdA. En la superposición de

los modelos de estructuras, con importantes diferencias generales, sí parecen coincidir en el motivo lazo-hélice-hélice de unión al ADN (Figura 6.16).

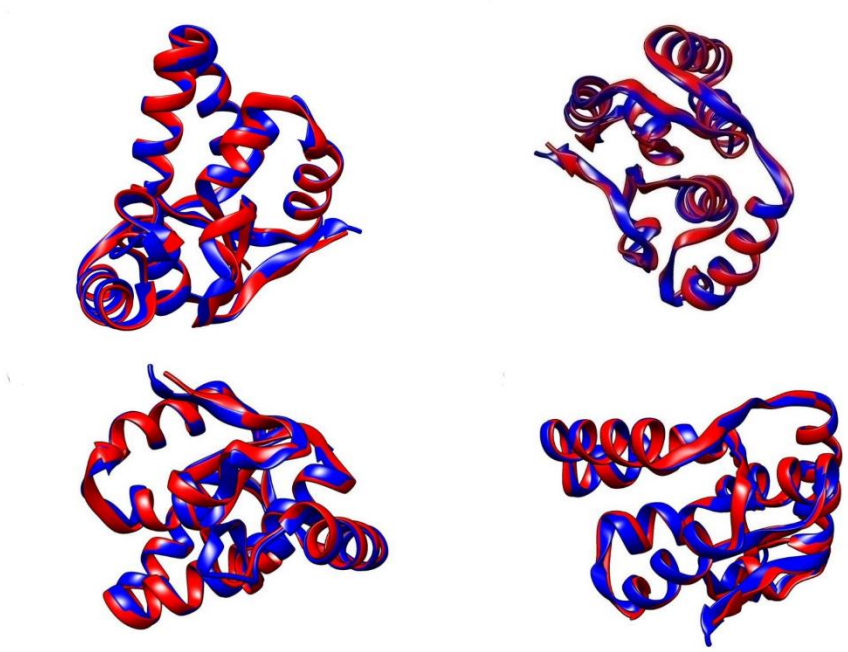


Figura 6.15. Comparativa estructural entre la toxina VapC30 (rojo) de *M. tuberculosis* y VapC28 (azul) de *M. llatzerense* MG13^T desde diferentes perspectivas.

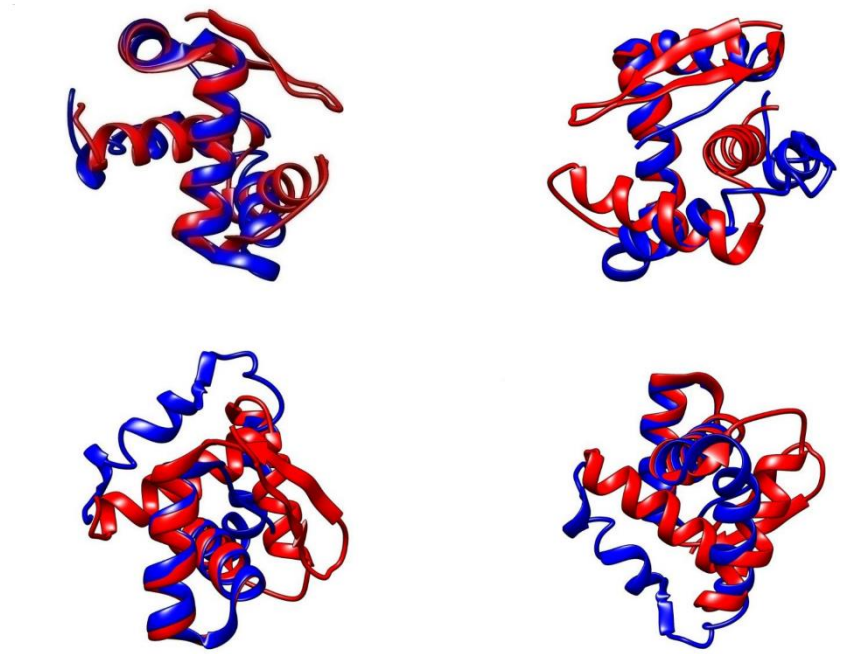


Figura 6.16. Comparativa estructural entre el represor Arc (rojo) y VapB28 (azul) de *M. llatzerense* MG13^T desde diferentes perspectivas.

Caracterización del centro activo de las ribonucleasas Vap

Teniendo en cuenta el diagrama oculto de Markov o HMM (Figura 6.17) para las proteínas con dominio PIN y atendiendo a los valores del C-Score proporcionados en el análisis con I-TASSER, por debajo de 0,3 en una escala de 0 a 1, su detección en las toxinas VapC27 y VapC28 no fue posible.

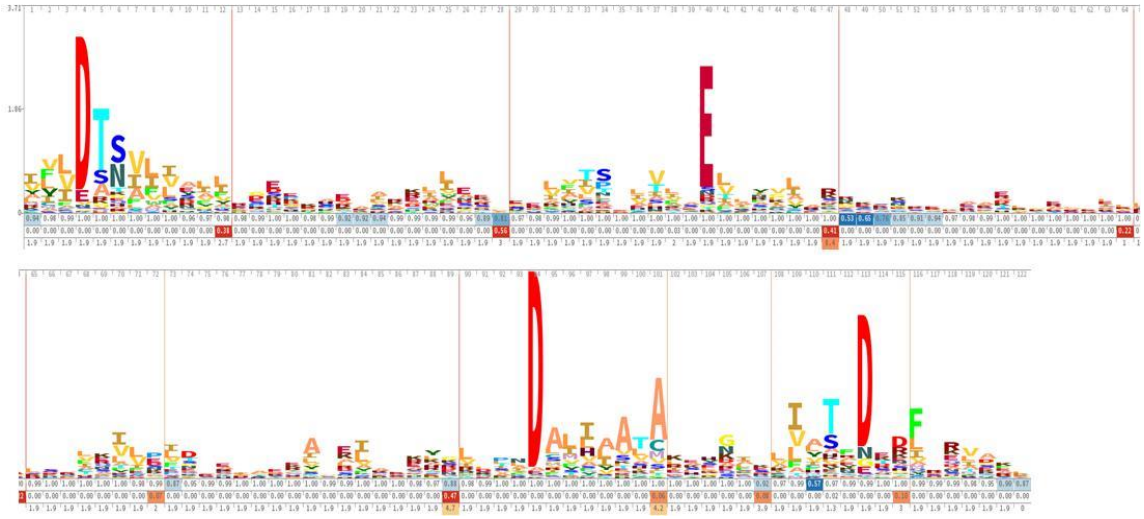


Figura 6.17. Diagrama de Markov de la familia de proteínas con dominio PIN (base de datos Pfam). En dicho diagrama se destacan principalmente el grado de conservación de los 4 aminoácidos que conforman su centro activo (mayor tamaño de letra, mayor grado de conservación).

Mediante el alineamiento de estructuras se identificó a los cuatro residuos implicados en el centro activo de las toxinas Vap de los genomas estudiados, según el HMM de las proteínas con dominio PIN. Se utilizaron como guía tres estructuras de proteínas de este tipo y aplicando un procedimiento descrito anteriormente en la bibliografía [241]. El resultado de este alineamiento reveló cómo la toxina VapC27 presentaba las cuatro posiciones de aminoácidos altamente conservados en secuencia, mientras que las toxinas VapC5 y VapC28 presentaban conservados tres de los aminoácidos, mientras que en el cuarto la glutamina era sustituida por glicina en el caso de la toxina VapC28 y por asparagina en el caso de la VapC5 (Figura suplementaria 2 y Tabla suplementaria 9, Anexo 3). La localización de estas posiciones en la estructura terciaria de la proteína permitió la reconstrucción del centro activo (Figura 6.18).

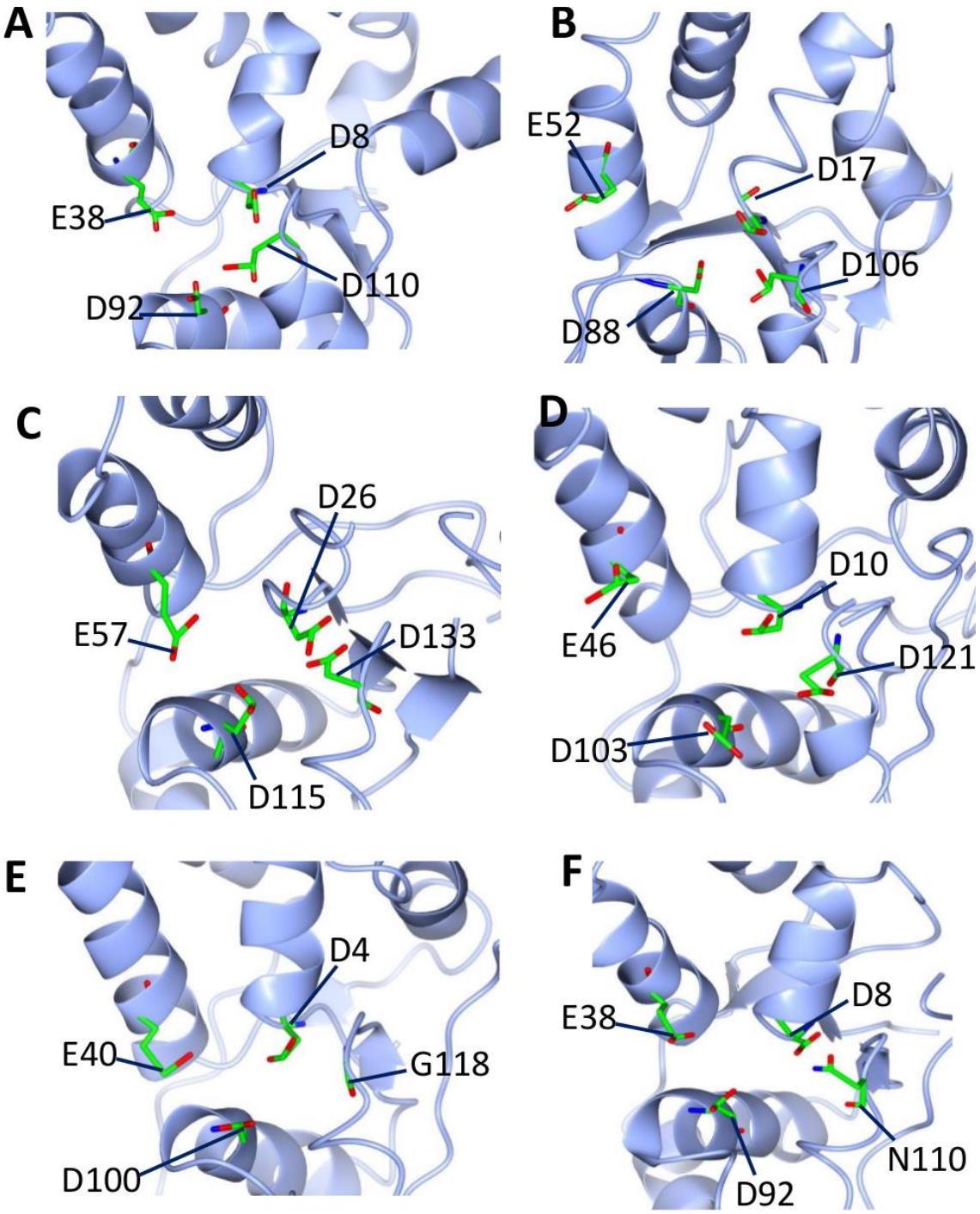


Figura 6.18. Representación de las posiciones de los residuos potencialmente implicados en la actividad de las toxinas donde D: Aspartato, E: glutamina, G: glicina y N: asparagina. A) Proteína con dominio PIN de *Pyrobaculum aerophilum* ATCC 51768, B) *Archaeoglobus fulgidus* DSM 4304, C) VapC5 de *M. tuberculosis* H37Rv, D) VapC27, E) VapC28 de *M. llatzerense* MG13^{T} y F) VapC5 de la cepa CR-UIB1.

6.3.5.2. Sistema proteína hipotética-toxina zeta

En base a los resultados obtenidos mediante BLAST, tanto en NCBI como UniProt, este sistema únicamente mostró similaridad con un pequeño grupo de proteínas. En el caso de la HP los porcentajes de identidad fueron superiores al 92 % e incluía un conjunto de proteínas idénticas asociadas al mismo número de acceso (WP_064408866.1), pertenecientes a las cepas *Mycobacterium* sp. QIA-37 (número de acceso CP010071.1), *M. chelonae* 1558 (número de acceso JAOI01, aislada de una muestra de esputo) y *M. chelonae* 15517 (número de acceso MLIR01, aislada de una muestra de tejido blando). En el caso de la toxina zeta las identidades se mantenían entre un 88 y un 99 %, en un 100 % de cobertura, con una ATPasa multiespecie (WP_070917639.1) representada nuevamente en las cepas 1558 y 15517 (indicadas en el caso anterior) y *M. chelonae* 15518 (número de acceso MLIS01, aislada de tejido blando); una ATPasa de *Mycobacterium* sp. QIA-37 (WP_064408867.1); y una HP (WP_064408867) de la cepa *M. chelonae* 203 (número de acceso MLID01). Los resultados de BLAST con el resto de proteínas utilizadas en la presente comparativa se mantuvieron por encima del 90 % de cobertura, pero con identidades inferiores al 65 %, entre las cuales se encontraron representantes de los grupos MCR y MCL. A partir del listado obtenido en UniProt, ordenado por identidad, las 50 primeras proteínas se utilizaron en la construcción de los dendrogramas para confirmar su posición y relaciones respecto al resto (Figura 6.19).

La representación gráfica del análisis jerárquico obtenido para la HP, y en el que se representaron las relaciones en base a la secuencia de aminoácidos de la antitoxina (Figura 6.19A, números de acceso en la Tabla suplementaria 7 del Anexo 3); se diferenciaron dos grupos principales: 1) proteínas representativas de MCL y MCR, donde a su vez destacan dos grandes agrupaciones constituidas por proteínas altamente similares de especies del complejo *M. avium* (MAC), y proteínas iguales presentes en las especies patógenas *M. tuberculosis*, *M. caprae*, *M. africanum*, *M. orygis* y *M. canetti*; y 2) un grupo constituido por la HP de la cepa *Mycobacterium* sp. MHSD3 que, junto con las proteínas con las que mostro valores de identidad altos, formó una agrupación individual con un valor de soporte estadístico del 100 % y mostrando una notable diferencia en cuanto a secuencia de aminoácidos con respecto a las otras proteínas comentadas.

El árbol obtenido para la toxina zeta (Figura 6.19B, números de acceso en la Tabla suplementaria 8 del Anexo 3) mostró una división en dos grupos principales. El primero reflejó una clara diferenciación entre ATPasas o UP procedentes de MCL y MCR, formándose a su vez dos subgrupos. El primer subgrupo aglutinaría las proteínas procedentes de MCL, a excepción de *M. phlei* que es una MCR cuya toxina apareció bien agrupada con las proteínas procedentes de *M. parascrofularaceum* y *M. europaeum*. En este subgrupo las secuencias procedentes de las especies *M. tuberculosis*, *M. caprae*, *M. africanum* y *M. orygis* mostraron ser prácticamente idénticas (valores de identidad por encima del 98 % sobre el 100 % de la secuencia), al igual que las proteínas de los representantes del MAC. Por su parte, en el segundo subgrupo de la rama superior se concentraron las secuencias procedentes de micobacterias de crecimiento rápido, a excepción de *M. vulneris*, una MCL cuya toxina apareció agrupada con las toxinas zeta y ATPasas de MCR como *M. neworleansense* o *M. conceptionense*.

El segundo grupo principal del dendrograma, con unos valores de soporte de agrupación del 100 %, se formó con las secuencias de las proteínas procedentes de la cepa *Mycobacterium* sp. MHSD3, *Mycobacterium* sp. QIA-37, las ATPasas de las cepas *M. chelonae* 1558 y 15517 y *M. chelonae* 15518; así como una HP de *M. chelonae* 203. Esta agrupación quedó claramente fuera de cualquiera de las agrupaciones del dendograma. En el análisis jerárquico para estos grupos de proteínas, los valores de soporte de agrupación derivados mostraron en general una gran solidez en las relaciones obtenidas.

Vista la gran similitud con respecto a las proteínas de la cepa MHSD3 y las cepas 1558, 15517, QIA-37 y 203, en el contexto de la genómica comparada se realizó un estudio de sintenia entre dichas cepas con respecto a las dos proteínas estudiadas (Figura 6.20). Así, se observó cierto grado de paralelismo en la organización de los genes que codifican para dichas proteínas. En todos los casos se identificaron genes codificantes para proteínas con dominios de trasnposasa corriente arriba del gen de la HP del supuesto STA. Genes relacionados con actividad β-lactamasa fueron hallados en las cepas MHSD3, 1558, 15517 y QIA-37 codificados corrientes abajo del gen de la zeta toxina u homólogos. Aunque en el caso de la cepa *M. chelonae* 203 no se encontraron genes relacionados con dicha actividad, si se encontró un gen que codificaba para una proteína con dominio

degradador de grupos hemo, relacionado con la respuesta al estrés frente a la presencia de hemina o peróxido.

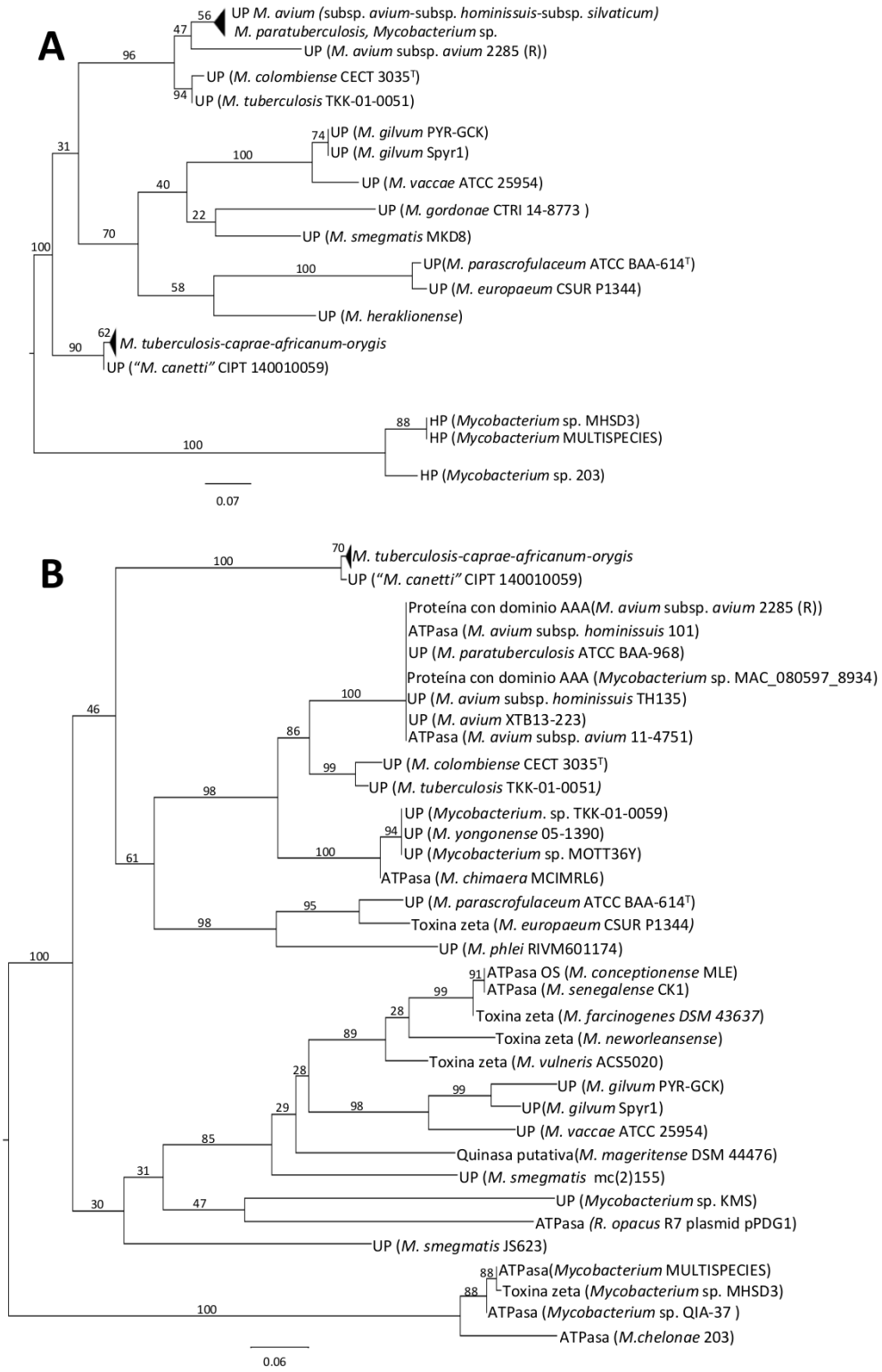


Figura 6.19. Dendogramas obtenidos mediante el algoritmo de MLE (100 iteraciones) a partir de las proteínas que mostraron más de un 50 % de similitud (y más de un 50 % de cobertura) con la proteína hipotética (A) y la toxina zeta (B). UP (Proteína no caracterizada).

No se pudo obtener mucha más información en cuanto a la sintenia de la cepa *M. chelonae* 15518, ya que la organización mostrada en la figura 24 corresponde a un *contig* de 3 kb. La sintenia de los genes de interés fue más conservada entre las cepas MHSD3, 1558 y 15517.

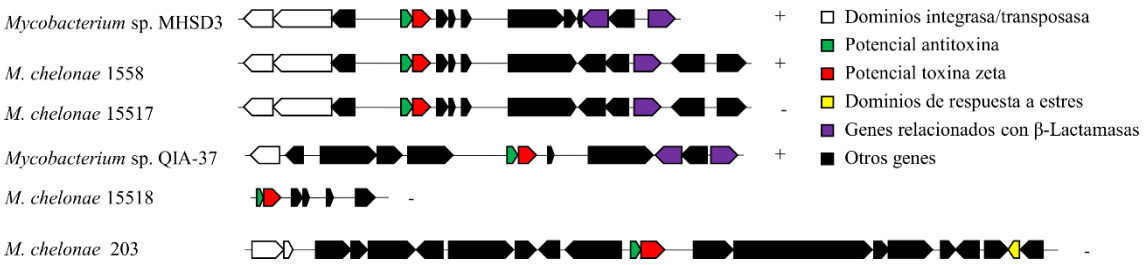


Figura 6.20. Sintenia observada en los bloques génicos que enmarcan los genes que representan el potencial operón HP.-toxina zeta. Se destaca en verde el gen de la potencial antitoxina (HP). de interés, en rojo el gen que codifica para la toxina zeta, en violeta los genes que contienen dominios relacionados con actividad β-lactamasa, en amarillo los genes con dominios de respuesta a estrés y en negro se representan los genes relacionados con otras funciones.

Con la ayuda de I-TASSER se consiguieron realizar las predicciones estructurales tanto de la toxina como de la antitoxina en el STA HP - toxina zeta (Figura 6.21). En el primer caso, la estructura se conformó por siete hélices α y cinco láminas β, mientras que en la antitoxina se encontraron cinco hélices α.

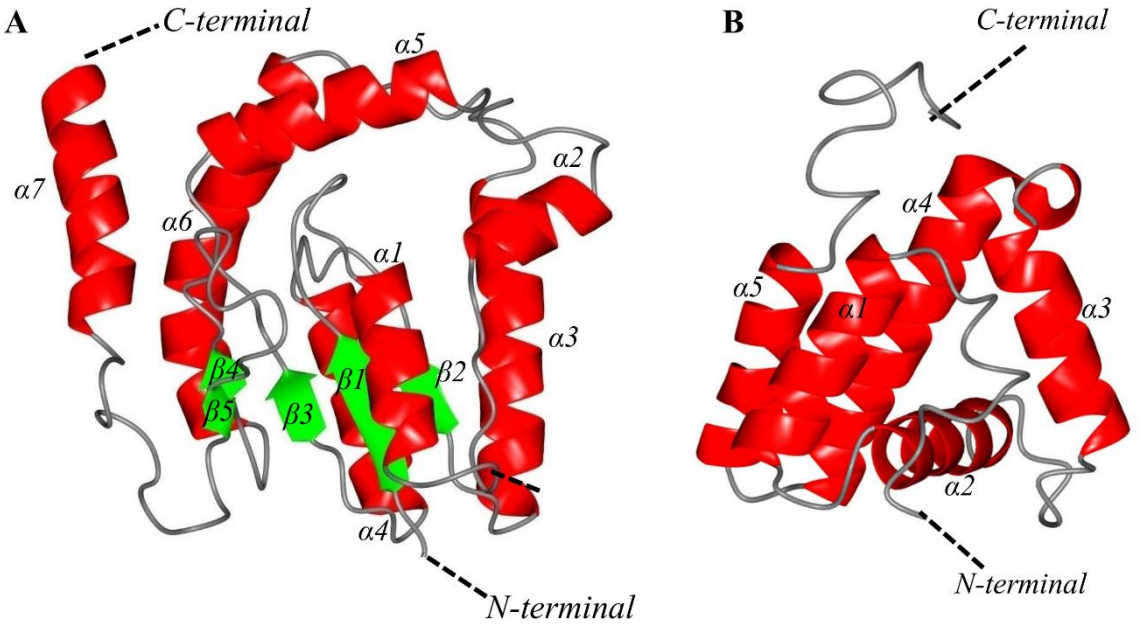


Figura 6.21. Predicción estructural de la toxina zeta (A) y de la potencial antitoxina (HP) (B) presente únicamente en el genoma de la cepa *Mycobacterium* sp. MHSD3. En rojo se destacan las hélices α y en verde las láminas β. Estos elementos se enumeran por orden de aparición desde el extremo amino-terminal al extremo carboxilo-terminal dentro de cada tipo.

A partir de las estructuras disponibles en bases de datos, el análogo estructural más próximo a la toxina de MHSD3 según I-TASSER resultó ser la toxina ζ del STA ε-ζ del plásmido pSM19035 de *Streptococcus pyogenes* [242] (cadena B del complejo bajo el número de acceso 1GVN en el PDB). Al superponer ambas estructuras, se observó la existencia de importantes similitudes en lo que respecta al núcleo central de la proteína, a la vez que presentaron notables diferencias en otras regiones (Figura 6.22).

A través de la información obtenida por I- TASSER, se consiguieron identificar los posibles residuos aminoacídicos que estarían implicados en el lugar de unión a ATP (Figura 6.23). El C-Score que se obtuvo para este potencial punto de unión al ATP identificado fue de 0,69.

Por lo que respecta a la potencial antitoxina el análogo estructural más cercano, según los cálculos de I- TASSER, fue la helicasa PriA de *Klebsiella pneumoniae* (con número de acceso 4NL4 en el PDB); una proteína de unión al ADN con un tamaño muy superior a la hipotética antitoxina (Figura 6.24) y que no tendría nada que ver con los sistemas toxina-antitoxina conocidos. En base a que la plataforma Pfam clasificó la antitoxina en la familia ParD-like, se decidió hacer la comparación con dicha antitoxina utilizando la única estructura disponible en el PDB y que correspondería a la proteína ParD de *Escherichia coli*, con número de acceso 2AN7 en el PDB (Figura 6.25). Como resultado de esta última comparación se pudo observar que estructuralmente no existen similitudes y, de hecho, no se pudieron superponer a consecuencia de estas sustanciales diferencias existentes. Además, el alineamiento de las secuencias proteicas de ParD y la antitoxina de MHSD3 mostraron a nivel de secuencia de aminoácidos ser también muy diferentes.

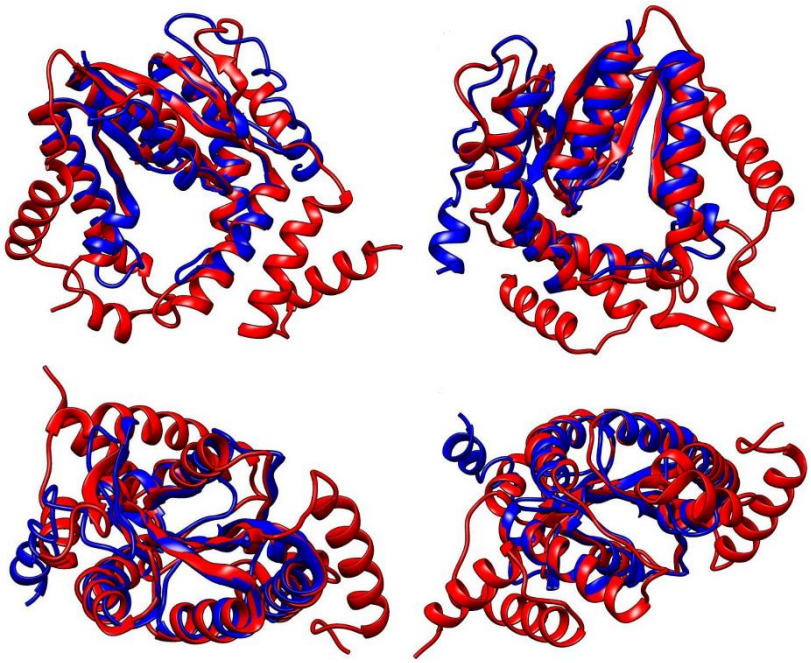


Figura 6.22. Comparativa estructural entre la toxina ζ del plásmido pSM1035 *S. pyogenes* (rojo) y la toxina zeta presente en el genoma de la cepa *Mycobacterium* sp. MHSD3 (azul). Se observa la similitud en el núcleo central de ambas proteínas.

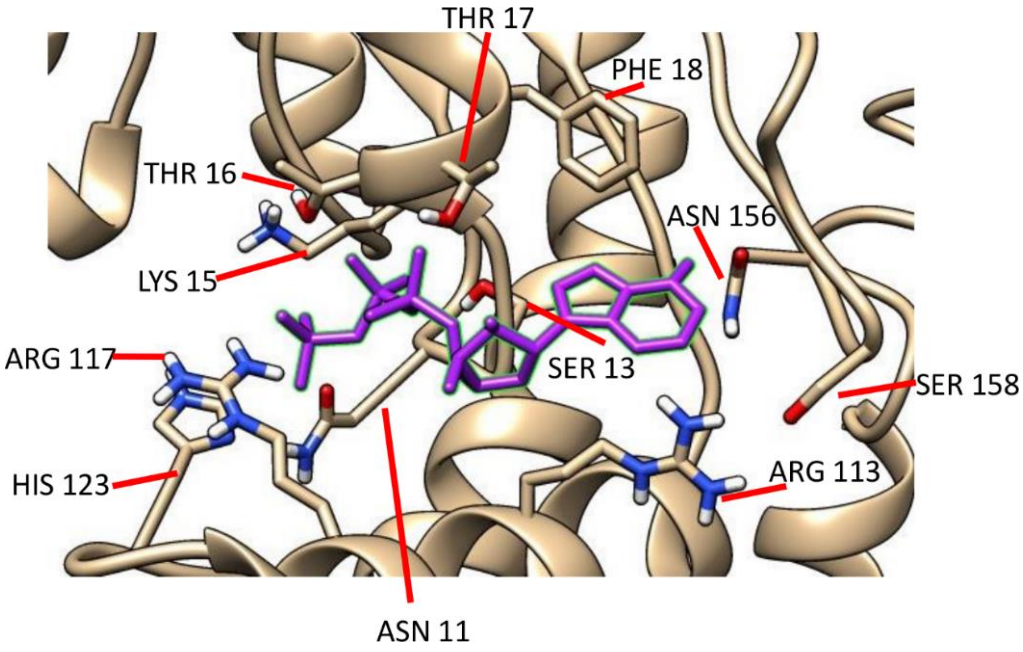


Figura 6.23. Residuos potencialmente implicados en la zona de unión a ATP (molécula destacada en color violeta). ARG: arginina, ASN: asparagina, LYS: Lisina, THR: treonina, HIS: histidina, PHE: fenilalanina.

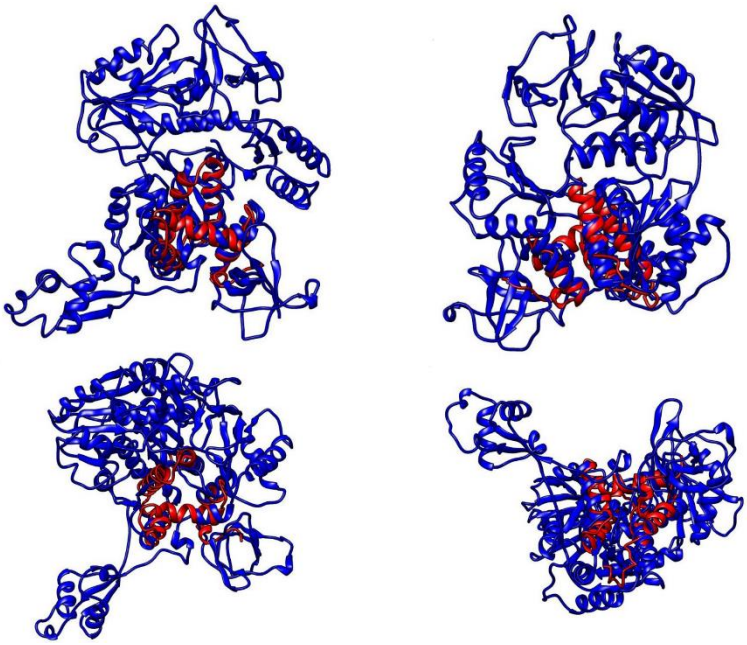


Figura 6.24. Comparativa estructural entre la helicasa PriA de *Klebsiella pneumoniae* (Azul) y la potencial antitoxina del genoma de la cepa *Mycobacterium* sp. MHSD3 (rojo) desde diferentes perspectivas.

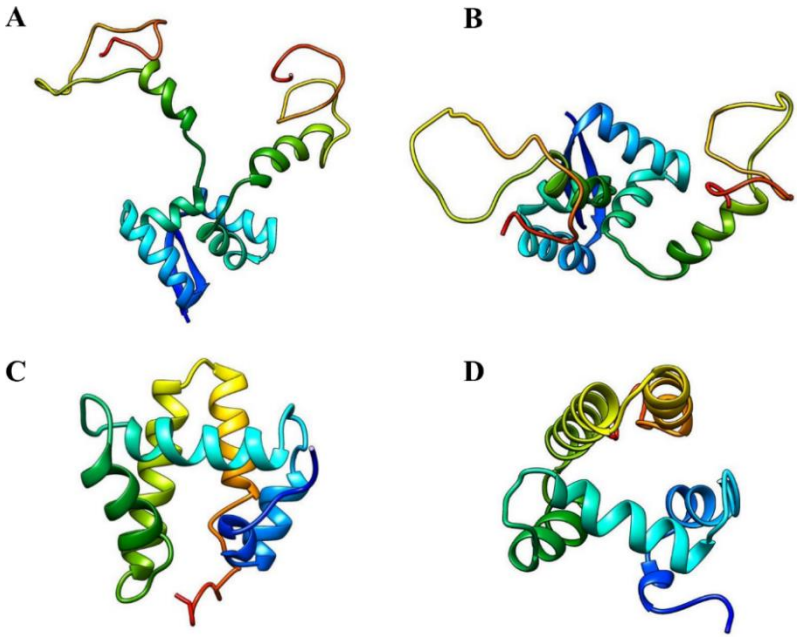


Figura 6.25. Comparativa estructural entre la antitoxina ParD (A, B) y la potencial antitoxina de MHSD3 (C, D).

6.4. Discusión

6.4.1. Sistemas toxina-antitoxina tipo Vap

La familia de proteínas VapBC es la más común y abundante dentro de las nueve familias de STA aceptadas a día de hoy [241]. Existen estudios en los que se apunta hacia una función enfocada a la ralentización o detención del crecimiento celular frente a determinados estímulos externos, permitiendo a la célula entrar en un estado de letargo que le permite sobrevivir en un ambiente hostil o de condiciones estresantes [243,244], pudiéndose relacionar también con la virulencia del microorganismo [245]. En consecuencia, en principio no se trata de sistemas letales para el huésped, presentando además un efecto reversible en el momento en que la condición adversa desaparece [246]. Esto puede ser especialmente importante en patógenos bien definidos como *M. tuberculosis*, donde la acción de estos sistemas les ayudaría a entrar en un estado de latencia (células persistentes) durante las etapas intracelulares del proceso de infección y podría explicar en cierta manera la gran cantidad de copias de *vapBC* que podemos encontrar en su genoma, donde más del 50 % de STA corresponden precisamente a los de tipo VapBC [238]. Este hecho se confirmó en el caso de *M. tuberculosis* CR-UIB2 donde se encontraron hasta 46 sistemas VapBC de los 68 STAs detectados, lo que representa un 67 %. Ese número tan elevado a buen seguro se debe a un importante papel que lleva a este patógeno a no sólo no desprenderse de ellos, sino incluso a acumularlos. En el caso de las bacterias ambientales, estos números no fueron tan elevados, pero siguen estando presentes, como es el caso de los sistemas VapBC27 y VapBC28 hallados en *M. llatzerense* $MG13^T$ y el de VapBC5 de *M. abscessus* subps. *bolletii* CR-UIB1.

Las toxinas VapC son ribonucleasas cuya estructura está caracterizada por la presencia del denominado dominio PIN. Las proteínas con dominio PIN están ampliamente distribuidas y aparecen en eucariotas, arqueas y bacterias, desempeñando generalmente en este último caso el papel de toxina de los operones TA [247]. Al parecer, ejercen su actividad ribonucleasa sobre ARN monocatenario en un proceso dependiente de Mg^{2+}, o en algunos casos de Mn^{2+} [248,249]. En la mayoría de ocasiones el gen que codifica para la toxina VapC presenta corriente arriba un gen, normalmente orientado en sentido 5'->3' solapando en su extremo el gen de la toxina, el cual codifica una proteína que suele

contener dominios de unión al ADN [250], especialmente de las superfamilias MetJ/Arc y ArbB/MazE [251]. Esto se debe a que las antitoxinas VapB inhiben la acción de la toxina uniéndose a ellas mediante una interacción proteína-proteína y, una vez formado este complejo, son capaces de unirse al ADN para auto-regular la expresión del operón TA [250].

En las toxinas VapC5, VapC27 y VapC28 se identificó a través de la base de datos Pfam la presencia del dominio PIN, al mismo tiempo que se clasificó a las antitoxinas VapB27 en la superfamilia ArbB/MazE (familia antitoxina MazE) y a VapB28 en la superfamilia MetJ/Arc (familia de factores de transcripción PKS). Por su parte, VapB5 fue incluida en la superfamilia Plasmid-antitox (familia antitoxina PhdYeFM). Toda esta información pone de manifiesto que, desde el punto de vista de los dominios funcionales, estas secuencias proteicas tendrían lo necesario para desempeñar la función que se les presupone como genes pertenecientes a operones TA tipo VapBC.

El ensayo experimental reprodujo perfectamente el comportamiento de un STA en el caso de VapBC27, mientras que en el sistema VapBC28 y aunque se demostró claramente el efecto tóxico sobre la población bacteriana, la capacidad de atenuación de la antitoxina no quedó patente de forma clara e irrefutable, debido al progreso similar en cuanto a número de UFC/ml que mostraron los respectivos cultivos al inducir la expresión de la toxina en solitario o de ambos elementos (toxina-antitoxina) a la vez. Cabe destacar que el efecto de inhibición del crecimiento, observado en el planteamiento experimental aplicado, sólo hace referencia a la contención por el secuestro que realizaría la antitoxina sobre la toxina, ya que el promotor natural del operón no está presente, por lo que el complejo no puede modular la expresión del mismo. Este hecho podría haber influido en la inhibición incompleta de la acción de la toxina VapC28. Por su parte, el STA VapBC5 no mostró efecto tóxico alguno en las condiciones experimentales aplicadas.

Los análisis por espectroscopia de masas por MALDI-TOF demostraron la expresión de la proteína VapC28, pero no se pudo obtener ningún resultado en cuanto a su antitoxina o a las proteínas del sistema VapBC27. La complejidad de la muestra y la ausencia de bandas diferenciales para estas tres proteínas en concreto, complicó el proceso de caracterización y derivó en los resultados obtenidos. Así, en el caso del sistema VapBC27 se podría concluir que se trata de un sistema cuya funcionalidad quedó demostrada

experimentalmente, pero para la cual no se pudo demostrar su presencia en los extractos de proteínas totales obtenidas. Métodos con mayor resolución o la introducción de colas de polihistidina a través de los vectores de expresión para su purificación podrían contribuir a demostrar la presencia de estas proteínas en la célula.

Desde el punto de vista estructural, las proteínas con dominio PIN, a pesar de ser una familia de proteínas con grandes diferencias en secuencia entre sus representantes, presentan una estructura altamente conservada. Así, la estructura básica está constituida por un sándwich α/β/α caracterizado por cinco láminas β paralelas en la región central del dominio. En su estructura se repiten unidades formadas por una lámina β y dos hélices α seguidas con giros hacia la derecha [247]. Esta gran conservación estructural ha facilitado en gran medida la predicción de las estructuras de las tres VapC problema analizadas, donde todas ellas presentaron la estructura terciaria característica, con las correspondientes variaciones respecto del modelo básico de esta familia. Así, las toxinas VapC27 y VapC28 exhibieron cuatro láminas β y siete hélices α, con tres repeticiones lamina-hélice-hélice en su estructura. Por su parte, VapC5 presenta 4 láminas β y 6 hélices α, con sólo dos repeticiones lamina-hélice-hélice en su secuencia según la predicción de I-TASSER. Evidentemente, las variaciones estructurales pueden tener un claro efecto sobre la acción de la proteína, especialmente si éstas afectan de alguna manera a la formación del centro activo. Las proteínas con dominio PIN se caracterizan por un centro activo muy conservado, constituido por un tetrámero de aminoácidos donde tres posiciones son altamente conservadas (dos asparaginas y una glutamina) y un cuarto aminoácido más variable [241], pero que según el HMM de la familia (Figura 6.17) el más común es una asparagina. La estructura conservada de estas proteínas hizo que, aunque sus ubicaciones en la estructura primaria de la proteína están espaciadas, dichos aminoácidos se agruparon siempre en la misma disposición, lo que en el alineamiento de las estructuras facilitó su identificación. VapC27 presentó los cuatro aminoácidos característicos, mientras que VapC5 y VapC28 presentaban tres altamente conservados y un cuarto variable, según la proteína. Además, todas las secuencias proteicas correspondientes presentaron un aminoácido polar en la posición +1 con respecto al primer aminoácido del centro activo, en este caso una serina, la cual parece ser esencial para su funcionamiento [247]. En principio estos resultados estarían en concordancia con la información bibliográfica consultada y explicarían que VapC27 y VapC28

aparentemente estuviesen en disposición de poder llevar a cabo su actividad, aunque no explicarían por qué la toxina VapC5 no mostró efecto tóxico alguno, toda vez que según los datos de secuencia y comparaciones estructurales su centro activo no parece estar alterado.

En contra de lo que ocurrió con las toxinas, las antitoxinas VapB no presentaron una estructura tan conservada, lo cual hizo difícil determinar su estructura. Además, son pocas las proteínas VapB que actualmente encontramos en las bases de datos con una estructura tridimensional corroborada experimentalmente. Así, la toxina VapB5 pudo ser comparada estructuralmente sólo con VapB5 de *M. tuberculosis*, una de las pocas disponible en las bases de datos [252]. Al compararlas, se observa que no tienen mucho en común, estando la VapB5 constituida por dos hélices α conectadas por una lazo mientras que la estructura secundaria de la VapB5 de CR-UIB1 consta de dos láminas β y tres hélices α. Las predicciones estructurales determinadas mediante I-TASSER de las antitoxinas VapB5, VapB27 y VapB28 deben ser tenidas en cuenta únicamente como lo que son, modelos teóricos y como tales se requiere de una confirmación experimental. Sin embargo, la gran variabilidad estructural entre las antitoxinas, incluso entre aquellas que pertenecen a un mismo tipo, podría relacionarse con la gran especificidad que presenta cada toxina por su respectiva toxina VapC, ya que para este tipo de toxinas se observa escasa o nula actividad cruzada [253].

Atendiendo a la distribución de los sistemas VapBC27 y VapBC28, parecen estar presentes en multitud de otras especies o cepas según los dendogramas obtenidos utilizando las secuencias de proteínas más similares encontradas en las bases de datos consultadas. En primer lugar, merece destacarse la presencia de UP correspondientes a una potencial toxina o antitoxina, tanto en el sistema VapBC27 como en el VapBC28, prácticamente idénticas y compartidas por las especies patógenas *M. tuberculosis*, *M. caprae*, *M. africanum* (sólo VapBC27), *M. orygis* y *M. bovis* (sólo VapBC28), por lo que parece existir una estrecha relación entre dichas proteínas de los integrantes del complejo *M. tuberculosis* (MTC). En lo que se refiere al STA de VapBC27, concretamente estas proteínas del MTC se incluyeron de forma estable en un gran grupo, conjuntamente con un segundo subgrupo de proteínas procedentes del género *Frankia*, actinobacterias que, de forma general, actúan como simbiontes de plantas [254]. La toxina VapC27 y

antitoxina VapB27 de *M. llatzerense* MG13^T se agruparon siempre con las respectivas proteínas procedentes de *M. mucogenicum.* Sin embargo, a pesar de que la toxina sí se agrupó de forma significativa con proteínas procedentes de otras micobacterias, la antitoxina no puedo ser asignada manifiestamente a un grupo determinado de proteínas.

En el caso del dendograma para las proteínas correspondientes al operón VapBC28, se percibieron algunas similitudes con el caso anterior. Por una parte, en las especies del MTC surgieron proteínas semejantes, tanto para la toxina como para la antitoxina. Además, la antitoxina VapB28 de *M. llatzerense* MG13^T no se agrupó ostensiblemente con otros grupos de proteínas. En cambio, la toxina VapC28 si se agrupó de forma muy clara (con un soporte de agrupación de 80 mediante la iteración del proceso de 100 veces) con las ribonucleasas procedentes de otras micobacterias utilizadas para el estudio. En este caso las secuencias proteicas procedentes de las cepas del género *Frankia* no parecieron guardar valores de agrupación tan altos como con las micobacterias patógenas como en el caso anterior.

6.4.2. Sistema toxina-antitoxina proteína hipotética-toxina zeta de *Mycobacterium* sp. MHSD3

Los STA epsilon-zeta (ε-ζ cuando se codifican en plásmidos) son sistemas de tipo II cuya toxina puede ejercer un efecto bacteriotóxico o bacteriostático [242]. Estas toxinas atacan específicamente a la sínstesis de la pared celular. Así, fosforilan el precursor del peptidoglucano UDP-N-acetilglucosamina (UNAG) formándose UNAG-3P, el cual se acumula en el citosol e impide la acción de la enzima UDP-N-acetilglucosamina enolpiruvil transferasa (MurA), clave para la formación del peptidoglicano [255]. Comúnmente se describen como operones PSK (del inglés *Post-Segregational Killing*) implicados en el mantenimiento y estabilización de plásmidos o elementos de fagos, de tal forma que aquellas células que los pierdan en el proceso de división celular mueren bajo el efecto de la toxina, ya que al perderse el operón se paraliza también la producción de la antitoxina necesaria para contrarrestar su efecto [256].

Actualmente existen dos grandes tipologías de este tipo de sistemas: 1) aquellos sistemas ε-ζ codificados en plásmidos y que contribuyen a la estabilización de los mismos; y 2) los sistemas épsilon-zeta codificados en el cromosoma bacteriano y que parecen contribuir a

la virulencia de los microorganismos patógenos en los que se encuentran [255,256]. El sistema observado en el aislamiento MHSD3 no puede clasificarse de forma definitiva en ninguno de los dos tipos con la información disponible. Tal como se ha descrito en la sección de resultados, justo delante del extremo 5' del gen que codificaría para la potencial toxina zeta se encontró un pequeño gen solapado que codificaba para una HP, en la cual mediante la consulta en Pfam se encontraron dominios relacionados con antitoxinas y que la situaban dentro de la familia de proteínas ParD-like (PF11903), siendo ParD la antitoxina del STA ParDE. Por este motivo y en base a que encajaba en tamaño y posición en las inmediaciones de la toxina, para completar el sistema se apuntó a esta proteína como potencial candidata a antitoxina. Por su parte, la toxina fue incluida en la familia PF13671 o AAA_33 (del inglés *ATPases Associated with diverse cellularActivities*), incluida en la superfamlia P-loop que contienen nucleósido trifosfato hidrolasa (IPR027417) o de forma abreviada P-loop NTPasa; y que está constituida por proteínas que presentan el dominio P-loop, generalmente característico de proteínas con actividad quinasa. Sin embargo, los integrantes de este grupo difieren mucho entre ellos en cuanto a actividad, estabilidad y mecanismo de acción. La característica común a todas ellas es que el dominio P-loop sirve de punto de unión al ATP, cuya utilización es clave para llevar a cabo su función. En el caso concreto de organismos procariotas, las acciones de estas proteínas suelen radicar en proteolisis o actividad chaperona [257].

El análisis preliminar de sus secuencias proteicas situó a ambas proteínas como dos elementos bastante genuinos en cuanto a estructura primaria. Sólo un pequeño conjunto de proteínas codificadas en los genomas de las cepas *Mycobacterium* QIA-37, *M. chelonae* 1558, 15517, 15518 y 203 reflejaron la presencia de sistemas prácticamente idénticos atendiendo a los valores de cobertura e identidad calculados y que tanto en la toxina como la antitoxina fueron cercanos al 100 % en cuanto a cobertura y con identidades superiores al 88 %. Merece destacarse el hecho que, a parte de estas secuencias, el siguiente mejor resultado obtenido correspondió a secuencias de proteínas que, aunque mostraban coberturas superiores al 90 %, las identidades obtenidas eran inferiores al 68 %, reflejando una caída abrupta en cuanto a identidad cuando se trata de proteínas que no proceden de cepas cercanas a *M. chelonae*.

Los dendrogramas obtenidos para la toxina y la antitoxina mostraron valores de soporte de agrupación que, en general, daban lugar a un dendrograma con nodos muy estables. En el caso de la secuencia candidata a toxina zeta se obtuvo una clara distinción entre aquellas proteínas procedentes de MCL y las originarias de MCR, formándose dos ramas principales bien destacadas. Como excepción, dentro de la agrupación dominada por secuencias de MCL apareció una proteína de *M. phlei*. Esta excepción podría deberse a un fenómeno de transferencia lateral. Por su parte, dentro de la rama obtenida para MCR aparece una proteína de *M. vulneris*, una micobacteria del grupo MCL. Esta distinción no es tan clara en el caso del árbol de la antitoxina, donde aparece una gran rama con especies de MCL patógenas y representantes del grupo MCR.

En ambos dendrogramas (Figuras 6.19A y 6.19B) se obtuvo una agrupación de proteínas idénticas compartidas por especies del MTC (*M. tuberculosis*, *M. caprae*, *M. africanum*, y *M. orygis*), aunque en este caso la secuencia procedente de *M. canetii* parece ser ligeramente diferente al resto. Asimismo, aparece una segunda agrupación dentro de la colección de MCL y que corresponde a proteínas altamente similares compartidas por subespecies del complejo *M. avium*, donde podemos encontrar importantes patógenos como *M. avium* subsp. *paratuberculosis* [258].

El alto porcentaje de identidad obtenido entre las secuencias proteicas de STA de la cepa MHSD3 y las secuencias procedentes de las cepas *Mycobacterium* sp. QIA-37, *M. chelonae* 1558, 15517, 15518 y 203 quedó reflejado en ambos dendogramas, agrupándose siempre conjuntamente en una misma rama (con valores que dan soporte a la agrupación del 100 %). Tanto en el caso de la toxina como la antitoxina la agrupación apareció como una rama completamente independiente de las otras agrupaciones de MCL y MCR. Todas estas premisas apuntaron al hecho de que, probablemente, se trataba de un STA nuevo desde el punto de vista de la secuencia, no en cuanto a organización y composición de sus genes.

El estudio genómico de la sintenia entre estas cepas permitió la identificación de genes con dominios de transposasas/integrasas a 2 kb corriente arriba del gen de la potencial antitoxina, y en su proximidad aparecieron genes implicados en resistencias a antibióticos (Figura 6.20), a excepción de la cepa *M. chelonae* 203 en cuyo genoma se pudo encontrar un gen potencialmente implicado en la respuesta frente a estrés provocado por agentes

como el peróxido, que es capaz de actuar ya sea como agente oxidante o como reductor. El mero hecho de la estabilización de estos genes de resistencia o de respuesta a condiciones de estrés, podría estar detrás de la justificación de la presencia de este sistema, favoreciendo la virulencia de estas cepas, por ejemplo, ayudándoles a hacer frente a los tratamientos con antibióticos. La cercanía de una transposasa lo dotaría de la característica de ser un elemento potencialmente móvil y, por lo tanto, de su eventual transmisibilidad.

El ensayo experimental con los elementos genéticos propuestos del sistema HP-toxina zeta de *Mycobacterium* sp. MHSD3 confirmó su funcionalidad básica como STA. Así, el efecto de inducir la expresión de la potencial toxina se tradujo en la inmediata atenuación del incremento de la DO del cultivo en el que se añadió sólo IPTG. Por otra parte, el elemento correspondiente a la secuencia considerada como hipotética antitoxina, no sólo no fue tóxica, sino que además atenuó el efecto de la toxina si atendemos a la evolución de los valores de DO obtenidos al inducir simultáneamente los dos elementos. Acorde a estos valores realmente actuó como antídoto en el sistema experimentalmente aplicado. La diferencia entre inducción de toxina o ambos produjo a la vez una manifiesta reducción en el cultivo de hasta dos órdenes de magnitud en la escala logarítmica, atendiendo a los resultados del recuento de UFC/ml. Por tanto, experimentalmente se confirmó que ambas proteínas presentan las actividades esperadas y compatibles con un STA. Además, para las dos bandas identificadas mediante electroforesis en gel de poliacrilamida, y correspondientes a los respectivos tamaños esperados para la toxina y la antitoxina; el análisis por MALDI-TOF MS corroboró de manera inequívoca su presencia e identificación en las correspondientes bandas de proteína recuperadas del gel.

Los análisis hechos a través de I-TASSER para la caracterización estructural de ambas proteínas precisaron que la antitoxina se trataba de una proteína constituida por cinco hélices α (Figura 6.21B); mientras que en el caso de la toxina correspondería a una proteína formada por siete hélices α y cinco láminas β (Figura 6.21A). El proceso de predicción estructural identificó como análogo estructural más próximo (con una cobertura del 99 % y un TM-Score de 0,809) la toxina ζ del sistema plasmídico ε-ζ de *Streptococcus pyogenes* [242]. Esta última es más larga en secuencia que la toxina problema, coincidiendo ambas sólo en la zona que correspondería al núcleo central de la

proteína, más concretamente en las posiciones que abarcan la región de las cinco hélices α y dos láminas β. En cualquier caso, la coincidencia no es exacta y son bastante diferentes en cuanto a la posición del resto de elementos. La toxina zeta, como proteína perteneciente a la familia AAA_33, debería disponer del respectivo punto de unión al ATP comentado previamente. Conviene aquí recordar que con la ayuda de I-TASSER se detectó un punto de coordinación con este sustrato que presentó un C-Score de 0,69, valor que da validez al resultado, sin llegar a certificar que los 11 residuos que se destacaron como implicados ciertamente formen parte del punto de unión al ATP. En base a los trabajos previos de identificación del centro activo de toxinas zeta de Mutschler y colaboradores [259], se intentó reproducir el alineamiento de estructuras incluyendo la secuencia proteica de la toxina correspondiente al aislamiento MHSD3. En este alineamiento, originalmente se identificaron las posiciones que son clave por una parte en la interacción con UNAG, así como el punto de interacción con el ATP también conocido como motivo Walker-A [260]. Los resultados no fueron concluyentes para la proteína problema.

En el caso de la antitoxina, el análogo estructural con el que se obtuvieron los mejores valores estadísticos fue a la helicasa PriA de *Klebsiella penumoniae* (con una cobertura de la secuencia problema del 83,2 % y un TM-Score de 0,548), una proteína considerablemente más grande y con la que no guarda ningún tipo de relación. Sin embargo, al ser una proteína de unión al ADN, la coincidencia estructural podría residir en la región con la cual interactúa con el ADN. Ésta podría ser una pista que nos llevaría a su proposición como candidato a actuar como regulador de la expresión del operón de la antitoxina problema. En este caso la coincidencia estructural se centró en una subunidad del dominio helicasa de dicha proteína, concretamente el definido como lóbulo helicasa I (*helicase lobe 1*) en la descripción experimental de su estructura [261]. Como fue incluido por Pfam dentro de la familia de proteínas similares a ParD, y puesto que la estructura de ParD de *Escherichia coli* también ha sido confirmada experimentalmente, se procedió a realizar un estudio comparativo de secuencias y modelado por homología entre ambas para comprobar si era posible realizar una predicción estructural más precisa. Los resultados mostraron diferencias concluyentes tanto en la secuencia como en la estructura entre ParD y la hipotética antitoxina de la cepa MHSD3.

En cualquier caso, vuelve a hacerse patente la dificultad de mejorar la predicción con modelos que respondan a una precisión razonable la estructura de la antitoxina, tal y como ocurrió con anterioridad para las antitoxinas VapB. Ello es debido en gran medida a las diferencias existentes entre sus estructuras terciarias, acentuadas por la dificultad de hallar una relación óptima perceptible en el alineamiento de la secuencia objeto de estudio y las posibles secuencias relacionadas existentes en la actualidad en las bases de datos.

6.4.3. Sistemas MT0933-MT0934 de *Mycobacterium llatzerense* y potenciales sistemas de 3 componentes (MT0933-Lipasa-MT0934)

El sistema MT0933-MT0934, también conocido como Rv0909-Rv0910, es un STA descrito originalmente en *M. tuberculosis* H37Rv como un sistema altamente conservado en dicha especie y que, al ser expresado en *M. smegmatis*, provoca una clara inhibición del crecimiento, presentándose a su vez como el representante de una nueva familia en el momento de su descubrimiento [238]. Tal y como se destacó en el apartado de resultados, los genes anotados bajo el nombre de estos componentes se han encontrado en diferentes configuraciones: una organización típica toxina-antitoxina, compuesta por dos pequeños genes muy cercanos o solapados entre ellos, en *M. llatzerense* MG13^{T}; una organización de tres componentes con la toxina y la antitoxina separadas por el gen de una lipasa y, por último, toxinas MT0934 sin una antitoxina asociada.

En todos los casos, el análisis Pfam situó a todas las antitoxinas, lipasas y toxinas en las mismas familias: familia antitoxina MT0933, lipasa secretora (incluida en la superfamilia alfa/beta hidrolasa) y poliquétido ciclasas 2, respectivamente. En relación a las antitoxinas no hay mucha información disponible más allá de la que hace referencia a que se trataría de una familia de antitoxinas pertenecientes a STAs. En el caso de la familia lipasa secretora, algunos de sus integrantes están implicados en el procesamiento de compuestos lipídicos con fines muy diversos. Para la que existe más información en este caso, es la familia asignada a la toxina MT0934. Las enzimas conocidas como poliquétido ciclasas están implicadas en la síntesis de lo que se conoce como poliquétidos, metabolitos secundarios de naturaleza lipídica biológicamente muy activos y que son producidos por determinados microorganismos para favorecer su propia supervivencia [262]. Entre estos compuestos se encuentran antibióticos de gran relevancia clínica como

son la tetraciclina, la eritromicina o la daunorubicina; así como otros compuestos que actúan como elementos estructurales de la pared bacteriana, sideróforos y fungicidas entre otros, aunque la función real de los poliquétidos dentro del orden *Actinomicetales* todavía no está clara del todo [263–266].

En general, las ciclasas se encargan del paso de circularización del poliquétido, formando un anillo aromático y eliminando agua hasta constituir el compuesto final [267]. Para ello estas proteínas tienen una estructura típica que incluye el dominio START (del inglés *STAr-Related Lipid Transfer*), conformado por un núcleo central de láminas β, sobre las cuales se situarían dos estructuras helicoidales, formándose un hueco que acomoda el sustrato. Existen diferentes variantes de esta estructura, entre las cuales la estructura más similar en este caso fue la correspondiente a la variante BA (del inglés *Birch Allergen*) o CSD (del inglés *Classic Start Domain*), constituida por tres hélices α y siete láminas β tal [268]. En todos los casos hallados en los genomas estudiados, una estructura muy similar entre ellos se pudo reproducir prácticamente sin variaciones considerables, mostrándose el caso de *M. llatzerense* $MG13^T$ a modo de ejemplo de la Figura suplementaria 3 (Anexo 3). No obstante, a fecha de conclusión de la presente tesis, y para estas supuestas toxinas, dentro del género *Mycobacterium* no había disponibles en las bases de datos modelos estructurales determinados experimentalmente para ser utilizados como plantillas relacionadas, y tampoco se pudo llevar a cabo la modelación con una exactitud razonable.

Tal y como se ha indicado previamente, el comportamiento del módulo genético Rv0910-0909 de *M. tuberculosis* H37Rv, mediante clonación y transformación en *M. smegmatis*, responde a un STA en el que se observa una inhibición del crecimiento bacteriano al inducir el gen Rv0910, atenuado al expresar conjuntamente su potencial antitoxina Rv0909 [238]. A pesar de estas evidencias experimentales en *M. tuberculosis* H37Rv, ninguno de los supuestos STA y en cualquiera de sus formatos hallados en los genomas estudiados mostró efecto tóxico alguno. Esto puede deberse a diferentes motivos. El más importante sería que podrían ser genes pertenecientes a sistemas más grandes involucrados, por ejemplo, en la producción de poliquétidos con distintos fines. Así, toda vez que en el presente estudio se clonaron y transformaron en *E. coli*, posiblemente al separarse de los genes acompañantes en el genoma original de procedencia, no estarían recibiendo el sustrato necesario para generar su producto final y, por lo tanto, su efecto.

De hecho, en los casos de *M. chelonae* y *M. immunogenum* muy cerca de dichos genes y en su configuración de tres componentes, aparece el gen de una poliquétido sintasa (PKS), con la cual muy probablemente están relacionados. Estas PKS son el eje central para el proceso de síntesis de los poliquétidos y, de hecho, haciendo un rastreo a partir de los datos del pangenoma de MCR se detectaron un total de 175 PKSs en hasta 45 de los 52 genomas utilizados. Estas evidencias apuntarían al grupo de las MCR como una potencial fuente de este tipo de compuestos, con gran interés biotecnológico y clínico.

En cualquier caso, en el contexto de los objetivos planteados, al no observarse el efecto tóxico esperado no se profundizó en su caracterización, pero sí se establecieron las bases para futuros estudios centrados en este grupo de proteínas dentro de las MCR.

6.4.4. Sistemas phd-Doc y toxinas zeta sin antitoxina.

Los sistemas phd-Doc, al igual que los sistemas épsilon/zeta, pueden relacionarse en algunos casos con la estabilización de plásmidos u otros elementos génicos móviles dentro de la población bacteriana [269]. En trabajos de laboratorio adicionales llevados a cabo con la toxina Doc se observó un efecto tóxico sobre la población bacteriana inmediatamente después de su inducción, produciéndose un frenado brusco en el incremento de la DO del cultivo con respecto al control (datos no mostrados, Trabajo de Fin de Grado del Sr. Víctor Fernández, 2015-2016). Sin embargo, en la secuencia de la hipotética proteína candidata a antitoxina y que encajaba en tamaño y ubicación en el genoma con una toxina típica, no se encontró ningún dominio en el análisis Pfam y tampoco fue capaz de contrarrestar el efecto tóxico, mostrando una curva de crecimiento semejante a la del cultivo con la toxina inducida.

Por su parte, las toxinas zeta sin antitoxina presentaron un tamaño completamente fuera de los parámetros que corresponderían a una toxina estándar, con más de 1 kb de longitud en su secuencia nucleotídica, lo que se traduce en una proteína de más de 330 aminoácidos. Sin embargo, a raíz de que se detectó en estas proteínas el dominio AAA, como en el caso del sistema épsilon/zeta de MHSD3 descrito previamente; y a pesar de que este dominio puede estar implicado en múltiples funciones, se decidió llevar a cabo los respectivos ensayos experimentales, los cuales permitieron descartar de manera definitiva su posible efecto tóxico. Por ello no se profundizó más en su caracterización.

7. Discusión general de los resultados

El camino para abordar el estudio de las capacidades ecológicas y clínicas de las micobacterias de crecimiento rápido, y más concretamente de las especies *M. chelonae*, *M. immunogenum* y el complejo *M. abscessus*, se inició con el primer objetivo propuesto de incrementar el número de genomas representativos de cada especie, especialmente de las dos primeras.

El primer paso para la secuenciación de un genoma pasa por la obtención de muestras de ADN en cantidad y calidad óptimas para dicho proceso. La obtención de ADN de micobacterias no es sencillo, debido fundamentalmente a la gruesa y resistente pared celular que las envuelve, complicado además por la presencia de elevadas cantidades de sustancias hidrofóbicas de base lipídica. Estas características obligan al uso de protocolos más agresivos, a menudo implicando tratamientos mecánicos, con el fin de romper esta barrera y liberar las moléculas de ADN. Los esfuerzos dedicados en los progresivos pasos de optimización aplicados, tanto a nivel de rotura celular como de la propia purificación del ADN obtenido, permitieron establecer un protocolo general de extracción eficaz para todas las cepas utilizadas. Los puntos más críticos fueron la inclusión en el protocolo definitivo de un pretratamiento basado en la rotura mecánica con ayuda de microesferas de vidrio, seguido de un tratamiento enzimático con proteinasa K utilizando el tampón de lisis ATL (Qiagen); y finalmente un paso de purificación altamente eficaz. La optimización del paso de rotura mecánica fue especialmente complicada, ya que implica una liberación temprana del ADN. Efectivamente, la degradación del ADN empieza prácticamente en el mismo momento de iniciarse el proceso, especialmente debido a los efectos de cizalladura mecánica que provocan los cambios bruscos y fuertes en la intensidad y/o dirección de las fuerzas físicas que intervienen, por otra parte, difíciles de evitar. Establecido un tiempo óptimo de cinco minutos para este paso, se consiguió minimizar la fractura física del ADN, lo cual es especialmente importante cuando se desea ADN de alto peso molecular requerido, por ejemplo, en plataformas de secuenciación como PacBio.

En general, los esfuerzos invertidos en la obtención de ADN de alta calidad se reflejaron en la buena calidad de las lecturas obtenidas, tanto en las lecturas obtenidas con la tecnología Illumina como las lecturas generadas con tecnología PacBio. Posiblemente

todo ello contribuyó decisivamente al cierre de los genomas de las cepas tipo de *M. chelonae* y *M. immunogenum*, así como a la obtención de ensamblajes de genomas a nivel de *draft* de alta calidad para el resto de cepas. En lo que se refiere a *M. llatzerense* MG 13^T hay que destacar que este fue un caso más complejo. La baja calidad de las lecturas procedentes de las primeras tandas de secuenciación en las plataformas 454 y PacBio, junto con la complejidad implícita del genoma en sí (presencia de elementos repetitivos, elementos de fagos, transposasas, integrasas, etc.) complicaron enormemente la resolución del ensamblaje, impidiendo cumplir el objetivo final para esta cepa de llevar a cabo el cierre definitivo del genoma con las suficientes garantías de minimización de errores en el ensamblaje. En este sentido, todos los genomas fueron pulidos y revisados a través de programas como REAPR o FRCurves, con el fin de asegurar bioinformáticamente y en la medida de lo posible la ausencia de incongruencias estructurales en la continuidad de las secuencias derivadas.

Los ensamblajes de genomas considerados de alta calidad fueron anotados con el fin de caracterizar y utilizar sus proteomas en estudios comparativos basados en el cálculo de pangenomas (incluidos el genoma esencial y el accesorio). Estos estudios se realizaron para obtener de una forma general las interrelaciones genómicas entre el complejo grupo de especies que conforman las MCR (incluyendo para ello los genomas de las 17 especies de MCR disponibles en las bases de datos en el momento de iniciar el estudio), o una visión más específica en la que el estudio derivó primero en el análisis de las cepas que conforman el grupo de especies *M. chelonae*, el complejo *M. abscessus* y *M. immunogenum*; y en segundo lugar en el análisis del pangenoma de la especie *M. immunogenum*.

El análisis detallado y comparado del genoma esencial en el contexto del grupo de MCR permitió establecer las relaciones evolutivas de las especies representadas, relaciones basadas en las posiciones homólogas de aminoácidos de todas las proteínas compartidas y codificadas por genes presentes en todos los genomas y en copia única. El contexto evolutivo adecuado del estudio se estableció primero mediante la elaboración de un árbol filogenético construido a partir de sus secuencias de ADNr 16S. Las relaciones establecidas en ambos casos se mantuvieron en líneas generales, aunque el poder discriminativo basado en el análisis del genoma esencial fue muy superior, debido a la

utilización de mayor cantidad de información. Esto fue especialmente relevante en la discriminación de especies estrechamente relacionadas como *M. chelonae*, el complejo *M. abscessus* y *M. immunogenum*, grupo en el que el ADN 16S tiene serias dificultades para su resolución. Además, en ambos casos se detectaron genomas que parecen no pertenecer a la especie inicialmente asignada, como *M. fortuitum* Z58, *M chelonae* 1518 o *M. rhodesiae* NBB3, lo que pone de manifiesto el especial cuidado que debe tenerse en confirmar las especies a las que pertenece un genoma como paso previo a un estudio más profundo, con el fin de evitar errores en los resultados y sobre todo en la extrapolación de conclusiones a partir de los mismos.

Por su parte, en la consideración del pangenoma del grupo MCR desde el punto de vista ecológico, destaca primero el hecho de que presentan como conjunto un pangenoma claramente abierto. Este hecho justificaría la gran diversidad de nichos ecológicos ocupados por las distintas especies que conforman el grupo, así como la gran diversidad funcional que han desarrollado durante el proceso evolutivo, lo cual queda patente en la ganancia y la pérdida de genes con el fin de adaptarse para prosperar en los distintos ambientes acorde a las presiones ambientales y necesidades de supervivencia sufridas en cada caso. La tendencia abierta de su pangenoma apunta a que la incorporación de cepas y nuevas especies en el estudio permitiría el incremento del número de familias proteicas diferentes que se pueden encontrar. Además, la representación gráfica del pangenoma en forma de dendograma (basado en la presencia o ausencia de proteínas), permitió ver cuán similares son las distintas especies en función de las proteínas compartidas del pangenoma en su conjunto, que no necesariamente en la secuencia de las mismas. Este hecho reveló que algunas especies se alejan de las que se encontrarían evolutivamente más cercanas desde el punto de vista del genoma esencial, debido a que comparten un mayor número de proteínas con otras especies evolutivamente más alejadas. Este dato podría ser un reflejo de la compartición de hábitat. En este sentido, especies que comparten un hábitat o nicho ecológico son más propensas a intercambiar genes o desarrollar roles similares, haciendo que funcionalmente se alejen de especies incluso evolutivamente más cercanas.

Avanzando hacia un aspecto más concreto, los estudios de genoma esencial realizados con genomas del conjunto de especies con relevancia clínica, es decir, *M. chelonae*, el

complejo *M. abscessus* y *M. immunogenum* dieron como resultado la obtención de una buena discriminación en cuanto a las relaciones evolutivas existentes entre los distintos genomas representantes de dichas especies. Así, las cepas de las especies *M. chelonae* y *M. immunogenum* fueron perfectamente separadas a nivel de especie entre ellas y con respecto al complejo *M. abscessus*. No obstante, el aspecto de más interesante se obtuvo precisamente dentro de este último complejo. Así, en base al genoma esencial *M. abscessus* se pudo subdividir hasta en tres agrupaciones, cada una marcada por cepas clave: 1) *M. abscessus* ATCC 19977^T, 2) *M. abscessus* subsp. *bolletii* CCUG 50184^T y BD^T, 3) *M. abscessus* subsp. *bolletii* CCUG 48898 (antes denominada *M. masiliense*). La topología del árbol parecía indicar la presencia de tres subespecies dentro del grupo, sospechas que fueron confirmadas por un estudio de reciente publicación [101]. En dicho estudio, los autores presentan un árbol construido sobre una matriz de distancias de valores de ANI que refleja una topología similar a la obtenida con el genoma esencial en el presente estudio, a pesar de no utilizar los mismos genomas. En cualquier caso, en ambas aproximaciones los tres grupos están marcados por las cepas de referencia mencionadas, siendo *M. abscessus* subsp. *bolletii* CCUG 48898 la propuesta como la cepa tipo de *M. abscessus* subsp. *masiliense.* Por lo tanto, la propuesta del estudio de subdividir el complejo de *M. abscessus* en *M. abscessus* subsp. *abscessus,* M. *abscessus* subsp. *bolletii* y *M. abscessus* subsp. *massilliense* se ve perfectamente respaldada por los datos basados en el genoma esencial obtenido en el actual trabajo. Además, estos resultados aportan información útil para una futura reclasificación de los genomas implicados, en función del genoma de la cepa tipo que domina la rama donde se agruparon.

El estudio del pangenoma sobre este mismo grupo reflejó además que, desde el punto de vista de compartición de agrupaciones basadas en familias proteicas, se conserva la topología del árbol obtenido con el genoma esencial. Evidentemente esto significa que los genomas evolutivamente más cercanos son también los que comparten más genes. Por otra parte, la tendencia claramente abierta de la evolución del pangenoma indicaría que, en su conjunto, estas especies pueden tener una gran capacidad de intercambio y variación de su repertorio genético, lo cual es importante al ser un grupo capaz de desarrollar infecciones oportunistas. Esta alta capacidad de intercambio conduce irremediablemente a la posibilidad de adquisición de mecanismos de patogenicidad que contribuyan a un aumento de su virulencia. Este fenómeno es más significativo cuando se trata de estudios

hechos sobre una sola especie, como es el caso de *M. tuberculosis*, donde la tendencia evolutiva de su pangenoma derivada del análisis comparativo con los genomas disponibles es claramente abierta. En este sentido es importante tener este hecho en cuenta también en los casos que afectan a las especies de MCR con relevancia clínica adquirida año tras año.

El estudio del pangenoma centrado en la especie *M. immunogenum* con genomas procedentes de cepas ambientales y clínicas, puso de relieve varios hechos de interés en esta especie. En primer lugar, el análisis del genoma esencial puso de manifiesto que todos los aislamientos eran prácticamente idénticos. Efectivamente, las cepas analizadas comparten más del 80 % de las proteínas, con una conservación de secuencia en las posiciones homólogas superior al 99 %. Además, el pangenoma presentó una tendencia claramente cerrada, con una rápida estabilización de la curva resultante tras añadir los primeros genomas al análisis. Por tanto, no se observaron muchas diferencias en la dotación génica de las cepas de *M. immunogenum* como para poder aportar nuevas familias proteicas a medida que se incorporaran más genomas al análisis, en clara contraposición a lo que ocurre, por ejemplo, en el pangenoma del patógeno primario *M. tuberculosis*. Un hecho a destacar que sí reveló el estudio del pangenoma de *M. immunogenum*, más concretamente en el apartado que afecta a los genes específicos de cada genoma, es la presencia en las cepas clínicas SMUC14 y CCUG 47286^{T} de una posible proteína de defensa contra antibióticos y de una posible proteína implicada en la invasión celular, respectivamente. En referencia a las cepas ambientales se trata de una diferencia importante, ya que mostraría la adquisición de funciones para la defensa frente a tratamientos biocidas o para a favorecer el proceso de infección en cepas que se están desarrollando en el cuerpo humano, toda vez que el microorganismo ha estado sometido a las presiones que le conducen por esta vía.

La información obtenida gracias al estudio comparativo del pangenoma aporta una visión general de la capacidad de adaptación de una especie o conjunto de especies, atendiendo a la tendencia abierta o cerrada de su pangenoma. Este tipo de información permite situar a las especies en su marco ecológico global como paso previo a la descripción detallada de sus capacidades clínicas. En relación a este último aspecto, son muchas las familias proteicas encontradas que potencialmente estarían relacionadas con todos aquellos

aspectos que contribuyen a definir la patogenicidad de una bacteria. En primer lugar, se encontraron resistencias contra hasta 18 tipos diferentes de antibióticos en el conjunto de todos los genomas secuenciados de las especies *M. chelonae*, *M. immunogenum* y *M. abscessus*, incluyendo un buen número de bombas de expulsión de antibióticos por genoma (de 10 a 14), sin olvidar mutaciones o la presencia de proteínas implicadas en la resistencia a cuatro agentes antituberculosos como son la isoniacida, etambutol, rifampicina y piracinamida. Por otra parte, también se hallaron numerosos elementos de resistencia a agentes externos, que como mínimo comprometen la viabilidad biológica del microorganismo, como metales pesados, condiciones ácidas e incluso desinfectantes de uso común. El conjunto es un buen reflejo del contexto genómico que podría explicar el camino seguido por estas MCR en su avance desde las aguas naturales o de un sistema artificial de distribución de aguas hasta llegar al paciente. Para ello, han tenido que superar las diferentes barreras interpuestas. Estas incluyen los desinfectantes añadidos en los sistemas de suministro de aguas o utilizados en la limpieza de las propias instalaciones. No hay que olvidar que estos microorganismos se han encontrado contaminando instrumental hospitalario y superficies, en gran medida debido a estas resistencias, llegando en última instancia a infectar al paciente. En relación a este último punto la propia información genómica derivada del presente estudio deja en evidencia la dificultad de tratar las infecciones producidas por estas micobacterias debido precisamente al elevado número de potenciales resistencias encontradas. Así, una de las resistencias a antibióticos más relevantes implica al grupo de las MBLs. En este contexto, cabe mencionar el hecho de la detección de representantes de MBLs en todos los genomas analizados en el presente trabajo, además de experimentalmente demostrar que son cepas productoras de este tipo de enzimas. El análisis comparado detallado basado en el alineamiento de secuencias de las mismas puso de manifiesto cómo, al compararse con el resto de MBLs presentes en la base de datos CARD, las secuencias de MBLs halladas en *M. chelonae*, *M. abscessus* subsp. *bolletii* y *M. immunogenum* formaban una agrupación independiente y estable con respecto al resto. Este dato apunta claramente hacia la posibilidad de tratarse un nuevo tipo de MBL y que contribuiría a complicar todavía más el perfil de resistencias de estas tres especies.

Continuando con las implicaciones clínicas, mención aparte merece el hecho de la detección de un buen número de potenciales factores de virulencia y que cubren varios

aspectos relacionados con el potencial patógeno de estos microorganismos. Entre ellos se encontraron elementos implicados en la captación y transporte del hierro, un proceso de gran importancia para los patógenos en general. También se hallaron elementos implicados en la supervivencia en condiciones ácidas, protección frente al peróxido de hidrógeno (producido, por ejemplo, por macrófagos), elementos de supervivencia a distintas condiciones de estrés (falta de nutrientes, uso de fuentes de carbono alternativas, hipoxia, presencia de agentes biocidas externos, etc.), elementos implicados en la potencial producción de agentes antimicrobianos como mecanismo para eliminar la competencia por nicho ecológico, o de genes implicados en la comunicación celular o mejor percepción de quórum (QS) y para actuar como sensores del ambiente circundante. Además, se detectó la presencia de proteínas potencialmente implicadas en la interacción con células epiteliales, formación de células persistentes, o genes relacionados con la penetración de células no fagocíticas de mamíferos (operones *mce*), así como de elementos que podrían contribuir a la remota, pero no descartable, posibilidad de formación de cápsula en estas especies.

Uno de los aspectos importantes que se ha considerado en este estudio y mencionado antes, por su posible relación con la patogenicidad de los microorganismos implicados, hace referencia a los elementos implicados tanto en sensar los parámetros físico-químicos del ambiente como la comunicación celular a través de la percepción de Quórum. En esta categoría se pudieron detectar proteínas implicadas, por ejemplo, en la síntesis de señales de percepción del Quórum o síntesis de compuestos cuya producción está gobernada por esta funcionalidad (como es el caso de la toxoflavina). Pero no sólo se encontraron elementos relacionados con la síntesis de señales implicadas en QS, sino también proteínas implicadas en el transporte a través de la membrana de dichas señales. Dichos transportadores controlarían el flujo de las señales de QS entre el interior y el exterior celular, permitiendo iniciar los cambios en la expresión génica necesarios para responder a dicha señal. En este contexto también se puso de manifiesto la presencia del SDC kdpB/C, cuya regulación puede estar gobernada por QS, entre otros factores. Este SDC permitiría a la bacteria determinar si se encuentra en el interior o exterior celular en función de los niveles de K^+ (kdpB/C) [207,208].

En cualquier caso, el QS puede ser vital en la modulación de la expresión génica de de elementos relacionados con la patogenicidad, como pueden ser aquellos que codifican factores de virulencia [2]. Por este motivo, en el presente trabajo se consideró importante tener en cuenta los mecanismos de QS como un aspecto más para definir la patogenicidad potencial de las especies *M. chelonae*, *M. immunogenum* y el complejo *M. abscessus*. Para añadir importancia a este hecho, en muchos de los aspectos referentes no solo al QS, sino también a los FV y resistencias expuestas previamente, la proteína equivalente en *M. tuberculosis* ha sido fruto de profundos estudios que se ha llegado a demostrar, o por lo menos mostrar indicios, de la implicación real de estos elementos en la patogénesis del agente causal de la tuberculosis [9,199,200,207,208], apuntándose en algunos casos como potenciales dianas para el futuro desarrollo de fármacos alternativos para combatir la infección. En paralelo, se podría pensar que ese mismo papel podría estar beneficiando la faceta de patógeno oportunista de *M. chelonae*, *M. immunogenum* o el complejo *M. abscessus*, y que podrían ser tratados al mismo nivel que en la especie patógena, pero adaptando los estudios a estas MCR.

No sólo es importante definir a nivel genómico los elementos potencialmente implicados en cualquiera de los aspectos de la patogenicidad destacados previamente, sino también la presencia de elementos reguladores de su expresión y que aseguren la acción de cada elemento en el momento y nivel de expresión adecuado. En este sentido se encontraron representantes de familias reguladoras que pueden estar implicadas en la modulación de la expresión de genes de resistencia a antibióticos, de defensa contra otros agentes químicos externos (peróxido de hidrógeno, metales pesados, etc.), situaciones de estrés (choques térmicos, hipoxia, falta de nutrientes), modulación de la expresión de FV o elementos implicados en la QS ya menconados. En definitiva, proteínas reguladoras de todos y cada uno de los aspectos destacados en la disección genómica de la patogenicidad de las cepas estudiadas. Si consideramos todos los aspectos en conjunto, estos genomas presentan además de un importante repertorio de elementos que pueden favorecer enormemente su faceta de patógeno oportunista, los elementos reguladores necesarios para modular la expresión de los mismos y en el momento más adecuado. La importancia de este hecho radica en que les permitiría adaptarse para hacer frente a diferentes situaciones y escenarios, favoreciendo tanto la penetración como la evasión de las

barreras defensivas interpuestas por el organismo hospedador, y en consecuencia la progresión de la infección.

Uno de los elementos a destacar en el presente estudio y que ha sido objeto de una demostración experimental de la funcionalidad de los elementos constituyentes, son los denominados sistemas toxina-antitoxina o STA. A pesar de no estar oficialmente reconocidos como factores de virulencia, algunos de los aspectos de su funcionalidad sí que podrían contribuir de forma significativa a la patogenicidad de un determinado microorganismo [225–228]. Esta podría ser parte de la justificación del hecho por el que *M. tuberculosis* acumula habitualmente tal número de estos sistemas en su genoma [238]. En la búsqueda sobre el genoma de *M. tuberculosis* CR-UIB2 de estos elementos se encontraron hasta 68 STA, de los cuales 48 correspondían a los STA de tipo II VapBC; un número que debe reflejar algún tipo de función biológica altamente beneficiosa para el patógeno en cuestión y que le lleva a acumularlos en su dotación genómica. Sabiendo que entre las funciones de los STA estarían las del mantenimiento de genes importantes para la virulencia (como por ejemplo resistencias a antibióticos) o la de inducción a un estado de letargo ante una condición ambiental adversa [242,244,255,256]; cobra fuerza la hipótesis de que estos sistemas pueden beneficiar de forma significativa la patogenicidad del agente causal de la tuberculosis, y por extensión de las MCR consideradas patógenos oportunistas que presenten este tipo de sistemas en sus genomas.

Es de destacar que de todos los potenciales sistemas encontrados en los genomas de MCR analizados, sólo realmente tres mostraron en *E. coli* un comportamiento compatible con un STA al aplicar el protocolo apuntado en materiales y métodos. Dos de ellos corresponden a los sistemas VapBC27 y VapBC28, encontrados en el genoma de *M. llatzerense* MG13^{T}. Las toxinas de ambos sistemas mostraron un claro efecto inhibitorio sobre el crecimiento de la población bacteriana en *E. coli*, pero sólo en el caso de VapBC27 se consiguió demostrar claramente el efecto atenuante esperado de la antitoxina asociada. El estudio de dominios de las respectivas secuencias proteicas situó a las antitoxinas en familias asociadas a este tipo de actividad, mientras que las toxinas se incluyeron en la familia de proteínas con dominio PIN, con actividad ribonucleasa. Desde el punto de vista estructural se consiguió modelar a estas toxinas VapC acorde a la estructura típica de dicha familia, así como precisar los aminoácidos que conforman el

centro activo de las mismas. Este proceso no fue concluyente en el caso de las antitoxinas, cuya variabilidad estructural es mayor y, por tanto, más difícil de predecir a partir de la comparación con las estructuras existentes en la actualidad.

El tercer STA que mostró un comportamiento acorde con los resultados esperados fue el correspondiente al sistema épsilon/zeta detectado en la cepa *Mycobacterium* sp. MHSD3. Aunque inicialmente sólo se detectó la toxina zeta como tal, se consiguió identificar una proteína hipotética (HP) corriente arriba de la toxina que encajaba con una posible antitoxina. La caracterización desde el punto de vista estructural y de dominios puso de relieve que ambos componentes estaban teóricamente en disposición de los elementos necesarios para constituir un STA real del tipo II épsilon/zeta: una proteína hipotética con dominios de antitoxina y una toxina con dominios pertenecientes a la familia AAA, con actividad ATPasa y característicos de las toxinas zeta. Además, también se consiguió precisar los potenciales aminoácidos candidatos a constituir el centro de unión al ATP en la toxina zeta. En cuanto al ensayo experimental, la toxina zeta mostró un claro efecto inhibitorio del crecimiento inmediatamente después de su inducción. Por su parte, la HP al ser inducida mostró el efecto atenuante esperado de dicha toxicidad. Además, al comparar las secuencias proteicas de la toxina y la hipotética antitoxina con las secuencias presentes en las bases datos, se consiguió definir un pequeño conjunto de secuencias con una elevada identidad con ellas y codificadas en genomas de cepas de *M. chelonae* (con más de un 80 % de identidad en el 100 % de cobertura). Sin embargo, la presencia de homólogos para esta proteína no fue exclusiva en cepas de *M. chelonae*, sino que también se encontraron en otras especies del género *Mycobacterium*, aunque en este caso las identidades no superaron el 67 % (con más del 90 % de cobertura). Por lo tanto, la presencia de homólogos de la toxina zeta y la P.H. de la cepa *Mycobacterium* sp. MHSD3 no solo se limita a MCR sino también a MCL, incluida *M. tuberculosis.* La demostración experimental de que el sistema de la cepa estudiada en este caso encaja con un STA daría indicios de que dichos homólogos pueden presentar el mismo comportamiento. Esta información conduce a la posibilidad de estar ante una nueva variante en secuencia del sistema épsilon/zeta. Además, en el genoma de la cepa MHSD3 se detectó que este sistema estaba codificado muy cerca de proteínas con dominios característicos de transposasas, lo que añadiría la posibilidad de ser transferible. Adicionalmente, en su vecindad se hallaron genes relacionados con actividad β-lactamasa, por lo que se podría

tratar de una resistencia a antibióticos cuya estabilidad en el genoma estaría siendo protegida por el STA épsilon/zeta.

El resto de genes relacionados con este tipo de sistemas no mostraron efecto tóxico alguno. Los sistemas MT0933-34 hallados junto con una lipasa codificada entre los dos componentes del sistema toxina-antitoxina merecen una mención aparte. Efectivamente, estos elementos, según los resultados obtenidos basados en la exploración de bases de datos de rutas metabólicas, podrían pertenecer a rutas de síntesis de poliquétidos, compuestos de alto interés económico debído a su potencial antimicrobiano, antifúngico o incluso contra la formación de biopelículas, entre otras funciones. En este sentido, las MCR podrían ser una buena fuente de este tipo de compuestos.

En definitiva, teniendo en cuenta toda la información expuesta, podemos concluir que se hallaron claras evidencias genómicas que pueden dar explicación a la capacidad de estas especies de desencadenar una infección en determinadas situaciones, así como de la capacidad de superar diversas de las medidas preventivas destinadas a evitar que estos microorganismos alcancen a potenciales pacientes en el ámbito hospitalario, y de resistir los tratamientos una vez ya los han colonizado. Sin embargo, tal y como se ha destacado anteriormente, en muchos casos estamos ante una exposición meramente descriptiva basada en la información disponible recopilada en el momento de la anotación genómica y, por lo tanto, inicialmente con un mero valor teórico. Por todo ello, y como valor añadido, es necesaria la demostración experimental de su funcionalidad para confirmar los hallazgos, por ejemplo, tal y como se hizo con los STA. Sólo de esa forma se pondrán las bases necesarias para llegar a estar en disposición de definir a estos elementos encontrados como importantes para la patogenicidad del microrganismo y convertirse en auténticas dianas alternativas contra las que actuar mediante el desarrollo de nuevas estrategias de tratamiento.

8. Conclusiones

1. La calidad del ADN en términos de pureza e integridad es generalmente un elemento clave para el proceso de secuenciación de genomas en plataformas de alto rendimiento, pero es especialmente crítico en la generación de librerías para aquellas plataformas que generan lecturas largas, como por ejemplo PacBio.
2. El enfoque basado en el *scaffolding* híbrido es capaz de completar el cierre de un genoma microbiano de 5 Mb de manera directa utilizando coberturas 50x de lecturas Illumina y una única celda SMRT.
3. El programa de ensamblaje Velvet, con lecturas illumina y una cobertura 50x, genera mejores ensamblajes que Newbler y SPAdes en términos de menor acumulación de errores.
4. Como paso previo a la realización de un análisis genómico comparativo, con el fin de detectar cepas mal clasificadas presentes en las bases de datos o establecer las relaciones evolutivas correctas para nuevos aislamientos, es importante el establecimiento del correcto contexto evolutivo de todos los microorganismos objeto de análisis.
5. El establecimiento de relaciones evolutivas basadas en el genoma esencial monocopia tiene un poder discriminativo muy superior a los análisis basados en ADNr 16S o MLSA, permitiendo una mayor resolución para grupos taxonómicamente complejos. Además, en el global del pangenoma de micobacterias de crecimiento rápido las cepas evolutivamente más cercanas no son siempre las que mayor número de proteínas comparten.
6. La gran variedad de nichos ecológicos ocupados por las micobacterias de crecimiento rápido se refleja en el gran número de familias de proteínas incluidas en la sección del pangenoma denominada como "*cloud*", representando hasta un 65,52 %.
7. El estudio de las relaciones evolutivas basadas en el genoma esencial con cepas de las especies *Mycobacterium chelonae*, *Mycobacterium immunogenum* y el complejo *Mycobacterium abscessus* concuerdan y confirman la reciente división del complejo *M. abscessus* en tres subespecies, y que vendrían definidas por las cepas tipo *Mycobacterium abscessus* subsp. *abscessus* ATCC 19977^{T}, *Mycobacterium abscessus* subsp. *bolletii* CCUG 50184^{T} y *Mycobacterium abscessus* subsp. *massiliense* CCUG 48898^{T}.

8. El análisis comparativo del pangenoma de cepas y aislamientos de *Mycobacterium tuberculosis* ha puesto de manifiesto que es claramente abierto, evidenciando una alta capacidad de intercambio de su repertorio genético.
9. El análisis del genoma esencial de la especie *Mycobacterium immunogenum* revelan una alta conservación del mismo, mostrando genomas que son prácticamente idénticos. Además, el pangenoma de esta especie tiene una tendencia claramente cerrada, y consecuentemente una moderada capacidad para incorporar o perder genes en comparación a otras especies analizadas del género.
10. Las cepas de *Mycobacterium immunogenum* de origen clínico presentan en sus genomas genes exclusivos relacionados con su procedencia y con posibles implicaciones en el desarrollo de su patogenicidad.
11. El perfil de resistencias definido genéticamente en los genomas de las cepas analizadas revela un amplio abanico de elementos con posibles implicaciones en la supervivencia frente a diversos agentes externos con potencial biocida, como por ejemplo antibióticos o metales pesados. Estos elementos deberían tener un peso relevante en las dificultades inherentes al control de la propagación o tratamiento de las infecciones de estos microorganismos.
12. Desde el punto de vista de los factores de virulencia, las micobacterias de crecimiento rápido analizadas disponen de toda una serie de mecanismos que favorecerían el desarrollo de la infección, la evasión del sistema inmune, la penetración a través de células epiteliales y su defensa frente a respuestas inmunológicas o situaciones ambientales de estrés (como falta de nutrientes, hipoxia o condiciones ácidas).
13. En lo referente a la percepción del Quórum, los genomas de micobacterias de crecimiento rápido analizados contienen, desde el punto de vista teórico, los elementos necesarios para la síntesis y el transporte de señales implicadas en los mecanismos de comunicación entre bacterias. En algunos casos, la información recabada por estos procesos les permitiría modular la expresión de una serie de genes, incluidos algunos factores de virulencia y mecanismos comprometidos en su supervivencia en el interior del hospedador y, por lo tanto, en su patogenicidad.

14. El análisis del reguloma de los genomas analizados revela la presencia de proteínas reguladoras implicadas en la modulación de la expresión de elementos de patogenicidad acorde a los requerimientos ambientales específicos.
15. Los sistemas VapBC27, VapBC28 de *Mycobacterium llatzerense* MG 13^T y épsilon/zeta de *Mycobacterium* sp. MHSD3; en base a secuencia muestran a priori todos los elementos necesarios para actuar como un sistema toxina-antitoxina. Además, en el ensayo experimental manifiestan un comportamiento característico de este tipo de operones, si bien es cierto que el efecto atenuante de la toxina VapB28 no ha podido ser demostrado con rotundidad.
16. El sistema épsilon/zeta, junto con los genes homólogos presentes en algunas cepas cercanas a *Mycobacterium chelonae*, representa un sistema muy diferente al resto de homólogos detectados en otras especies de micobacterias.
17. Los hipotéticos sistemas toxina-antitoxina MT0933-34, con una lipasa intercalada entre la toxina y la antitoxina, parecen más bien pertenecer a sistemas de síntesis de poliquétidos y, por lo tanto, no corresponderían a toxinas reales. En cualquier caso, este último echo apuntaría a las micobacterias de crecimiento rápido como potenciales fuentes naturales de este tipo de compuestos.

9. Bibliografía

1. Euzéby JP. LPSN - List of Prokaryotic names with Standing in Nomenclature. 1997-2017. Available: http://www.bacterio.net/
2. Madigan M, Martinko J, Stahl D, Clark D. Brock Biology of Microorganisms (13th Edition).San Francisco. Pearson Education; 2010.
3. Cook G, Berney M, Gebhard S. Physiology of mycobacteria. Adv Microb Physiol. 2009;55: 81–182. doi:10.1016/S0065-2911(09)05502-7
4. Beran V, Havelkova M, Kaustova, J Dvorska, L Pavlik I. Cell wall deficient forms of mycobacteria: a review. Vet Med. 2006;2006: 365–389. Available: http://vri.cz/docs/vetmed/51-7-365.pdf
5. Katoch VM. Infections due to non-tuberculous mycobacteria (NTM). Indian J Med Res. 2004;120: 290–304. doi:10.1016/j.phrs.2009.12.004
6. Hogben AJ. A historical portrait of tuberculosis. Lancet Infect Dis. 2013;13: 1020. doi:10.1016/S1473-3099(13)70354-2
7. WHO | Tuberculosis. WHO. World Health Organization; 2017; Available: http://www.who.int/mediacentre/factsheets/fs104/en/
8. Daniel TM. The history of tuberculosis. Respir Med. 2006;100: 1862–1870. doi:10.1016/j.rmed.2006.08.006
9. Ollinger J, O'Malley T, Ahn J, Odingo J, Parish T. Inhibition of the sole type I signal peptidase of *Mycobacterium tuberculosis* is bactericidal under replicating and nonreplicating conditions. J Bacteriol. 2012;194: 2614–2619. doi:10.1128/JB.00224-12
10. Mirsaeidi M, Machado RF, Garcia JGN, Schraufnagel DE. Nontuberculous mycobacterial disease mortality in the United States, 1999-2010: a population-based comparative study. PLoS One. 2014;9: e91879. doi:10.1371/journal.pone.0091879
11. De Groote M a, Huitt G. Infections due to rapidly growing mycobacteria. Clin Infect Dis. 2006;42: 1756–1763. doi:10.1086/504381
12. Ruíz-Aragon J, García-Agudo L, Flores S, Rodríguez MJ, Marín P, García-Martos P. Sensibilidad a los antimicrobianos de micobacterias de crecimiento rápido. Rev Esp Quimioterap. 2007;20: 429–432.
13. Lee I. Etymologia: *Mycobacterium abscessus* subsp. *bolletii*. Emerg Infect Dis. 2014;20(3):379. https://dx.doi.org/10.3201/eid2003.ET2003

14. Sala A, Calderon V, Bordes P, Genevaux P. TAC from *Mycobacterium tuberculosis*: a paradigm for stress-responsive toxin-antitoxin systems controlled by SecB-like chaperones. Cell Stress Chaperones. 2013;18: 129–35. doi:10.1007/s12192-012-0396-5

15. Wilson RW, Steingrube V a, Böttger EC, Springer B, Brown-Elliott B a, Vincent V, et al. *Mycobacterium immunogenum* sp. nov., a novel species related to *Mycobacterium abscessus* and associated with clinical disease, pseudo-outbreaks and contaminated metalworking fluids: an international cooperative study on mycobacterial taxonomy. Int J Syst Evol Microbiol. 2001;51: 1751–1764. doi:10.1099/00207713-51-5-1751

16. Set R, Rokade S, Agrawal S, Shastri J. Antimicrobial susceptibility testing of rapidly growing mycobacteria by microdilution - experience of a tertiary care centre. Indian J Med Microbiol. 2010;28: 48–50. doi:10.4103/0255-0857.58729

17. Svetlíková Z, Skovierová H, Niederweis M, Gaillard J-L, McDonnell G, Jackson M. Role of porins in the susceptibility of *Mycobacterium smegmatis* and *Mycobacterium chelonae* to aldehyde-based disinfectants and drugs. Antimicrob Agents Chemother. 2009;53: 4015–8. doi:10.1128/AAC.00590-09

18. Chua KYL, Bustamante A, Jelfs P, Chen SC-A, Sintchenko V. Antibiotic susceptibility of diverse *Mycobacterium abscessus* complex strains in New South Wales, Australia. Pathology. England; 2015;47: 678–682. doi:10.1097/PAT.0000000000000327

19. Kennedy BS, Bedard B, Younge M, Tuttle D, Ammerman E, Ricci J, et al. Outbreak of *Mycobacterium chelonae* infection associated with tattoo ink. N Engl J Med. 2012;367: 1020–4. doi:10.1056/NEJMoa1205114

20. Sergeant a, Conaglen P, Laurenson IF, Claxton P, Mathers ME, Kavanagh GM, et al. *Mycobacterium chelonae* infection: a complication of tattooing. Clin Exp Dermatol. 2013;38: 140–2. doi:10.1111/j.1365-2230.2012.04421.x

21. Wallace R, Brown B, Onyi G. Skin, soft tissue, and bone infections due to *Mycobacterium chelonae chelonae*: importance of prior corticosteroid therapy, frequency of disseminated infections, and resistance to oral antimicrobials other than clarithromycin. J Infect desease. 1992;166: 405–412. Available: http://jid.oxfordjournals.org/content/166/2/405.short

22. Song Y, Wu J, Yan H, Chen J. Peritoneal dialysis-associated nontuberculous *Mycobacterium* peritonitis: a systematic review of reported cases. Nephrol Dial Transplant. 2012;27: 1639–44. doi:10.1093/ndt/gfr504
23. Eid AJ, Berbari EF, Sia IG, Wengenack NL, Osmon DR, Razonable RR. Prosthetic joint infection due to rapidly growing mycobacteria: report of 8 cases and review of the literature. Clin Infect Dis. 2007;45: 687–94. doi:10.1086/520982
24. Lee RP, Cheung KW, Chiu KH, Tsang ML. *Mycobacterium chelonae* infection after total knee arthroplasty: a case report. J Orthop Surg (Hong Kong). 2012;20: 134–6. doi: 10.1177/230949901202000130
25. Chbeir E, Casas L, Toubia N, Tawk M, Brown B. Adult cystic fibrosis presenting with recurrent non-tuberculous mycobacterial infections. Lancet. 2006;367: 2006. doi: 10.1016/S0140-6736(06)68851-X
26. Scola B La, Raoult D, Drancourt M. Amoebal coculture of "*Mycobacterium massiliense*" sp. nov. from the sputum of a patient with hemoptoic pneumonia. J Clin Microbiol. 2004;42: 5493–5501. doi:10.1128/JCM.42.12.5493
27. Leao SC, Tortoli E, Viana-Niero C, Ueki SYM, Lima KVB, Lopes ML, et al. Characterization of mycobacteria from a major brazilian outbreak suggests that revision of the taxonomic status of members of the *Mycobacterium chelonae-M. abscessus* group is needed. J Clin Microbiol. 2009;47: 2691–2698. doi:10.1128/JCM.00808-09
28. Adékambi T, Drancourt M. Dissection of phylogenetic relationships among 19 rapidly growing *Mycobacterium* species by 16S rRNA, hsp65, sodA, recA and rpoB gene sequencing. Int J Syst Evol Microbiol. 2004;54: 2095–2105. doi:10.1099/ijs.0.63094-0
29. Lee M-R, Yang C-Y, Chang K-P, Keng L-T, Yen DH-T, Wang J-Y, et al. Factors associated with lung function decline in patients with non-tuberculous mycobacterial pulmonary disease. PLoS One. 2013;8: e58214. doi:10.1371/journal.pone.0058214
30. Griffith DE, Aksamit T, Brown-Elliott BA, Catanzaro A, Daley C, Gordin F, et al. An official ATS/IDSA statement: Diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. Am J Respir Crit Care Med. 2007;175:

367–416. doi:10.1164/rccm.200604-571ST

31. Lee MR, Sheng WH, Hung CC, Yu CJ, Lee LN, Hsueh PR. *Mycobacterium abscessus* complex infections in humans. Emerg Infect Dis. 2015;21: 1638–1646. doi:10.3201/eid2109.141634
32. Nakanaga K, Hoshino Y, Era Y, Matsumoto K, Kanazawa Y, Tomita A, et al. Multiple cases of cutaneous *Mycobacterium massiliense* infection in a "hot spa" in Japan. J Clin Microbiol. 2011;49: 613–617. doi:10.1128/JCM.00817-10
33. Kothavade RJ, Dhurat RS, Mishra SN, Kothavade UR. Clinical and laboratory aspects of the diagnosis and management of cutaneous and subcutaneous infections caused by rapidly growing mycobacteria. Eur J Clin Microbiol Infect Dis. 2013;32: 161–188. doi:10.1007/s10096-012-1766-8
34. Moorthy RS, Valluri S, Rao NA. Nontuberculous mycobacterial ocular and adnexal infections. Surv Ophthalmol. 2012;57: 202–235. doi:10.1016/j.survophthal.2011.10.006
35. Girgis DO, Karp CL, Miller D. Ocular infections caused by non-tuberculous mycobacteria: update on epidemiology and management. Clin Experiment Ophthalmol. 2012;40: 467–475. doi:10.1111/j.1442-9071.2011.02679.x
36. Biggs HM, Chudgar SM, Pfeiffer CD, Rice KR, Zaas a. K, Wolfe CR. Disseminated *Mycobacterium immunogenum* infection presenting with septic shock and skin lesions in a renal transplant recipient. Transpl Infect Dis. 2012;14: 415–421. doi:10.1111/j.1399-3062.2012.00730.x
37. Sampaio JLM, Nassar D, De Freitas D, Höfling-Lima AL, Miyashiro K, Lopes Alberto F, et al. An outbreak of keratitis caused by *Mycobacterium immunogenum*. J Clin Microbiol. 2006;44: 3201–3207. doi:10.1128/JCM.00656-06
38. Shedd IV AD, Edhegard KD, Lugo-Somolinos A. *Mycobacterium immunogenum* skin infections: Two different presentations. Int J Dermatol. 2010;49: 941–944. doi:10.1111/j.1365-4632.2009.04363.x
39. del Castillo M, Palmero DJ, Palmero D, Lopez B, Paul R, Ritacco B, et al. Mesotherapy-associated outbreak caused by *Mycobacterium immunogenum*. Emerg Infect Dis. 2009;15: 357–359. doi:10.1136/adc.87.3.202
40. Wallace RJ, Zhang Y, Wilson RW, Mann L, Rossmoore H. Presence of a single

genotype of the newly described species *Mycobacterium immunogenum* in industrial metalworking fluids associated with hypersensitivity pneumonitis. Appl Environ Microbiol. 2002;68: 5580–5584. doi:10.1128/AEM.68.11.5580-5584.2002

41. Carson LA, Petersen NJ, Favero MS, Aguero SM. Growth characteristics of atypical mycobacteria in water and their comparative resistance to disinfectants. Appl Environ Microbiol. 1978;36: 839–846.
42. Thomson R, Tolson C, Carter R, Coulter C, Huygens F, Hargreaves M. Isolation of nontuberculous mycobacteria (NTM) from household water and shower aerosols in patients with pulmonary disease caused by NTM. J Clin Microbiol. 2013;51: 3006–11. doi:10.1128/JCM.00899-13
43. Gomila M, Ramirez A, Lalucat J. Diversity of environmental *Mycobacterium* isolates from hemodialysis water as shown by a multigene sequencing approach. Appl Environ Microbiol. 2007;73: 3787–97. doi:10.1128/AEM.02934-06
44. Hussein Z, Landt O, Wirths B, Wellinghausen N. Detection of non-tuberculous mycobacteria in hospital water by culture and molecular methods. Int J Med Microbiol. 2009;299: 281–90. doi:10.1016/j.ijmm.2008.07.004
45. Hall-Stoodley L, Lappin-Scott H. Biofilm formation by the rapidly growing mycobacterial species *Mycobacterium fortuitum*. FEMS Microbiol Lett. 1998;168: 77–84.
46. Hall-Stoodley L, Stoodley P. Biofilm formation and dispersal and the transmission of human pathogens. Trends Microbiol. 2005;13: 7–10. doi:10.1016/j.tim.2004.11.004
47. Schulze-Robbecke R, Buchholtz K. Heat susceptibility of aquatic mycobacteria. Appl Environ Microbiol. United States; 1992;58: 1869–1873.
48. Maxam a M, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74: 560–4. doi:10.1073/pnas.74.2.560
49. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94: 441–448. doi:http://dx.doi.org/10.1016/0022-2836(75)90213-2
50. Mardis ER. Next-Generation Sequencing Platforms. Annu Rev Anal Chem. 2013;6: 287–303. doi:10.1146/annurev-anchem-062012-092628

51. McCarthy A. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. Chem Biol. 2010;17: 675–676. doi:10.1016/j.chembiol.2010.07.004

52. Mokrousov I, Chernyaeva E, Vyazovaya A, Sinkov V, Zhuravlev V, Narvskaya O. Next-Generation Sequencing of *Mycobacterium tuberculosis*. Emerg Infect Dis. 2016;22: 1127–1129. doi:10.3201/eid2206.152051

53. Henriques Normark B, Normark S. Evolution and spread of antibiotic resistance. J Intern Med. 2002;252: 91–106. doi:10.1046/j.1365-2796.2002.01026.x

54. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 2005; 175–185. doi:10.1101/gr.8.3.175

55. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 2005; 175–185. doi:10.1101/gr.8.3.175

56. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). 2011.

57. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008;24: 2818–2824. doi:10.1093/bioinformatics/btn548

58. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;30: 31–7. doi:10.1093/bioinformatics/btt310

59. Powell DR, Seemann T. VAGUE: A graphical user interface for the Velvet assembler. Bioinformatics. 2013;29: 264–265. doi:10.1093/bioinformatics/bts664

60. Simpson JT, Wong K, Jackman SD, Shein JE, Jones SJ, Birol I. ABySS : A parallel assembler for short read sequence data. 2009; 1117–1123. doi:10.1101/gr.089532.108

61. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Meth. 2013;10: 563–569. doi: 10.1038/nmeth.2474

62. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13: 238. doi:10.1186/1471-2105-13-238

63. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18: 821–829. doi:10.1101/gr.074492.107
64. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-Cell sequencing. J Comput Biol. 2012;19: 455–477. doi:10.1089/cmb.2012.0021
65. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 2014;15: 211. doi:10.1186/1471-2105-15-211
66. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: Algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009;25: 1968–1969. doi:10.1093/bioinformatics/btp347
67. Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD. A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat Protoc. 2012;7: 1260–1284. doi:10.1038/nprot.2012.068
68. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106: 19126–31. doi:10.1073/pnas.0906412106
69. Boetzer M, Pirovano W, Zerbino D, Birney E, Simpson J, Wong K, et al. Toward almost closed genomes with GapFiller. Genome Biol. 2012;13: R56. doi:10.1186/gb-2012-13-6-r56
70. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, et al. Tablet-next generation sequence assembly visualization. Bioinformatics. 2009;26: 401–402. doi:10.1093/bioinformatics/btp666
71. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013;14: R47. doi:10.1186/gb-2013-14-5-r47
72. Vezzi F, Narzisi G, Mishra B. Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. PLoS One. 2012;7: 1–11. doi:10.1371/journal.pone.0052210
73. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29: 1072–1075. doi:10.1093/bioinformatics/btt086

74. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, et al. Update on RefSeq microbial genomes resources. Nucleic Acids Res. 2015;43: D599–D605. doi:10.1093/nar/gku1062

75. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. England; 2014;30: 2068–2069. doi:10.1093/bioinformatics/btu153

76. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013;41: 36–42. doi:10.1093/nar/gks1195

77. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 2005;102: 13950–5. doi:10.1073/pnas.0506758102

78. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F. The bacterial pan-genome: A new paradigm in microbiology. Int Microbiol. 2010;13: 45–57. doi:10.2436/20.1501.01.110

79. Tatusov RL, Koonin E V, Lipman DJ. A genomic perspective on protein families. Science. 1997;278: 631–637.

80. Vesth T, Lagesen K, Acar Ö, Ussery D. CMG-Biotools, a free workbench for basic comparative microbial genomics. PLoS One. 2013;8. doi:10.1371/journal.pone.0060120

81. Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23: 2947–2948. doi:10.1093/bioinformatics/btm404

82. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and mehtods to estimate maximum-likelihood phylogenies: asessing the performance of PhyML 3.0. Syst Biol. 2010;59: 307–321. doi: 10.1093/sysbio/syq010

83. Pearson WR. An Introduction to sequence similarity ("Homology") searching. Int J Res. 2014;1: 1286–1292. doi:10.1002/0471250953.bi0301s42.An

84. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

85. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene

clusters to infer functional coupling. Proc Natl Acad Sci U S A. 1999;96: 2896–2901. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC15866/

86. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin E V, et al. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. 2010;26: 1481–1487. doi:10.1093/bioinformatics/btq229
87. Li L, Jr CJS, Roos DS. OrthoMCL : Identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 2178–2189. doi:10.1101/gr.1224503.candidates
88. Contreras-moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. 2013;79: 7696–7701. doi:10.1128/AEM.02411-13
89. Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW. Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. Genome Biol. 2007;8: R267. doi:10.1186/gb-2007-8-12-r267
90. Kaas RS, Friis C, Ussery DW, Aarestrup FM, Otto T, Oryan M, et al. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. BMC Genomics. 2012;13: 577. doi:10.1186/1471-2164-13-577
91. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7: 539. doi:10.1038/msb.2011.75
92. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17: 540–552. doi:10.1093/oxfordjournals.molbev.a026334
93. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches : A fast , accurate and powerful alternative. Syst Biol. 2006;55: 539–552. doi: 10.1080/10635150600755453
94. Galperin MY, Makarova KS, Wolf YI, Koonin E V. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43: D261-9. doi:10.1093/nar/gku1223
95. Greninger AL, Langelier C, Cunningham G, Keh C, Melgar M, Chiu CY. Two

rapidly growing mycobacterial species Isolated from a brain abscess : first whole-genome sequences of *Mycobacterium immunogenum* and *Mycobacterium llatzerense*. 2015;53: 2374–2377. doi:10.1128/JCM.00402-15

96. Buist G, Steen A, Kok J, Kuipers OP. LysM, a widely distributed protein motif for binding to (peptido)glycans. Mol Microbiol. 2008;68: 838–847. doi:10.1111/j.1365-2958.2008.06211.x
97. Větrovský T, Baldrian P. The Variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One. 2013;8: 1–10. doi:10.1371/journal.pone.0057923
98. Falkinham JO 3rd. Nontuberculous mycobacteria in the environment. Clin Chest Med. 2002;23: 529–551.
99. Gomila M, Ramirez A, Gascó J, Lalucat J. *Mycobacterium llatzerense* sp. nov., a facultatively autotrophic, hydrogen-oxidizing bacterium isolated from haemodialysis water. Int J Syst Evol Microbiol. 2008;58: 2769–73. doi:10.1099/ijs.0.65857-0
100. Dröge M, Pühler A, Selbitschka W. Horizontal gene transfer among bacteria in terrestrial and aquatic habitats as assessed by microcosm and field studies. Biol Fertil Soils 1999; 221–245.
101. Cariani L, Vasireddy S, Wallace Jr. RJ, Cardoso Leao S, Teri A, Turenne CY, et al. Emended description of *Mycobacterium abscessus, Mycobacterium abscessus* subs. abscessus, *Mycobacterium abscessus* subsp. *bolletii* and designation of *Mycobacterium abscessus* subsp. *massiliense* comb. nov. Int J Syst Evol Microbiol. 2016; 4471–4479. doi:10.1099/ijsem.0.001376
102. Smith I. *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. Clin Microbiol Rev. 2003;16: 463–496. doi:10.1128/CMR.16.3.463
103. Periwal V, Patowary A, Vellarikkal SK, Gupta A. Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of *Mycobacterium tuberculosis* pangenome. 2015; 1–26. doi:10.1371/journal.pone.0122979
104. Narayanan S, Deshpande U. Whole-genome sequences of four clinical isolates of *Mycobacterium tuberculosis* from Tamil Nadu , South India. Genome Announc. 2013;1: 4430. doi:10.1128/genomeA.00186-13.Copyright
105. Sarkar R, Lenders L, Wilkinson KA, Wilkinson RJ, Nicol MP. Modern lineages

of *Mycobacterium tuberculosis* exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages macrophages. PLoS One. 2012;7: 6–13. doi:10.1371/journal.pone.0043170

106. Brown-Elliott B, Wallace R. Clinical and taxonomic status of pathogenic nonpigmented or late-pigmenting rapidly growing mycobacteria. Clin Microbiol Rev. 2002;15: 716–746. doi:10.1128/CMR.15.4.716
107. McArthur AG, Waglechner N, Nizam F, Yan A, Azad M a, Baylay AJ, et al. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother. 2013;57: 3348–57. doi:10.1128/AAC.00419-13
108. Yong D, Lee K, Yum JH, Shin HB, Rossolini GM, Chong Y. Imipenem-EDTA disk method for differentiation of metallo-β-Lactamase-producing clinical isolates of *Pseudomonas* spp. and *Acinetobacter* spp. J Clin Microbiol. 2002;40: 3798–3801. doi:10.1128/JCM.40.10.3798
109. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33: D325-8. doi:10.1093/nar/gki008
110. Okonechnikov K, Golosova O, Fursov M. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28: 1166–7. doi:10.1093/bioinformatics/bts091
111. Barakat M, Ortet P, Whitworth DE. P2RP: a Web-based framework for the identification and analysis of regulatory proteins in prokaryotic genomes. BMC Genomics. 2013;14: 269. doi:10.1186/1471-2164-14-269
112. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H-R, et al. The Pfam protein families database. Nucleic Acids Res. 2008;36: D281-8. doi:10.1093/nar/gkm960
113. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. 2000;28: 231–4. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102444/
114. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F, et al. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res. 2015;43: W104–W108. doi:10.1093/nar/gkv401
115. Langille MG, Brinkman FS. IslandViewer: an integrated interface for

computational identification and visualization of genomic islands. Bioinformatics. 2009;25: 664–5. doi:10.1093/bioinformatics/btp030

116. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: A Fast Phage Search Tool. Nucleic Acids Res. 2011;39: 347–352. doi:10.1093/nar/gkr485
117. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol. 2016;428: 726–731. doi:10.1016/j.jmb.2015.11.006
118. Snel B, Lehmann G, Bork P, Huynen MA. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. 2000;28: 3442–4. doi:10.1093/nar/28.18.3442
119. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43: D447–D452. doi:10.1093/nar/gku1003
120. Maurer FP, Castelberg C, Quiblier C, Böttger EC, Somoskövi A. Erm(41)-dependent inducible resistance to azithromycin and clarithromycin in clinical isolates of *Mycobacterium abscessus*. J Antimicrob Chemother. 2014;69: 1559–1563. doi:10.1093/jac/dku007
121. Rodriguez GM, Smith I. Mechanisms of iron regulation in mycobacteria: role in physiology and virulence. Mol Microbiol. 2003;47: 1485–1494. doi:10.1046/j.1365-2958.2003.03384.x
122. Soncini FC, Véscovi EG, Solomon F, Groisman EA. Molecular basis of the magnesium deprivation response in *Salmonella typhimurium*: Identification of PhoP-Regulated genes. J Bacteriol. 1996;178: 5092–5099.
123. Teimourpour R, Sadeghian A, Meshkat Z, Esmaelizad M, Sankian M, Jabbari AR. Construction of a DNA vaccine encoding Mtb32C and HBHA genes of *Mycobacterium tuberculosis*. Jundishapur J Microbiol. 2015;8. doi:10.5812/jjm.21556
124. Richter L, Saviola B. The lipF promoter of *Mycobacterium tuberculosis* is upregulated specifically by acidic pH but not by other stress conditions. Microbiol Res. 2010;164: 228–232. doi:10.1016/j.micres.2007.06.003.The
125. Bandyopadhyay P, Steinman HM. Catalase-peroxidases of *Legionella pneumophila*: Cloning of the katA gene and studies of KatA function. J Bacteriol.

2000;182: 6679–6686. doi:10.1128/JB.182.23.6679-6686.2000

126. Primm TP, Andersen SJ, Mizrahi V, Avarbock D, Rubin H, Barry III CE. The stringent response of *Mycobacterium tuberculosis* is required for long-term survival. J Bacteriol. 2000;182: 4889–4898. doi:10.1128/JB.182.17.4889-4898.2000
127. McKinney JD, Honer zu Bentrup K, Munoz-Elias EJ, Miczak A, Chen B, Chan WT, et al. Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. Nature. 2000;406: 735–738. doi:10.1038/35021074
128. Tzeng YL, Swartley JS, Miller YK, Nisbet RE, Liu LJ, Ahn JH, et al. Transcriptional regulation of divergent capsule biosynthesis and transport operon promoters in serogroup B *Neisseria meningitidis*. Infect Immun. 2001;69: 2502–2511. doi:10.1128/IAI.69.4.2502-2511.2001
129. Yendapally R, Lee RE. Design, synthesis, and evaluation of novel ethambutol analogues. Bioorg Med Chem Lett. 2008;18: 1607–11. doi:10.1016/j.bmcl.2008.01.065
130. Nahid P, Dorman SE, Alipanah N, Barry PM, Brozek JL, Cattamanchi A, et al. Executive Summary: Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis. Clin Infect Dis. 2016;63: 853–867. doi:10.1093/cid/ciw566
131. Greninger AL, Cunningham G, Yu JM, Hsu ED, Chiu CY, Miller S. Draft Genome Sequence of *Mycobacterium obuense* Strain UC1, Isolated from Patient Sputum. Genome Announc. 2015;3: e00691-15. doi:10.1128/genomeA.00691-15
132. Greninger A, Cunningham G, Yu J, Hsu E, Chiu C, Miller S. Draft Genome Sequence of *Mycobacterium arupense* Strain GUC1. 2015;3: 2012–2013. doi:10.1128/genomeA.00630-15
133. Houben ENG, Korotkov K V., Bitter W. Take five - Type VII secretion systems of mycobacteria. Biochim Biophys Acta - Mol Cell Res. 2014;1843: 1707–1716. doi:10.1016/j.bbamcr.2013.11.003
134. Siegrist MS, Steigedal M, Ahmad R. Mycobacterial Esx-3 requires multiple components for iron acquisition. MBio. 2014;3: 1–10. doi:10.1128/mBio.01073-

14

135. Siegrist MS, Unnikrishnan M, McConnell MJ, Borowsky M, Cheng T-Y, Siddiqi N, et al. Mycobacterial Esx-3 is required for mycobactin-mediated iron acquisition. Proc Natl Acad Sci U S A.. 2009;106: 18792–18797. doi:10.1073/pnas.0900589106
136. Quadri LE, Sello J, Keating TA, Weinreb PH, Walsh CT. Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. Chem Biol. 1998;5: 631–645. doi:10.1016/S1074-5521(98)90291-5
137. Belisle JT, Vissa VD, Sievert T, Takayama K, Brennan PJ, Besra GS. Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. Science. 1997;276: 1420–1422. doi:10.1126/science.276.5317.1420
138. Puech V, Guilhot C, Perez E, Tropis M, Armitige LY, Gicquel B, et al. Evidence for a partial redundancy of the fibronectin-binding proteins for the transfer of mycoloyl residues onto the cell wall arabinogalactan termini of *Mycobacterium tuberculosis*. Mol Microbiol. 2002;44: 1109–1122. doi:10.1046/j.1365-2958.2002.02953.x
139. Gebhardt H, Meniche X, Tropis M, Krämer R, Daffé M, Morbach S. The key role of the mycolic acid content in the functionality of the cell wall permeability barrier in *Corynebacterineae*. Microbiology. 2007;153: 1424–1434. doi:10.1099/mic.0.2006/003541-0
140. Mobley HL, Island MD, Hausinger RP. Molecular biology of microbial ureases. Microbiol Rev. 1995;59: 451–480. doi:10.2741/1350
141. Rutherford JC. The emerging role of urease as a general microbial virulence factor. PLoS Pathog. 2014;10: 1–3. doi:10.1371/journal.ppat.1004062
142. Mollenhauer-Rektorschek M, Hanauer G, Sachs G, Melchers K. Expression of UreI is required for intragastric transit and colonization of gerbil gastric mucosa by *Helicobacter pylori*. Res Microbiol. 2002;153: 659–666.
143. Marcus EA, Moshfegh AP, Sachs G, Scott DR. The periplasmic alpha-carbonic anhydrase activity of *Helicobacter pylori* is essential for acid acclimation. J Bacteriol. 2005;187: 729–738. doi:10.1128/JB.187.2.729-738.2005
144. Smoot DT, Mobley HL, Chippendale GR, Lewison JF, Resau JH. *Helicobacter*

pylori urease activity is toxic to human gastric epithelial cells. Infect Immun. 1990;58: 1992–1994.

145. Parsons CL, Stauffer C, Mulholland SG, Griffith DP. Effect of ammonium on bacterial adherence to bladder transitional epithelium. J Urol. 1984;132: 365–366.
146. Musher DM, Griffith DP, Yawn D, Rossen RD. Role of urease in pyelonephritis resulting from urinary tract infection with *Proteus*. J Infect Dis. 1975;131: 177–181.
147. Camacho LR, Ensergueix D, Perez E, Gicquel B, Guilhot C. Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. Mol Microbiol. 1999;34: 257–267. doi:10.1046/j.1365-2958.1999.01593.x
148. Ripoll F, Pasek S, Schenowitz C, Dossat C, Barbe V, Rottman M, et al. Non mycobacterial virulence genes in the genome of the emerging pathogen *Mycobacterium abscessus.* PLoS One. 2009;4. doi:10.1371/journal.pone.0005660
149. Hassan HM, Fridovich I. Mechanism of the antibiotic action of pyocyanine. J Bacteriol. 1980;141: 156–163. doi:10.1126/science.105.2734.549
150. Pierson LS, Pierson EA. Metabolism and function of phenazines in bacteria: Impacts on the behavior of bacteria in the environment and biotechnological processes. Appl Microbiol Biotechnol. 2010;86: 1659–1670. doi:10.1007/s00253-010-2509-3
151. Pethe K, Alonso S, Biet F, Delogu G, Brennan MJ, Locht C, et al. The heparin-binding haemagglutinin of *M. tuberculosis* is required for extrapulmonary dissemination. Nature. 2001;412: 190–194. doi:10.1038/35084083
152. Reyrat JM, Kahn D. *Mycobacterium smegmatis*: an absurd model for tuberculosis?. Trends Microbiol. 2017;9: 472–473. doi:10.1016/S0966-842X(01)02168-0
153. Delogu G, Bua A, Pusceddu C, Parra M, Fadda G, Brennan MJ, et al. Expression and purification of recombinant methylated HBHA in *Mycobacterium smegmatis*. FEMS Microbiol Lett. 2004;239: 33–39. doi:10.1016/j.femsle.2004.08.015
154. Schmitz KR, Sauer RT. Substrate delivery by the AAA+ ClpX and ClpC1 unfoldases activates the mycobacterial ClpP1P2 peptidase. Mol Microbiol.

2014;93: 617–628. doi:10.1111/mmi.12694

155. Ollinger J, O'malley T, Kesicki EA, Odingo J, Parish T. Validation of the essential ClpP protease in *Mycobacterium tuberculosis* as a novel drug target. J Bacteriol. 2012;194: 663–668. doi:10.1128/JB.06142-11
156. Collins DM. In search of tuberculosis virulence genes. Trends Microbiol. 1996;4: 426–430. doi: http://dx.doi.org/10.1016/0966-842X(96)10066-4
157. Pym AS, Domenech P, Honore N, Song J, Deretic V, Cole ST. Regulation of catalase-peroxidase (KatG) expression, isoniazid sensitivity and virulence by furA of *Mycobacterium tuberculosis*. Mol Microbiol. 2001;40: 879–889.
158. Larue K, Ford RC, Willis LM, Whitfield C. Functional and structural characterization of polysaccharide co-polymerase proteins required for polymer export in ATP-binding cassette transporter-dependent capsule biosynthesis pathways. J Biol Chem. 2011;286: 16658–16668. doi:10.1074/jbc.M111.228221
159. Jarvis GA. Recognition and control of neisserial infection by antibody and complement. Trends Microbiol. 2017;3: 198–201. doi:10.1016/S0966-842X(00)88921-0
160. Jarvis GA, Vedros NA. Sialic acid of group B *Neisseria meningitidis* regulates alternative complement pathway activation. Infect Immun. 1987;55: 174–180.
161. Daffe M, Etienne G. The capsule of *Mycobacterium tuberculosis* and its implications for pathogenicity. Tuber Lung Dis. 1999;79: 153–169. doi:10.1054/tuld.1998.0200
162. Schwebach JR, Casadevall A, Schneerson R, Dai Z, Wang X, Robbins JB, et al. Expression of a *Mycobacterium tuberculosis* arabinomannan antigen in vitro and in vivo. Infect Immun. 2001;69: 5671–5678. doi:10.1128/IAI.69.9.5671-5678.2001
163. Schwebach JR, Glatman-freedman A, Dai Z, Robbins JB, Schneerson R, Casadevall A, et al. Glucan is a component of the *Mycobacterium tuberculosis* surface that is expressed in vitro and in vivo glucan is a component of the *Mycobacterium tuberculosis* surface that is expressed in vitro and in vivo. Infect Immun. 2002;70: 2566–2575. doi:10.1128/IAI.70.5.2566
164. Daffe M, Draper P. The envelope layers of mycobacteria with reference to their pathogenicity. Adv Microb Physiol. 1998;39: 131–203.

165. Jishage M, Kvint K, Shingler V, Nyström T. Regulation of σ factor competition by the alarmone ppGpp. Genes Dev. 2002;16: 1260–1270. doi:10.1101/gad.227902

166. Hoyt S, Jones GH. relA is required for actinomycin production in *Streptomyces antibioticus*. J Bacteriol. 1999;181: 3824–3829.

167. Chakraburtty R, Bibb M. The ppGpp synthetase gene (relA) of *Streptomyces coelicolor* A3(2) plays a conditional role in antibiotic production and morphological differentiation. J Bacteriol. 1997. pp. 5854–5861.

168. Martinez-Costa OH, Arias P, Romero NM, Parro V, Mellado RP, Malpartida F. A relA/spoT homologous gene from *Streptomyces coelicolor* A3(2) controls antibiotic biosynthetic genes. J Biol Chem. 1996;271: 10627–10634.

169. Mechold U, Malke H. Characterization of the stringent and relaxed responses of *Streptococcus equisimilis*. J Bacteriol. 1997;179: 2658–2667.

170. Mechold U, Cashel M, Steiner K, Gentry D, Malke H. Functional analysis of a relA/spoT gene homolog from *Streptococcus equisimilis*. J Bacteriol. 1996;178: 1401–1411.

171. Wehmeier L, Schafer A, Burkovski A, Kramer R, Mechold U, Malke H, et al. The role of the *Corynebacterium glutamicum* rel gene in (p)ppGpp metabolism. Microbiology. 1998;144 (Pt 7: 1853–1862. doi:10.1099/00221287-144-7-1853

172. Dahl JL, Kraus CN, Boshoff HIM, Doan B, Foley K, Avarbock D, et al. The role of RelMtb-mediated adaptation to stationary phase in long-term persistence of *Mycobacterium tuberculosis* in mice. Proc Natl Acad Sci U S A. 2003;100: 10026–31. doi:10.1073/pnas.1631248100

173. Dunn MF, Ramírez-Trujillo JA, Hernández-Lucas I. Major roles of isocitrate lyase and malate synthase in bacterial and fungal pathogenesis. Microbiology. 2009;155: 3166–3175. doi:10.1099/mic.0.030858-0

174. Wayne LG, Lin KY. Glyoxylate metabolism and adaptation of *Mycobacterium tuberculosis* to survival under anaerobic conditions. Infect Immun. 1982;37: 1042–1049.

175. Kinnear SM, Marques RR, Carbonetti NH. Differential regulation of Bvg-activated virulence factors plays a role in *Bordetella pertussis* pathogenicity. Infect Immun. 2001;69: 1983–1993. doi:10.1128/IAI.69.4.1983-1993.2001

176. Johnson CR, Newcombe J, Thorne S, Borde HA, Eales-Reynolds LJ, Gorringe AR, et al. Generation and characterization of a PhoP homologue mutant of *Neisseria meningitidis*. Mol Microbiol. 2001;39: 1345–1355.

177. Arruda S, Bomfim G, Knights R, Huima-Byron T, Riley LW. Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. Science. 1993;261: 1454–1457.

178. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998;393: 537–544. doi:10.1038/31159

179. Tekaia F, Gordon S V, Garnier T, Brosch R, Barrell BG, Cole ST. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. Tuber Lung Dis. 1999;79: 329–342. doi:10.1054/tuld.1999.0220

180. Casali N, Riley LW. A phylogenomic analysis of the *Actinomycetales* mce operons. BMC Genomics. 2007;8: 60. doi:10.1186/1471-2164-8-60

181. Høiby N, Bjarnsholt T, Givskov M, Molin S, Ciofu O. Antibiotic resistance of bacterial biofilms. Int J Antimicrob Agents. 2010;35: 322–332. doi:http://doi.org/10.1016/j.ijantimicag.2009.12.011

182. Cuthbertson L, Nodwell JR. The TetR Family of Regulators. Microbiol Mol Biol Rev. 2013;77: 440–475. doi:10.1128/MMBR.00018-13

183. Cai SJ, Inouye M. EnvZ-OmpR interaction and osmoregulation in *Escherichia coli*. J Biol Chem. 2002;277: 24155–24161. doi:10.1074/jbc.M110715200

184. Feng X, Oropeza R, Kenney LJ. Dual regulation by phospho-OmpR of ssrA/B gene expression in *Salmonella* pathogenicity island 2. Mol Microbiol. England; 2003;48: 1131–1143.

185. Feng X, Oropeza R, Walthers D, Kenney LJ. OmpR phosphorylation and its role in signaling and pathogenesis. ASM News. 2003;69: 390–395.

186. Sachdeva P, Misra R, Tyagi AK, Singh Y. The sigma factors of *Mycobacterium tuberculosis*: Regulation of the regulators. FEBS J. 2010;277: 605–626. doi:10.1111/j.1742-4658.2009.07479.x

187. Hu Y, Coates AR. Transcription of two sigma 70 homologue genes, sigA and sigB, in stationary-phase *Mycobacterium tuberculosis*. J Bacteriol. 1999;181: 469–476.

188. Manganelli R, Dubnau E, Tyagi S, Kramer FR, Smith I. Differential expression of 10 sigma factor genes in *Mycobacterium tuberculosis*. Mol Microbiol. 1999;31: 715–724.

189. Cappelli G, Volpe E, Grassi M, Liseo B, Colizzi V, Mariani F. Profiling of *Mycobacterium tuberculosis* gene expression during human macrophage infection: upregulation of the alternative sigma factor G, a group of transcriptional regulators, and proteins with unknown function. Res Microbiol. 2006;157: 445–455. doi:10.1016/j.resmic.2005.10.007

190. Volpe E, Cappelli G, Grassi M, Martino A, Serafino A, Colizzi V, et al. Gene expression profiling of human macrophages at late time of infection with *Mycobacterium tuberculosis*. Immunology. 2006;118: 449–460. doi:10.1111/j.1365-2567.2006.02378.x

191. Gicquel G, Bouffartigues E, Bains M, Oxaran V, Rosay T, Lesouhaitier O, et al. The extra-cytoplasmic function sigma factor SigX modulates biofilm and virulence-related properties in *Pseudomonas aeruginosa*. PLoS One. 2013;8: e80407. Available: https://doi.org/10.1371/journal.pone.0080407

192. Potvin E, Sanschagrin F, Levesque RC. Sigma factors in *Pseudomonas aeruginosa*. FEMS Microbiol Rev. 2008;32: 38–55. doi:10.1111/j.1574-6976.2007.00092.x

193. Choi Y, Park HY, Park SJ, Park SJ, Kim SK, Ha C, et al. Growth phase-differential quorum sensing regulation of anthranilate metabolism in *Pseudomonas aeruginosa*. Mol Cells. 2011;32: 57–65. doi:10.1007/s10059-011-2322-6

194. Lee CE, Goodfellow C, Javid-Majd F, Baker EN, Shaun Lott J. The crystal structure of TrpD, a metabolic enzyme essential for lung colonization by *Mycobacterium tuberculosis*, in complex with its substrate phosphoribosylpyrophosphate. J Mol Biol. 2006;355: 784–797. doi:http://doi.org/10.1016/j.jmb.2005.11.016

195. Baker TI, Crawford IP. Anthranilate synthetase. Partial purification and some kinetic studies on the enzyme from *Escherichia coli*. J Biol Chem. 1966;241: 5577–5584.

196. Ito J, Cox EC, Yanofsky C. Anthranilate synthetase, an enzyme specified by the

tryptophan operon of *Escherichia coli*: purification and characterization of component I. J Bacteriol. United States; 1969;97: 725–733.

197. Pabst MJ, Kuhn JC, Somerville RL. Feedback regulation in the anthranilate aggregate from wild type and mutant strains of *Escherichia coli*. J Biol Chem. United States; 1973;248: 901–914.
198. Chun H, Choi O, Goo E, Kim N, Kim H, Kang Y, et al. The quorum sensing-dependent gene katG of *Burkholderia glumae* is important for protection from visible light. J Bacteriol. 2009;191: 4152–4157. doi:10.1128/JB.00227-09
199. Forrellad MA, Klepp LI, Gioffré A, Sabio y García J, Morbidoni HR, de la Paz Santangelo M, et al. Virulence factors of the *Mycobacterium tuberculosis* complex. Virulence. 2013;4: 3–66. doi:10.4161/viru.22329
200. Danelishvili L, Stang B, Bermudez L. Identification of *Mycobacterium avium* genes expressed during in vivo infection and the role of the Oligopeptide transporter OppA in virulence Lia. Microb Pathog. 2014; 67–76. doi:10.1016/j.micpath.2014.09.010
201. Igarashi K, Kashiwagi K. Polyamine transport in bacteria and yeast. Biochem J. 1999;344 Pt 3: 633–642. doi:10.1042/0264-6021:3440633
202. Anderson JJ, Oxender DL. Genetic separation of high- and low-affinity transport systems for branched-chain amino acids in *Escherichia coli* K-12. J Bacteriol. 1978;136: 168–174.
203. Braunstein M, Brown AM, Kurtz S, Jacobs WR, Carolina N, Hill C, et al. Two nonredundant SecA homologues function in mycobacteria. J Bacteriol. 2001;183: 6979–6990. doi:10.1128/JB.183.24.6979
204. Thakur, Preeti ,Nagavara Gantashala prasad, Choudhary Eira,Singh Nirependra AZM and AN. The preprotein translocase YidC controls respiratory metabolism in *Mycobacterium tuberculosis*. Sci Rep. 2016;412115: 1–14. doi:10.1038/srep24998
205. Zhao L, Xue T, Shang F, Sun H, Sun B. *Staphylococcus aureus* AI-2 quorum sensing associates with the KdpDE two-component system to regulate capsular polysaccharide synthesis and virulence. Infect Immun. 2010;78: 3506–3515. doi:10.1128/IAI.00131-10
206. Xue T, You Y, Hong D, Sun H, Sun B. The *Staphylococcus aureus* KdpDE two-

component system couples extracellular K+ sensing and Agr signaling to infection programming. Infect Immun. 2011;79: 2154–2167. doi:10.1128/IAI.01180-10

207. Freeman ZN, Dorus S, Waterfield NR. The KdpD/KdpE Two-Component System: integrating K+ homeostasis and virulence. PLoS Pathog. 2013;9. doi:10.1371/journal.ppat.1003201
208. Hou, J. Y., Graham, J. E. and Clark-Curtiss JE. *Mycobacterium avium* genes expressed during growth in human macrophages detected by selective capture of transcribed sequences (SCOTS). Infect Immun. 2002;70: 3714–3726. doi:10.1128/IAI.70.7.3714
209. Ban N. The structural basis of FtsY recruitment and GTPase activation by SRP RNA. 2014;52: 643–654. doi:10.1016/j.molcel.2013.10.005
210. Lammertyn E, Van Mellaert L, Meyen E, Lebeau I, De Buck E, Anné J, et al. Molecular and functional characterization of type I signal peptidase from *Legionella pneumophila*. Microbiology. 2004;150: 1475–1483. doi:10.1099/mic.0.26973-0
211. Capitani G, De Biase D, Aurizi C, Gut H, Bossa F, Grutter MG. Crystal structure and functional analysis of *Escherichia coli* glutamate decarboxylase. Embo J. 2003;22: 4027–4037. doi:10.1093/emboj/cdg403
212. Cotter PD, Hill C. Surviving the Acid Test: Responses of Gram-positive bacteria to low pH. Food Technol Biotechnol. 2010;48: 296–307. doi:10.1128/MMBR.67.3.429
213. Prabhakaran K, Harris EB, Kirchheimer WF. Glutamic acid decarboxylase in *Mycobacterium leprae.* Arch Microbiol. Germany; 1983;134: 320–323.
214. Maurizi MR. Proteases and protein degradation in *Escherichia coli*. Experientia. 1992;48: 178–201.
215. Laskowska E, Kuczynska-Wisnik D, Skorko-Glonek J, Taylor A. Degradation by proteases Lon, Clp and HtrA, of *Escherichia coli* proteins aggregated in vivo by heat shock; HtrA protease action in vivo and in vitro. Mol Microbiol. 1996;22: 555–571.
216. Takaya A, Tabuchi F, Tsuchiya H, Isogai E, Yamamoto T. Negative regulation of quorum-sensing systems in *Pseudomonas aeruginosa* by ATP-dependent Lon

protease. J Bacteriol. 2008;190: 4181–4188. doi:10.1128/JB.01873-07

217. Van Melderen L, Thi MH, Lecchi P, Gottesman S, Couturier M, Maurizi MR. ATP-dependent degradation of CcdA by Lon protease. Effects of secondary structure and heterologous subunit interactions. J Biol Chem. 1996;271: 27730–27738.

218. Sala A, Calderon V, Bordes P, Genevaux P. TAC from M*ycobacterium tuberculosis*: a paradigm for stress-responsive toxin-antitoxin systems controlled by SecB-like chaperones. Cell Stress Chaperones. 2013;18: 129–35. doi:10.1007/s12192-012-0396-5

219. Yamaguchi Y, Park J-H, Inouye M. Toxin-antitoxin systems in bacteria and archaea. Annu Rev Genet. 2011;45: 61–79. doi:10.1146/annurev-genet-110410-132412

220. Fozo EM, Hemm MR, Storz G. Small toxic proteins and the antisense RNAs that repress them. Microbiol Mol Biol Rev. 2008;72: 579–89, Table of Contents. doi:10.1128/MMBR.00025-08

221. Unterholzner S. Toxin–antitoxin systems: Biology, identification, and application. Mob Genet Elements. 2013;3: 1–13. doi:10.4161/mge.26219

222. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010;8: 317–327. doi:10.1038/nrmicro2315

223. Masuda H, Tan Q, Awano N, Wu KP, Inouye M. YeeU enhances the bundling of cytoskeletal polymers of MreB and FtsZ, antagonizing the CbtA (YeeV) toxicity in *Escherichia coli*. Mol Microbiol. 2012;84: 979–989. doi:10.1111/j.1365-2958.2012.08068.x

224. Wang X, Lord DM, Cheng H-Y, Osbourne DO, Hong SH, Sanchez-Torres V, et al. A new type V toxin-antitoxin system where mRNA for toxin GhoT is cleaved by antitoxin GhoS. Nat Chem Biol. 2012;8: 855–861. doi:10.1038/nchembio.1062

225. Hanna Engelberg-Kulka and Gad Glaser. Addiction modules and programmed cell death and antideath in bacterial cultures. Annu Rev Microbiol. 1999;53: 43–70. doi:10.1227/01.NEU.0000143034.62913.59

226. Gerdes K. Toxin-antitoxin modules may regulate synthesis of macromolecules during nutritional stress. 2000;182: 561–572. Available:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC94316/

227. Yamaguchi Y, Inouye M. Regulation of growth and death in *Escherichia coli* by toxin-antitoxin systems. Nat Rev Microbiol. 2011;9: 779–790. doi:10.1038/nrmicro2651

228. Engelberg-Kulka H, Amitai S, Kolodkin-Gal I, Hazan R. Bacterial programmed cell death and multicellular behavior in bacteria. PLoS Genet. 2006;2: 1518–1526. doi:10.1371/journal.pgen.0020135

229. Sevin EW, Barloy-Hubler F. RASTA-Bacteria: a web-based tool for identifying toxin-antitoxin loci in prokaryotes. Genome Biol. 2007;8: R155. doi: 10.1186/gb-2007-8-8-r155

230. Shao Y, Harrison EM, Bi D, Tai C, He X, Ou HY, et al. TADB: A web-based resource for Type 2 toxin-antitoxin loci in bacteria and archaea. Nucleic Acids Res. 2011;39: 606–611. doi:10.1093/nar/gkq908

231. Sberro H, Leavitt A, Kiro R, Koh E, Peleg Y, Qimron U, et al. Discovery of functional toxin/antitoxin systems in bacteria by shotgun cloning. Mol Cell. 2013;50: 136–148. doi:10.1016/j.molcel.2013.02.002

232. Angers-Loustau A, Rainy J, Wartiovaara K, Dong X, Stothard P, Forsythe I, et al. PlasmaDNA: a free, cross-platform plasmid manipulation program for molecular biology laboratories. BMC Mol Biol. 2007;8: 77. doi:10.1186/1471-2199-8-77

233. Kalendar R, Lee D, Schulman AH. FastPCR Software for PCR, In Silico PCR, and oligonucleotide assembly and analysis. Methods Mol Biol. 2014. pp. 271–302. doi:10.1007/978-1-62703-764-8_18

234. Consortium TU. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014;42: D191-8. doi:10.1093/nar/gkt1140

235. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2011;5: 725–738. doi:10.1038/nprot.2010.5.I-TASSER

236. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004;57: 702–710. doi:10.1002/prot.20264

237. Li Z, Natarajan P, Ye Y, Hrabe T, Godzik A. POSA: A user-driven, interactive multiple protein structure alignment server. Nucleic Acids Res. 2014;42: 240–

245. doi:10.1093/nar/gku394

238. Ramage HR, Connolly LE, Cox JS. Comprehensive functional analysis of *Mycobacterium tuberculosis* Toxin-Antitoxin Systems: implications for pathogenesis, stress responses, and evolution. PLoS Genet. 2009;5: e1000767. doi:10.1371/journal.pgen.1000767

239. Bordes P, Sala AJ, Ayala S, Texier P, Slama N, Cirinesi A-M, et al. Chaperone addiction of toxin–antitoxin systems. Nat Commun. 2016;7: 13339. doi:10.1038/ncomms13339

240. Dienemann C, Bøggild A, Winther KS, Gerdes K, Brodersen DE. Crystal structure of the VapBC toxin-antitoxin complex from *Shigella flexneri* reveals a hetero-octameric DNA-binding assembly. J Mol Biol. 2011;414: 713–722. doi:10.1016/j.jmb.2011.10.024

241. Lee I-G, Lee SJ, Chae S, Lee K-Y, Kim J-H, Lee B-J. Structural and functional studies of the *Mycobacterium tuberculosis* VapBC30 toxin-antitoxin system: implications for the design of novel antimicrobial peptides. Nucleic Acids Res. 2015;43: 7624–37. doi:10.1093/nar/gkv689

242. Zielenkiewicz U, Cegłowski P. The toxin-antitoxin system of the streptococcal plasmid pSM19035. J Bacteriol. 2005;187: 6094–6105. doi:10.1128/JB.187.17.6094-6105.2005

243. Demidenok OI, Kaprelyants AS, Goncharenko A V. Toxin-antitoxin vapBC locus participates in formation of the dormant state in *Mycobacterium smegmatis*. FEMS microbiology letters. 2014. pp. 69–77. doi:10.1111/1574-6968.12380

244. Ahidjo BA, Kuhnert D, McKenzie JL, Machowski EE, Gordhan BG, Arcus V, et al. VapC toxins from *Mycobacterium tuberculosis* are ribonucleases that differentially inhibit growth and are neutralized by cognate VapB antitoxins. PLoS One. 2011;6: e21738. doi:10.1371/journal.pone.0021738

245. Ren D, Walker AN, Daines DA. Toxin-antitoxin loci vapBC-1 and vapXD contribute to survival and virulence in nontypeable *Haemophilus influenzae*. BMC Microbiol. 2012;12: 263. doi:10.1186/1471-2180-12-263

246. McKenzie JL, Robson J, Berney M, Smith TC, Ruthe A, Gardner PP, et al. A VapBC toxin-antitoxin module is a posttranscriptional regulator of metabolic flux in mycobacteria. J Bacteriol. 2012;194: 2189–204. doi:10.1128/JB.06790-11

247. Arcus VL, McKenzie JL, Robson J, Cook GM. The PIN-domain ribonucleases and the prokaryotic VapBC toxin-antitoxin array. Protein Eng Des Sel. 2011;24: 33–40. doi:10.1093/protein/gzq081

248. Arcus VL, Backbro K, Roos A, Daniel EL, Baker EN. Distant structural homology leads to the functional characterization of an archaeal PIN domain as an exonuclease. J Biol Chem. 2004;279: 16471–16478. doi:10.1074/jbc.M313833200

249. Bunker RD, McKenzie JL, Baker EN, Arcus VL. Crystal structure of PAEO151 from *Pyrobaculum aerophilum*, a PIN-domain (VapC) protein from a toxin-antitoxin operon. Proteins Struct Funct Genet. 2008;72: 510–518. doi:10.1002/prot.22048

250. Jin G, Pavelka MS, Butler JS. Structure-function analysis of VapB4 antitoxin identifies critical features of a minimal VapC4 toxin-binding module. J Bacteriol. 2015;197: 1197–1207. doi:10.1128/JB.02508-14

251. Gerdes K, Christensen SK, Løbner-Olesen A. Prokaryotic toxin-antitoxin stress response loci. Nat Rev Microbiol. 2005;3: 371–82. doi:10.1038/nrmicro1147

252. Miallau L, Faller M, Janet C, Arbing M, Guo F, Cascio D, et al. Structure and proposed activity of a member of the VapBC family of toxin-antitoxin systems VapBC-5 from *Mycobacterium tuberculosis.* J Biol Chem. 2009;284: 276–283. doi:10.1074/jbc.M805061200

253. Walling LR, Butler JS. Structural determinants for antitoxin identity and insulation of cross talk between homologous toxin-antitoxin systems. J Bacteriol. 2016;198: 3287–3295. doi:10.1128/JB.00529-16

254. Benson DR, Silvester WB. Biology of *Frankia* strains, actinomycete symbionts of actinorhizal plants. Microbiol Rev. 1993;57: 293–319. doi:0146-0749/93/020293-27$02.00/0

255. Mutschler H, Gebhardt M, Shoeman RL, Meinhart A. A novel mechanism of programmed cell death in bacteria by toxin-antitoxin systems corrupts peptidoglycan synthesis. PLoS Biol. 2011;9. doi:10.1371/journal.pbio.1001033

256. Meinhart A, Alonso JC, Strater N, Saenger W. Crystal structure of the plasmid maintenance system epsilon/zeta: functional mechanism of toxin zeta and inactivation by epsilon 2 zeta 2 complex formation. Proc Natl Acad Sci U S A.

2003;100: 1661–1666. doi:10.1073/pnas.0434325100\r0434325100 [pii]

257. Erzberger JP, Berger JM. Evolutionary relationships and structural mechanisms of AAA+ proteins. Annu Rev Biophys Biomol Struct. 2006;35: 93–114. doi:10.1146/annurev.biophys.35.040405.101933
258. Hermon-Taylor J, Bull TJ, Sheridan JM, Cheng J, Stellakis ML, Sumar N. Causation of Crohn's disease by *Mycobacterium avium* subspecies paratuberculosis. Can J Gastroenterol. 2000;14: 521–539.
259. Mutschler H, Meinhart A. Epsilon/Zeta systems: Their role in resistance, virulence, and their potential for antibiotic development. J Mol Med. 2011;89: 1183–1194. doi:10.1007/s00109-011-0797-4
260. Mitchell MS, Rao VB. Novel and deviant Walker A ATP-binding motifs in bacteriophage large terminase-DNA packaging proteins. Virology. 2004;321: 217–221. doi:10.1016/j.virol.2003.11.006
261. Bhattacharyya B, George NP, Thurmes TM, Zhou R, Jani N, Wessel SR, et al. Structural mechanisms of PriA-mediated DNA replication restart. Proc Natl Acad Sci U S A. 2014;111: 1373–8. doi:10.1073/pnas.1318001111
262. Hill P, Heberlig G, Boddy C. Sampling terrestrial environments for bacterial polyketides. Molecules. 2017;22: 707. doi:10.3390/molecules22050707
263. Shen Y, Volrath SL, Weatherly SC, Elich TD, Tong L. A mechanism for the potent inhibition of eukaryotic acetyl-coenzyme A carboxylase by soraphen A, a macrocyclic polyketide natural product. Mol Cell. 2004;16: 881–891. doi:10.1016/j.molcel.2004.11.034
264. George KM. Mycolactone : A polyketide toxin from *Mycobacterium ulcerans* required for virulence. Science. 2008;854. doi:10.1126/science.283.5403.854
265. Hong H, Demangel C, Pidot SJ, Leadlay PF, Stinear T. Mycolactones: immunosuppressive and cytotoxic polyketides produced by aquatic mycobacteria. Nat Prod Rep. 2008;25: 447–454. doi:10.1039/b803101k
266. Hertweck C. The biosynthetic logic of polyketide diversity. Angew Chem Int Ed Engl. 2009;48: 4688–4716. doi:10.1002/anie.200806121
267. Meurer G, Gerlitz M, Wendt-Pienkowski E, Vining LC, Rohr J, Hutchinson CR. Iterative type II polyketide synthases, cyclases and ketoreductases exhibit context-dependent behavior in the biosynthesis of linear and angular

decapolyketides. Chem Biol. 1997;4: 433–43. doi:10.1016/S1074-5521(97)90195-2

268. Iyer LM, Koonin E V, Aravind L. Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily . Proteins. 2001;43: 134–144. doi:10.1002/1097-0134(20010501)43:2<134::AID-PROT1025>3.0.CO;2-I

269. Gazit E, Sauer RT. The Doc toxin and Phd antidote proteins of the bacteriophage P1 plasmid addiction system form a heterotrimeric complex. J Biol Chem. 1999;274: 16813–16818. doi:10.1074/jbc.274.24.16813

Anexo 1

Tabla suplementaria 1. Genomas utilizados para el estudio de genoma esencial y pangenoma del grupo de micobacterias de crecimiento rápido. Se incluyen los números de acceso para cada uno, incluyendo los correspondientes a los plásmidos en los casos en los que están presentes.

Cepa	Números de acceso
Mycobacterium sp CR-UIB1	No disponible
Mycobacterium sp. MG2	No disponible
Mycobacterium sp. MG8	No disponible
Mycobacterium sp. MHSD2	No disponible
Mycobacterium sp. MHSD3	NADK00000000
M. abscessus ATCC 19977^{T}	NC_010397 NC_010394 (plasmido)
M. abscessus FLAC004	NZ_CP014951
M. abscessus FLAC003	NZ_CP014950
M. abscessus NOV0213	NZ_CP013049
M. abscessus UC22	NZ_CP012044
M. abscessus subsp *bolletii* 50594	NC_021282 NC_021278 (Plasmido 1) NC_021279 (Plasmido 2)
M. abscessus subsp *bolletii* CCUG 50184^{T}	LDMY00000000
M. abscessus subsp *bolletii* CCUG 48898	NZ_AP014547
M. abscessus subsp *bolletii* 103	JAOK00000000
M. chelonae CCUG 47445^{T}	NZ_CP007220
M. chelonae 1518	JAOI00000000
M. chelonae ATCC 35752	CP010946
M. chubbuense NBB4	NC_018027 NC_018022 (pMYCCH.01) NC_018023 (pMYCCH.02)
M. chubuense DSM 44219	NZ_JYNX01000000
M. fortuitum Z58	NZ_JASW00000000
M. fortuitum DSM 46621	NZ_ALQB01000000
M. fortuitum ATCC 6841	CRVX00000000
M. gilvum PYR-GCK	NC_009338 NC_009339 (pMFLV01) NC_009340 (pMFLV02) NC_009341 (pMFLV03)
M. gilvum Spyr1	NC_014814 NC_014811 (pMSPYR101) NC_014812 (pMSPYR102)
M. hassiacum DSM 44199	NZ_AMRA00000000
M. hassiacum DSM 44199	NZ_ARBU01000000
M. immunogenum CCUG 47286^{T}	NZ_CP011530
M. immunogenum SMUC14	JXUU01000000
M. llatzerense CLUC14	JXST00000000

Continuación de la Tabla suplementaria 1. Genomas utilizados para el estudio de genoma esencial y pangenoma del grupo de micobacterias de crecimiento rápido. Se incluyen los números de acceso para cada uno, incluyendo los correspondientes a los plásmidos en los casos en los que están presentes.

Cepa	Números de acceso
***M. llatzerense* MG13^T**	No disponible
***M. marinum* Europe**	ANPL00000000
M. marinum MB2	NZ_ANPM01000000
M. marinum M	NC_010612 NC_010604 (pMM23)
M. masiliense GO06	NC_018150
M. neoaurum VKM-Ac-1815D	NC_023036
M. neoaurum ATCC 25795^T	NZ_JMDW01000000
M. neoaurum DSM 44704^T	CCDR0100000000
M. neoaurum MN4	JXYZ01000000
M. phlei RIVM601174	NZ_AJFJ01000000
M. rhodesiae NBB3	NC_016604
M. rhodesiae JS60	NZ_AGIQ01000000
M. smegmatis JS623	NC_019966 NC_019957 (pMYCSM01) NC_019958 (pMYCSM02) NC_019959 (PMYCSM03)
M. smegmatis MC2-155	NC_008596
M. smegmatis MC2-155	NC_018289
M. smegmatis MC2-155	NZ_CP009494
M. smegmatis MC2-51	NZ_JAJD01000000
M. smegmatis MKD8	NZ_A0CJ01000000
M. smegmatis INHR1	NZ_CP009495
M. smegmatis INHR2	NZ_CP009496
M. thermoresistibile ATCC 19527^T	AGVE01000000
M. vaccae ATCC 25954	ALQA01000000
M. vanbaalenii PYR-1	NC_008726

Tabla suplementaria 2. Genomas utilizados para el estudio de genoma esencial y pangenoma del grupo de *abscessus-chelonae-immunogenum*. Se incluyen los números de acceso para cada uno, incluyendo los correspondientes a los plásmidos en los casos en los que están presentes. Los genomas correspondientes a *M. immunogenum* incluidos en esta tabla son los utilizados para el mismo estudio centrado en esta especie.

Cepa	Números de acceso
M. abscessus 625	FSPH00000000
M. abscessus 652	FVXK00000000
M. abscessus 666	FVXM00000000
M. abscessus 676	FSMP00000000
M. abscessus 690	FVXK00000000
M. abscessus 746	FSJU00000000
M. abscessus 748	FVCA00000000
M. abscessus FLAC003	NZ_CP014950

Continuación de la Tabla suplementaria 2. Genomas utilizados para el estudio de genoma esencial y pangenoma del grupo de *abscessus-chelonae-immunogenum*. Se incluyen los números de acceso para cada uno, incluyendo los correspondientes a los plásmidos en los casos en los que están presentes. Los genomas correspondientes a *M. immunogenum* incluidos en esta tabla son los utilizados para el mismo estudio centrado en esta especie.

Cepa	Números de acceso
M. abscessus FLAC004	NZ_CP014951
M. abscessus FLAC005	NZ_CP014952
M. abscessus FLAC007	NZ_CP014953
M. abscessus FLAC008	NZ_CP014954
M. abscessus FLAC031	NZ_CP014957
M. abscessus FLAC013	CP014955
M. abscessus FLAC045	CP014958
M. abscessus FLAC048	NZ_CP014959
M. abscessus FLAC049	NZ_CP014960
M. abscessus UC95	LGCJ00000000
M. abscessus 4529	CP009616
M. abscessus ATCC 19977^{T}	CU458896 CU458745 (plasmido)
M. abscessus DJO 44274	CP009615
M. abscessus subsp *bolletii* 103	CP009407
M. abscessus subsp *bolletii* 1513	JAOJ00000000
M. abscessus subsp *bolletii* 1S 151 0930	AKUI00000000
M. abscessus subsp *bolletii* 1S 152 0914	AKUJ00000000
M. abscessus subsp *bolletii* 1S 153 0915	AKUK00000000
M. abscessus subsp *bolletii* 1S 154 0310	AKUL00000000
M. abscessus subsp *bolletii* 2B 0107	AKUN00000000
M. abscessus subsp *bolletii* 2B 0626	AKUM00000000
M. abscessus subsp *bolletii* 2B 0912 R	AKUV00000000
M. abscessus subsp *bolletii* 2B 0912 S	AKUW00000000
M. abscessus subsp *bolletii* 2B 1231	AKUO00000000
M. abscessus subsp *bolletii* 50594	CP004374 CP004375 (plasmid1) CP004376 (plasmid2)
M. abscessus subsp *bolletii* BD^{T}	AHAS00000000
M. abscessus subsp *bolletii* CCUG 48898	AP014547
M. abscessus subsp *bolletii* CRM 0020	ATFQ00000000
M. abscessus subsp *bolletii* INCQS 00594	AUVF00000000 CP003376 (pMAB01)
M. abscessus subsp *bolletii* M18	AJSC00000000
M. abscessus subsp *bolletii* MA 1948	CP009408
M. abscessus subsp *bolletii* M2B 307	AKUU00000000
M. abscessus subsp *bolletii* MC1518	CP009613
M. abscessus subsp *bolletii* MM1513	CP009447
M. abscessus subsp *bolletii* str GO 06	CP003699.2

Continuación de la Tabla suplementaria 2. Genomas utilizados para el estudio de genoma esencial y pangenoma del grupo de *abscessus-chelonae-immunogenum*. Se incluyen los números de acceso para cada uno, incluyendo los correspondientes a los plásmidos en los casos en los que están presentes. Los genomas correspondientes a *M. immunogenum* incluidos en esta tabla son los utilizados para el mismo estudio centrado en esta especie.

Cepa	Números de acceso
M. abscessus subsp *bolletii* CCUG 50184^T	LDMY00000000
M. abscessus UC22	NZ_CP012044
M. abscessus NOV0213	NZ_CP013049
M. chelonae ATCC 35752^T	CP010946
M. immunogenum CCUG 47286^T	NZ_CP011530
M. immunogneum H088	LJFT00000000
M. immunogenum SMUC14	JXUU01000000
M. immunogneum H097	LJFU00000000
M. immunogneum HXV	LJFX00000000
M. immunogneum H106	LJFV00000000
M. immunogneum H008	LJFO00000000
M. immunogneum HXXI	LJFY00000000
M. immunogneum H074	LJFR00000000
M. immunogneum H076	LJFS00000000
M. immunogneum H060	LJFP00000000
M. immunogneum H068	LJFQ00000000
M. immunogneum H113	LJFW00000000
M. chelonae CCUG 47445^T	NZ_CP007220
Mycobacterium sp. MG2	No disponible
Mycobacterium sp. MG8	No disponible
Mycobacterium sp. MHSD2	No disponible
Mycobacterium sp. MHSD3	NADK00000000
Mycobacterium sp. CR-UIB1	No disponible

Tabla suplementaria 3. Genomas utilizados para el estudio de genoma esencial y pangenoma de la especie *M. tuberculosis*. Se incluyen los números de acceso para cada uno.

Cepa	Números de acceso
M. tuberculosis H37Rv	AL123456.3
M. tuberculosis CDC1551	AE000516.2
M. tuberculosis H37Ra	CP000611
M. tuberculosis F11	CP000717
M. tuberculosis KZN 1435	CP001658
M. tuberculosis str. Haarlem	CP001664
M. tuberculosis KZN 4207	CP001662
M. tuberculosis KZN 605	CP001976
M. tuberculosis W-148	CP012090
M. tuberculosis CTRI-2	CP002992
M. tuberculosis CCDC5180	CP001642
M. tuberculosis 7199-99	HE663067
M. tuberculosis str. Erdman = ATCC 35801	AP012340
M. tuberculosis str. Beijing/NITR203	CP005082
M. tuberculosis EAI5/NITR206	CP005387
M. tuberculosis CCDC5079	CP002884
M. tuberculosis EAI5	CP006578
M. tuberculosis HKBS1	CP002871
M. tuberculosis BT2	CP002882
M. tuberculosis BT1	CP002883
M. tuberculosis K	CP007803
M. tuberculosis KIT87190	CP007809
M. tuberculosis ZMC13-264	CP009100
M. tuberculosis ZMC13-88	CP009101
M. tuberculosis 96075	CP009426
M. tuberculosis 96121	CP009427
M. tuberculosis 49-02	HG813240
M. tuberculosis H37RvSiena	CP007027
M. tuberculosis str. Kurono	AP014573
M. tuberculosis SCAID 187.0	CP012506
M. tuberculosis F1	CP010329
M. tuberculosis F28	CP010330
M. tuberculosis 2242	CP010335
M. tuberculosis 2279	CP010336
M. tuberculosis 22115	CP010337
M. tuberculosis 37004	CP010338
M. tuberculosis 22103	CP010339
M. tuberculosis 26105	CP010340
M. tuberculosis str. Haarlem/NITR202	CP004886
M. tuberculosis CAS/NITR204	CP005386

Tabla suplementaria 4. Listado de los códigos de las categorías funcionales de los COG relacionados con las funciones específicas que representan en cada caso.

Codigo COG	Categoría funcional
J	Traducción, estructura ribosomal y biogénesis
A	Procesamiento y modificación del RNA
K	Transcripción
L	Replicación, recombinación y reparación
B	Estructura y dinámica de la cromatina
D	Control del ciclo cellular, division cellular, partición del cromosoma
Y	Estructura nuclear
V	Mecanismos de defensa
T	Mecanismos de transducción de señal
M	Biogénesis de la pared cellular, membrana y envoltura
N	Motilidad celular
Z	Ctoesqueleto
W	Estructuras extracelulares
U	Tráfico intracellular, secreción y transporte vesicular.
O	Modificación post-transcripcional, recambio de proteínas, chaperonas
X	Mobiloma: profagos, transposones
C	Producción y conversión de energía
G	Transporte y metabolismo de carbohidratos
E	Transporte y metabolism de aminoácidos
F	Transporte y metabolismo de nucleótidos
H	Transporte y metabolismo de coenzimas
I	Transporte y metabolismo de lípidos
P	Transporte y metabolismo de iones inorgánicos
Q	Biosíntesis, transporte y catabolismo de metabolitos secundarios
R	Solo predicción de función general
S	Función desconocida

Anexo 2

Tabla suplementaria 1. Listado completo de todos los FT encontrados en las cepas de MCR estudiadas, incluyendo las listadas en la tabla reducida el capítulo 3.

	CCUG 47445^T	CCUG 47286^T	CCUG 50184^T	MG2	MG8	MHSD2	MHSD3	CR-UIB1
AlpA	0	0	1	0	0	0	0	0
AraC	13	22	17	13	14	19	14	20
ArgR	1	1	1	1	1	1	1	1
ArsR	14	11	13	14	14	13	17	14
AsnC	4	4	4	5	5	4	5	5
Crp	1	1	2	1	1	2	1	3
CsoR	2	2	2	2	2	2	2	2
DeoR	0	1	1	0	0	0	0	0
DtxR	1	2	2	2	2	2	1	2
FeoC	0	0	0	0	0	0	0	1
Fur	3	2	3	3	3	2	3	2
GntR	12	15	14	10	11	16	13	16
HrcA	1	1	1	1	1	1	1	1
HxlR	6	6	5	5	4	5	5	7
IclR	7	9	9	6	6	9	7	14
KorB	0	1	0	0	0	0	0	0
LacI	0	1	2	0	0	1	0	1
LexA	1	1	1	1	1	1	1	1
LuxR	3	2	2	3	3	2	2	2
LysR	16	22	21	14	15	16	17	21
MarR	21	27	21	19	19	21	20	21
MerR	8	8	7	8	8	6	7	7
Mga	0	0	0	0	0	0	0	1
NrdR	1	1	1	1	1	1	1	1
PadR	7	5	6	6	6	6	6	6

Continuación de la Tabla suplementaria 1. Listado completo de todos los FT encontrados en las cepas de MCR estudiadas, incluyendo las listadas en la tabla reducida el capítulo 3.

	CCUG 47445^T	**CCUG 47286^T**	**CCUG 50184^T**	**MG2**	**MG8**	**MHSD2**	**MHSD3**	**CR-UIB1**
Rok	1	1	1	1	1	1	1	1
PucR	1	2	2	1	1	2	1	2
RpiR	0	1	1	0	0	0	0	0
Rrf2	2	1	1	1	1	1	2	2
Sarp	0	1	1	0	0	1	0	1
TetR	144	155	149	143	144	146	145	138
TrmB	3	1	3	1	1	3	1	3
WhiB	8	10	9	7	7	8	7	6
Xre	18	33	15	14	14	14	18	18
No clasif.	7	8	5	5	4	5	10	5

Tabla suplementaria 2. Listado completo de todos los FT encontrados en las cepas cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv., incluyendo las listadas en la tabla reducida el capítulo 3.

	CR-UIB2	H37Rv
AbrB	2	2
AraC	6	6
ArgR	1	1
ArsR	12	12
AsnC	2	2
Crp	2	2
CsoR	3	3
DtxR	2	2
Fur	2	2
GntR	7	7
HrcA	1	1
HxlR	2	2
IclR	2	2
KorB	1	1
LacI	1	1
LexA	1	1
LuxR	6	6
LysR	4	4
MarR	9	9
MerR	3	3
NrdR	1	1
PadR	3	3
Rrf2	1	1
Sarp	3	2
TetR	50	49
TrmB	1	1
WhiB	7	7
Xre	15	15
No clasifi.	5	5

Tabla suplementaria 3. Familias de reguladores de respuesta encontrados en los genomas de las cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv. Se indica el número de representantes encontrados en cada caso.

	Amir_NasR	CheY	NarL	OmpR
CR-UIB2	1	1	2	9
H37Rv	1	1	2	9

Tabla suplementaria 4. Factores sigma identificados en el análisis del reguloma de los genomas de las cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv. Se indica el número de representantes encontrados en cada caso.

	CR-UIB2	H37Rv
SigA	1	1
SigB	1	1
SigC	1	1
SigD	1	1
SigE	1	1
SigF	1	1
SigG	1	1
SigH	1	1
SigI	1	1
SigJ	1	1
SigK	1	1
SigL	1	1
SigM	1	1

Tabla suplementaria 5. Reguladores negativos (Factores antisigma) identificados en la prospección de los proteomas de las cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv. Se indica la presencia (+) o ausencia (-) de cada factor en los distintos genomas.

	CR-UIB2	H37Rv
Anti-Sig (RshA)	+	+
Anti-SigF(RsfA)	+	+
Anti-SigE(RseA)	+	+
Anti-SigM(RsmA)	+	+
Anti-SigD(RsdA)	+	+
Anti-SigL(RslA)	+	+
Anti-SigF(RsbW)	+	+
Anti-SigF(RsfB)	+	+
Anti-SigK(RskA)	+	+

Tabla suplementaria 6. Número de elementos relacionados con SDC encontrados en los genomas de las cepas *M. tuberculosis* CR-UIB2 y *M. tuberculosis* H37Rv. Se indica el número de histidina quinasas, reguladores de respuesta y el número de SDC completos formados entre ellos. Se indica también el número de elementos para los cuales no se ha hallado el elemento relacionado que completaría el SDC.

	CR-UIB2	**H37Rv**
HK	14	15
RR	13	13
TCS	9	10
HK solitarias	5	5
RR solitarios	4	3
Prot. fosfotransferasa	1	1

Anexo 3

Tabla suplementaria 1. Cebadores diseñados para la amplificación de los elementos de los STA. En rojo se muestran las dianas de restricción pertinentes en cada caso. Previa a esta diana se insertaron 6 nucleótidos coincidentes con la secuencia original del genoma para incrementar la eficiencia de hibridación en la PCR inicial.

Diana	Cebador F (5'->3')	Cebador R (5'->3')
VapB28	TACCAAccATGGCCCTGAACATCAAAGA	TCTCGTAAGCTTGGATGGCGATGATTGATGAC
VapC28	CTGGGAccatggAGATGATCATCGATACGTCA	CGCCAAAAGCTTGGTGCCGTTCTACAAGTTCTG
VapB27	CTAATCCCATGGAAGCTGTCATCGACTCA	GACATGaagcttATGCGAGGTCACAAGCAAC
VapC27	TTCGGGcatatgCTGACATCGCCGCGGGTCA	GCTTCCctcgagGTCGTCCATTGGCTATCCAG
MT0933	ATAGATGAATTCaaACCATGGGATTTCTGGACAAG	CAGGCCAAGCTTGGAGACAGCGGTACTTCGATG
MT0934	CAGTAGcCATGGCAAAGCTCGCAAGTTC (A)	TTCACGAAGCTTGCTGTCTAAACCGACGCGAAC
KnB A	GGAGAACCTCcCATGGCTGATTTCAA (A)	AGCGCCAAGCTTAGCAGCGCCAGTCGGTTACTGA
Mch L	TGGCGCCATATGGCGGCGCTGGTGTTGGTCA	acgataCTCGAGGACATCAACAGACGGCTA
Mch MT0934	TCGTTTccatgGTGGCAGCCAAACTGAATTCc	AGAACGAAGCTTATGGCGATTACGAGGCAGATG
Mch Toxin2	AACCGTcCATGGGACACATTGAAGCAAC	TGTCGTAAGCTTCAAGATCGTCAGGCATCTTC
Mch Toxin3	CTTCCGcatATGTCCGAGCTGCACTCCA	CTGTGActcgagTCGACGTCGGCAGACTCGGTTA
Mim MT0933	gCAAGCcCATGGCTGATTTCAAG	cgcaagaagcttcgccagtcggttactgag
MimLip	GCCGCCcatatGGTGCTGTCCGACGCGCCGTTG	acgtggctcgagtaccggagcgatacgaac
Mimm MT0934	CCTTTGccatgGTGGCAGCCAAACTGAATTCC	gataccaagcttcacaaaccgtcgcttatg
MG2 MT0933	AACCTCcCATGGCTGATTTCAAGGGCCTCATC	CGCCAGaagcttCGCCAGTCGGTTACTGAG
MG2 LIP	CTCtCtcatatgGGGTGAGGTGCTTTCCGA	AGACTActcgagGAGCGACATCAACAGACGGCTA
MG2 MT0934	CGTCTGccatgGTGGCAGCCAAACTGAATTCc	ACGGTCaagcttGTTTTCGCAGGTAGAACG
MG2 Ztox	agaTCAGGGGCAtATGGCCGGATA	CCGCCActcgagGACGATGATCGAGATCAG
MD2 MT0933	ACAAGCcCATGGCTGATTTCAAGGGCCTCA	CAGTCCaagcttATCAACACCAGCACCGCAAGCAG
MD2 lip	TTGCCAcatatgCGGTCGAGAAGGTCGAAGGTG	ACGATActcgagAGATCCGAACAGACCGCTA
MD2 MT0934	CGTCTGccatgGTGGCAGCCAAACTGAATTCC	TCACAAaagcttGCTTATGTTTTCGCAGGTAGAC
MD2 MT0934S	CTTTCGcatATGTCTGAGCTGCACTCCAGCATC	GTCCTGctcgagTTCCGTGAACGTGGTTCGTG
MD2 Z	TGGtGGcatATGACCGGATGTCCGCAG	TCTACCctcgagTTGTCAGGCCCTCGTGGACAAC
MD2 MT0934S	CTTCCGcatATGTCCGAGCTGCACTCCAGCATC	CTTTCCctcgagAAAGTGATCGACGTCGGCAGA
Ztoxin2	TACCCAccatgGTGAAACGGCTCGATCTGATCGTC	TTTGACaagcttTACTCACACATCGGACGCTA
Z2 Antitox	gagATCcCATGGCGGCTCCGGTAGA	TTGGGAaagcttATCAGATCGAGCCGTTTCAC
Doc	GAAGACcatATGATCACGTTCTATCTAACTGC	CGTGtTctcgagCCTCGCATGGCCTTAGTCGTTG
H1phd	GACATCccatggTCATGAGTATGACACAGCCCGAGAA	cgtggtaagcttCAGTTAGATAGAACGTGATCATC
Ztoxin	agaTCAGGGGCAtATGGCCGGATA	CCGCCActcgagGACGATGAcCGAGATCAG
VapB5	GCTACCccatgGTGGTGAACACCGTAGGCCTGCGTGA	GGTCACaagcttATTGCGAGCTCGCCGTCGATC
VapC5	TGGtCGccatgGTGAGGGCGGTGCTCGATAC	AaAAGGaagcttCGTTGTAATACACAGTGCTCACGG

Tabla sumplementaria 2. Porcentajes de identidad entre las secuencias de las proteínas MT0933, la lipasa y MT0934, utilizando la secuencia del genoma de la cepa tipo de *M. chelonae* CCUG 47445^T como referencia.

Cepa	MT0933	Lipasa	MT0934
CCUG 47445^T	100	100	100
MHSD3	100	91	100
MG2	98	91	100
MG8	98	92	100
CR-UIB-1	95	85	91
MHSD2	95	85	91
CCUG 47286^T	87	87	91

Tabla suplementaria 3. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapB27 de *M. llatzerense* MG13^T. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***M. gilvum* DSM 45189 / LMG 24558 / Spyr1**	Regulador transcripcional, Familia AbrB	ADU01177.1
***Mycobacterium sp.* KMS**	Regulador transcripcional, Familia AbrB	ABL90056.1
***M. chlorophenolicum* DSM 43826^T**	Antitoxina VapB27	KMO67373.1
***M. chubuense* DSM 44219^T**	Antitoxina VapB27	KMO84272.1
***M. gilvum* PYR-GCK**	Regulador transcripcional, Familia AbrB	ABP47662.1
***Acidipropionibacterium acidipropionici* ATCC 4875 / DSM 20272 / JCM 6432**	Regulador transcripcional, Familia AbrB	AFV91064.1
***Kineosphaera limosa* NBRC 100340**	Proteína no caracterizada	GAB94656.1
***Gordonia rhizosphera* NBRC 16068**	Proteína no caracterizada	GAB91680.1
***Microlunatus phosphovorus* ATCC 700054 / DSM 10555 / JCM 9379 / NBRC 101784 / NCIMB 13414 / VKM Ac-1990 / NM-1**	Proteína no caracterizada	BAK36694.1
***Tetrasphaera japonica* T1-X7**	Regulador transcripcional MraZ	CCH80362.1
***Microbacterium sp.* Leaf203**	Regulador transcripcional, Familia AbrB	KQM40320.1
***Frankia sp.* CcI6**	Regulador transcripcional, Familia AbrB	ESZ99906.1
***Frankia sp.* BMG5.23**	Regulador transcripcional, Familia AbrB	KDA40896.1
***Frankia casuarinae* DSM 45818 / CECT 9043 / CcI3**	Regulador transcripcional, Familia AbrB	ABD11927.1
***Frankia sp.* CcI6**	Regulador transcripcional, Familia AbrB	ETA03209.1
***Frankia sp.* BMG5.23**	Proteína no caracterizada	KDA41364.1
***Frankia casuarinae* DSM 45818 / CECT 9043 / CcI3**	Regulador transcripcional, Familia AbrB	ABD12648.1
***Frankia sp.* QA3**	Proteína de unión al ADN, Familia AbrB	EIV91044.1

Continuación de la Tabla suplementaria 3. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapB27 de *M. llatzerense* MG13^T. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***M. tuberculosis* CAS/NITR204**	Proteína no caracterizada	AGL26080.1
***M. orygis* 112400015**	Proteína no caracterizada	EMT37144.1
***M. bovis* BCG / Pasteur 1173P2**	Proteína no caracterizada	CAL70630.1
***M. tuberculosis* complex multispicies**	Antitoxina VapB27	WP_003403139.1
***M. tuberculosis* ATCC 35801 / TMC 107 / Erdman**	Proteína no caracterizada	BAL64474.1
***M. tuberculosis* ATCC 25618 / H37Rv**	Antitoxina	AIR13323.1
***M. tuberculosis* ATCC 25177 / H37Ra**	Proteína no caracterizada	ABQ72330.1
***M. canettii* CIPT 140010059**	Proteína no caracterizada	CCC42945.1
***M. africanum* GM041182**	Proteína no caracterizada	CCC25678.1
M. bovis BCG-1	Antitoxina	WP_003403139.1
M. tuberculosis DK9897	Antitoxina	APR56067.1
***M. tuberculosis* C**	Antitoxina VapB27	EAY59007.1
***M. tuberculosis* BTB05-013**	Antitoxina VapB27	KCN23102.1
***M. tuberculosis* MAL010121.**	Antitoxina VapB27	KBH07155.1
***M. bovis* BCG**	Putativa Antitoxina vapb27	CUI10382.1
***M. tuberculosis* OFXR-27**	Antitoxina VapB27	KBJ95943.1
***M. tuberculosis* BTB05-348**	Antitoxina VapB27	KCN32362.1
***M. tuberculosis* M1034**	Antitoxina VapB27	KAX84884.1
***M. tuberculosis* T46**	Antitoxina	EFD12147.1
***M. tuberculosis* W-148**	Antitoxina VapB27	EGE49364.1
***M. tuberculosis* ATCC 25618 / H37Rv**	VapB27	CCP43338.1
***Blastococcus saxobsidens* DD2**	Putativa Regulador transcripcional	CCG05636.1
***M. tuberculosis* CDC 1551 / Oshkosh**	Proteína no caracterizada	AAK44853.1
***M. caprae* MB2**	Proteína no caracterizada	CEJ50622.1
***M. bovis* MB1**	Putativa Proteína no caracterizada	CEJ34357.1
***Frankia alni* ACN14a**	Proteína no caracterizada	CAJ58885.1
***Frankia sp.* EUN1f**	Regulador transcripcional, Familia AbrB	EFC80503.1
***Frankia sp.* R43**	Antitoxina	KPM50928.1
***M. tuberculosis* Haarlem/NITR202**	Proteína no caracterizada	AGL22354.1
***M. tuberculosis* D00501624**	Putativa antitoxina	CNX55889.1
***M. bovis* BCG**	Proteína con el dominio de la superfamilia Antitoxina-Maz abrB	ALA769811
M. mucogenicum1199456.5	Regulador transcripcional, Familia AbrB	OBA79891.1
***Frankia sp.* QA3**	Proteína de union al ADN, Familia AbrB	EIV92738.1

Tabla suplementaria 4. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapC27 de *M. llatzerense* MG13^T. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Nº de acceso
***Mycobacterium* sp. MCS**	Ribonucleasa VapC	ABG06943.1
***Mycobacterium* sp. KMS**	Ribonucleasa VapC	ABL90057.1.
***Mycobacterium* sp. JLS**	Ribonucleasa VapC	ABN96655.1
***Mycobacterium* sp. Spyr1**	Ribonucleasa VapC	ADU01176.1.
***Gordonia rhizosphera* NBRC 16068**	Proteína no caracterizada	GAB91679.1.
***Propionibacterium acidipropionici* ATCC 4875 / DSM 20272 / JCM 6432 / NBRC 12425 / NCIMB 8070**	Ribonucleasa VapC	AFV91065.1.
***Microlunatus phosphovorus* ATCC 700054 / DSM 10555 / JCM 9379 / NBRC 101784 / NCIMB 13414 / VKM Ac-1990 / NM-1**	Ribonucleasa VapC	BAK36693.1.
***Kineosphaera limosa* NBRC 100340**	Ribonucleasa VapC	GAB94657.1.
***Frankia* sp. EAN1pec**	Ribonucleasa VapC	ABW11781.1.
***Frankia* sp. CcI3**	Ribonucleasa VapC	ABD12649.1.
***Frankia* sp. Thr**	Ribonucleasa VapC	EYT91509.1
***Frankia* sp. CcI6**	Ribonucleasa VapC	ETA03208.1.
***Frankia* sp. QA3**	Ribonucleasa VapC	EIV91045.1.
***Frankia* sp. BMG5.23**	Ribonucleasa VapC	KDA41363.1.
***Tetrasphaera japonica* T1-X7**	Proteína no caracterizada	CCH80361.1.
***Frankia alni* ACN14a**	Ribonucleasa VapC	CAJ58884.1.
Fran*kia* sp. EUN1f	Ribonucleasa VapC	EFC80502.1.
***Frankia* sp. CeD**	Proteína de union al ADN	KEZ37000.1
***Microlunatus phosphovorus* ATCC 700054 / DSM 10555 / JCM 9379 / NBRC 101784 / NCIMB 13414 / VKM Ac-1990 / NM-1**	Ribonucleasa VapC	BAK36962.1.
***Frankia* sp. Allo2**	Proteína con dominio PIN	KFB02672.1
***Frankia* sp. CcI6**	Ribonucleasa VapC	ESZ99905.1.
***Frankia* sp. BMG5.23**	Ribonucleasa VapC	KDA40897.1.
***Frankia* sp. CeD**	Proteína con dominio PIN	KEZ34451.1
***M. gastri* 'Wayne'**	Ribonucleasa VapC	ETW24979.1.
***M. tuberculosis* Haarlem/NITR202**	Ribonucleasa VapC	AGL22353.1.
***M. tuberculosis* ATCC 35801 / TMC 107 / Erdman**	Ribonucleasa VapC	BAL64473.1
***M. tuberculosis* SUMu007**	Ribonucleasa VapC	EFP35917.1
***M. tuberculosis* HKBS1**	Ribonucleasa VapC	AHJ41271.1
***M. tuberculosis* str. Beijing/NITR203**	Ribonucleasa VapC	AGJ66614.1
***M. tuberculosis* 7199-99**	Ribonucleasa VapC	CCG10461.1
***M. tuberculosis* CAS/NITR204**	Ribonucleasa VapC	AGL26079.1.
***M. tuberculosis* RGTB423**	Ribonucleasa VapC	AFE11915.1
***M. tuberculosis* BT1**	Ribonucleasa VapC	AHJ49571.1
***M. tuberculosis* RGTB327**	Ribonucleasa VapC	AFE15568.1

Continuación de la Tabla suplementaria 4. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapC27 de *M. llatzerense* MG13^T. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***M. bovis* BCG str. Korea 1168P**	Ribonucleasa VapC	AGE66546.1
***M. tuberculosis* K**	Ribonucleasa VapC	AIB47162.1
***M. bovis* BCG / Pasteur 1173P2**	Ribonucleasa VapC	CAL70629.1
***M. tuberculosis* CTRI-2**	Ribonucleasa VapC	AEM99050.1
***M. bovis* ATCC BAA-935 / AF2122/97**	Ribonucleasa VapC	CDO41856.1
***M. bovis* BCG str. ATCC 35743**	Ribonucleasa VapC	AHM06305.1
***M. tuberculosis* ATCC 25177 / H37Ra**	Ribonucleasa VapC	ABQ72329.1.
***M. tuberculosis* EAI5/NITR206**	Ribonucleasa VapC	AGL30032.1
***M. bovis* BCG str. Mexico**	Ribonucleasa VapC	AET17900.1
***M. tuberculosis* KZN 605**	Ribonucleasa VapC	AFM47989.1
***M. bovis* BCG / Tokyo 172 / ATCC 35737 / TMC 1019**	Ribonucleasa VapC	BAH24907.1
***M. tuberculosis* BT2**	Ribonucleasa VapC	AHJ45423.1
***M. tuberculosis* EAI5**	Ribonucleasa VapC	AGQ34131.1
***M. canettii* CIPT 140010059**	Ribonucleasa VapC	CCC42944.1
***M. africanum* GM041182**	Ribonucleasa VapC	CCC25677.1
***M. canettii* CIPT 140070008**	Ribonucleasa VapC	CCI89776.1
***M. mucogenicum* (secuencia de referencia)**	Ribonucleasa	WP_061001017.1

Tabla suplementaria 5. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapB28 de *M. llatzerense* MG13^T. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***M. tuberculosis* TKK-01-0050**	Antitoxina VapB28	KBZ57868.1
***M. tuberculosis* TKK_04_0075**	Antitoxina VapB28	KAT96417.1
***M. tuberculosis* TKK_04_0061**	Antitoxina VapB28	KAT54855.1
***M. tuberculosis* TKK-01-0014**	Antitoxina VapB28	KBY44366.1
***M. tuberculosis* HKBS1**	Antitoxina	AHJ41281.1
***M. tuberculosis* 7199-99**	Putativa antitoxina VAPB28	CCG10471.1
***M. tuberculosis* BT1**	Antitoxina	AHJ49581.1
***M. bovis* BCG str. Korea 1168P**	Proteína no caracterizada	AGE66555.1
***M. tuberculosis* K**	Antitoxina	AIB47173.1
***M. bovis* BCG str. ATCC 35743**	Factor de transcripción	AHM06314.1
***M. bovis* ATCC BAA-935 / AF2122/97M.**	Posible antitoxina vapb28	CDO41866.1
***M. tuberculosis* KZN 605**	Antitoxina	AFM47998.1
***M. tuberculosis* BT2**	Antitoxina	AHJ45433.1
***M. tuberculosis* EAI5**	Antitoxina	AGQ34138.1
***M. canettii* CIPT 140010059M.**	Proteína no caracterizada	CCC42949.1.

Continuación de la Tabla suplementaria 5. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapB28 de *M. llatzerense* MG13^T. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***M. africanum* GM041182M.**	Proteína no caracterizada	CCC25687.1.
***M. tuberculosis* str. Haarlem**	Antitoxina VapB28	EBA41123.1
***M. canettii* CIPT 140070008**	Proteína no caracterizada	CCI89781.1
***M. tuberculosis* M.M. ATCC 25618 / H37RvM.**	Antitoxina	AIR13332.1.
***M. tuberculosis* M.M. F11M.**	Antitoxina VapB28	NC_009565.1
***M. canettii* CIPT 140070010**	Proteína no caracterizada	CCI98308.1
***M. tuberculosis* M.M. KZN 1435 / MDRM.**	Antitoxina	ACT23648.1
***M. tuberculosis* KZN 4207**	Antitoxina VapB28	AEB02748.1
***M. tuberculosis* MAL020120**	Antitoxina VapB28	KBH56077.1
***M. tuberculosis* TB_RSA111**	Antitoxina VapB28	KCJ37189.1
***M. tuberculosis* KT-0006**	Antitoxina VapB28	KCF40132.1
***M. tuberculosis* M1295**	Antitoxina VapB28	KAY55550.1
***M. tuberculosis* TB_RSA114**	Antitoxina VapB28	KCJ41358.1
***M. tuberculosis* BTB04-452**	Antitoxina VapB28	KCN16458.1
***M. tuberculosis* TB_RSA174**	Antitoxina VapB28	KCL73960.1
***M. tuberculosis* NRITLD12**	Antitoxina VapB28	KBU89020.1
***M. tuberculosis* M1272**	Antitoxina VapB28	KAY40302.1
***M. tuberculosis* MD17517**	Antitoxina VapB28	KAP40402.1
***M. tuberculosis* NRITLD59**	Antitoxina VapB28	KBU74088.1
***M. tuberculosis* TKK_03_0089**	Antitoxina VapB28	KAS19154.1
***M. tuberculosis* TB_RSA144**	Antitoxina VapB28	KCK53991.1
***M. tuberculosis* BTB12-313**	Antitoxina VapB28	KCR46489.1
***M. tuberculosis* UT0014**	Antitoxina VapB28	KBS99421.1
***M. tuberculosis* M1410**	Antitoxina VapB28	KAZ49575.1
***M. tuberculosis* TKK_04_0114**	Antitoxina VapB28	KAW85920.1
***M. tuberculosis* XDR KZN 605**	Antitoxina VapB28	KBM93954.1
***M. tuberculosis* TKK_04_0141**	Antitoxina VapB28	KAM87474.1
***M. tuberculosis* MD17888**	Antitoxina VapB28	KAP63965.1
***M. tuberculosis* TKK_03_0111**	Antitoxina VapB28	KAS74900.1
***M. tuberculosis* TB_RSA16**	Antitoxina VapB28	KBO02588.1
***M. tuberculosis* XTB13-177**	Antitoxina VapB28	KCU78933.1
***M. tuberculosis* TKK-01-0066**	Antitoxina VapB28	KCA16365.1
***M. tuberculosis* MD14002**	Antitoxina VapB28	KAQ74311.1
***M. tuberculosis* M1475**	Antitoxina VapB28	KBA35660.1
***M. tuberculosis* TKK_02_0057**	Antitoxina VapB28	KBX02525.1

Tabla suplementaria 6. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapC28 de *M. llatzerense* MG13^T. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Nº de acceso
***Mycobacterium* sp. Root135**	Ribonucleasa VapC	KQY07927.1.
***Mycobacterium heraklionense* Davo**	Ribonucleasa VapC	KLO26365.1.
***Mycobacterium vanbaalenii* (strain DSM 7251 / PYR-1)**	Ribonucleasa VapC	ABM12006.1.
***Mycobacterium orygis* 112400015**	Ribonucleasa VapC	EMT37154.1.
***Mycobacterium bovis* (strain BCG / Pasteur 1173P2)**	Ribonucleasa VapC	CAL70640.1.
***Mycobacterium tuberculosis* (strain ATCC 35801 / TMC 107)**	Ribonucleasa VapC	BAL64484.1.
***Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv)**	Ribonucleasa VapC	AIR13333.1.
***Mycobacterium tuberculosis* (strain ATCC 25177 / H37Ra)**	Ribonucleasa VapC	ABQ72340.1.
Mycobacterium africanum (strain GM041182)	Ribonucleasa VapC	CCC25688.1.
***Mycobacterium caprae* MB2**	Ribonucleasa VapC	CEJ50629.1.
***Mycobacterium bovis* 1595**	Ribonucleasa VapC	AKR00203.1.
***Mycobacterium tuberculosis* DK9897**	Ribonucleasa VapC	APR56075.1
***Mycobacterium tuberculosis* (strain C**	Ribonucleasa VapC	EAY59014.1
***Mycobacterium tuberculosis* BTB05-013**	toxina VapC28	KCN23112.1
***Mycobacterium tuberculosis* MAL010121**	toxina VapC28	KBH07165.1
***Mycobacterium tuberculosis* OFXR-27**	toxina VapC28	KBJ95953.1
***Mycobacterium tuberculosis* BTB05-348**	toxina VapC28	KCN32372.1
***Mycobacterium tuberculosis* M1034**	toxina VapC28	KAX84894.1
***Mycobacterium tuberculosis* T46**	toxina	EFD12157.1
***Mycobacterium tuberculosis* W-148**	toxina VapC28	EGE49374.1
***Mycobacterium bovis* BCG**	Ribonucleasa VapC	AMC49203.1.
***Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv)**	Ribonucleasa VapC28	CCP43348.1.
***Mycobacterium tuberculosis* (strain CDC 1551 / Oshkosh)**	Ribonucleasa VapC28	AAK44860.1.
***Mycobacterium bovis* (strain ATCC BAA-935 / AF2122/97)**	Ribonucleasa VapC Mb0625	CDO41867.1.
***Mycobacterium canettii* (strain CIPT 140010059)**	Ribonucleasa VapC	CCC42950.1.
***Mycobacterium gastri* 'Wayne'**	Ribonucleasa VapC	ETW26661.1.
***Mycobacterium kansasii* 732**	Ribonucleasa VapC	EUA14932.1.
***Mycobacterium tuberculosis* CAS/NITR204**	Proteína no caracterizada	AGL26090.1.
***Mycobacterium tuberculosis* 0B076XDR**	Ribonucleasa	AIH51309.1
mine drainage metagenome	Proteína no caracterizada	CBH75433.1.
***Novosphingobium* sp. AAP93**	Ribonucleasa VapC	KPF81215.1.
***Rhodovulum* sp. PH10**	Ribonucleasa VapC	EJW12600.1.
mine drainage metagenome	Proteína no caracterizada	CBI02253.1.
mine drainage metagenome	Proteína no caracterizada	CBI02899.1.
***Frankia* sp. G2**	Ribonucleasa VapC	CUU59094.1.
***Gordonia araii* NBRC 100433**	Ribonucleasa VapC	GAB09124.1.
***Blastococcus saxobsidens* (strain DD2)**	Ribonucleasa VapC	CCG03066.1.

Continuación de la Tabla suplementaria 6. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la antitoxina VapC28 de *M. llatzerense* MG13^T. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***Novosphingobium* sp. Rr 2-17**	Ribonucleasa VapC	EIZ80440.1
***Blastococcus saxobsidens* (strain DD2)**	Ribonucleasa VapC	CCG01302.1.
***Frankia* sp. ACN1ag**	Ribonucleasa VapC	KQC37972.1.
***Leptospira alstonii* serovar Pingchang str. 80-412**	Ribonucleasa VapC	EQA82603.1.
***Leptospira alstonii* serovar Sichuan str. 79601**	Ribonucleasa VapC	EMJ94433.1.
***Agromyces* sp. Soil535**	Ribonucleasa VapC	KRE22857.1.
***Frankia* sp. CpI1-S**	Ribonucleasa VapC	KJE22333.1.
***Frankia* sp. CpI1-P**	Proteína hipotética	KQM05129.1
***Luteipulveratus halotolerans* C296001**	Ribonucleasa VapC	KNX38628.1.
***Nocardia vulneris* W9851**	Ribonucleasa VapC	KIA65694.1.
***Knoellia sinensis* KCTC 19936**	Ribonucleasa VapC	KGN32866.1.
***Knoellia subterranea* KCTC 19937**	Ribonucleasa VapC	KGN37371.1.

Tabla suplementaria 7. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la proteína hipotética de la cepa *Mycobacterium* MHSD3. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***M. canettii* CIPT 140010059**	Proteína hipotética	WP_014000191.1
***M. tuberculosis* Haarlem/NITR202**	Proteína hipotética	AGL22129.1
***M. orygis* 112400015**	Proteína hipotética	EMT37385.1
***M. bovis* BCG / Pasteur 1173P2**	Proteína hipotética	CAL70390.1
***M. bovis* ATCC BAA-935 / AF2122/97**	Proteína hipotética	CDO41612.1
***M. tuberculosis* ATCC 35801 / TMC 107 / Erdman**	Proteína hipotética	WP_003401860.1
***M. africanum* GM041182**	Proteína hipotética	WP_003401860.1
***M. bovis* BCG-1**	Proteína hipotética	WP_014000191.1
***M. tuberculosis* ATCC 25618 / H37Rv**	Proteína hipotética	CCP43097.1
***M. tuberculosis* MT43**	Proteína hipotética	WP_003401860.1
***M. tuberculosis* C**	Proteína hipotética	EAY58797.1
***M. tuberculosis* BTB05-013**	Proteína hipotética	KCN22862.1
***M. tuberculosis* MAL010121**	Proteína hipotética	WP_003401860.1
***M. tuberculosis* OFXR-27**	Proteína hipotética	WP_003401860.1
***M. tuberculosis* M1034**	Proteína hipotética	WP_003401860.1
***M. tuberculosis* MD15956**	Proteína hipotética	KAQ14453.1
***M. tuberculosis* W-148**	Proteína hipotética	WP_003401860.1
***M. bovis* BCG 26**	Proteína hipotética	AMC48905.1
***M. colombiense* CECT 3035^T**	Proteína hipotética	WP_007770931.1
***M. tuberculosis* TKK-01-0051**	Proteína hipotética	KBZ62316.1
***M. avium* subsp. *paratuberculosis* 08-8281**	Proteína hipotética	WP_003876750.1
***M. avium* subsp. *avium* 2285 (R)**	Proteína hipotética	WP_009976725.1

Continuación de la Tabla suplementaria 7. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la proteína hipotética de la cepa *Mycobacterium* MHSD3. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***M. avium* XTB13-223**	Proteína hipotética	WP_009976725.1
***M. avium* MAV_120709_2344**	Proteína hipotética	WP_009976725.1
***M. avium* subsp. *hominissuis* 101**	Proteína hipotética	WP_009976725.1
***M. avium* 05-4293**	Proteína hipotética	ETA92680.1
***M. avium* 10-5560**	Proteína hipotética	ETB53215.1
***M. avium* subsp. *hominissuis* TH135**	Proteína hipotética	WP_009976725.1
***M. avium* subsp. *avium* 11-4751**	Proteína hipotética	ETB21122.1
***M. avium* subsp. *hominissuis* 100**	Proteína hipotética	WP_009976725.1
***M. avium* subsp. *avium* 10-9275**	Proteína hipotética	WP_009976725.1
***M. avium* subsp. *silvaticum* ATCC 49884**	Proteína hipotética	ETB09947.1
***M. avium* 10-5581**	Proteína hipotética	ETA97490.1
***M. avium* 104**	Proteína hipotética	ABK67732.1
***Mycobacterium* sp. MAC_080597_8934**	Proteína hipotética	WP_033710872.1
***M. paratuberculosis* ATCC BAA-968 / K-10**	Proteína hipotética	AAS04094.1
***M. avium* subsp. *hominissuis* A5**	Proteína hipotética	KDO96018.1
***M. avium* subsp. *paratuberculosis* 10-5864**	Proteína hipotética	ETB03997.1
***M. avium* subsp. *paratuberculosis* 10-4404**	Proteína hipotética	WP_003878042.1
***M. avium* subsp. *hominissuis* 10-5606**	Proteína hipotética	ETB41458.1
M. tuberculosis	Proteína hipotética	WP_031730560.1
***M. parascrofulaceum* ATCC BAA-614**	Proteína hipotética	EFG79139.1
M. heraklionense	Proteína hipotética	WP_047317834.1
***M. vaccae* ATCC 25954**	Proteína hipotética	WP_003929789.1
***M. europaeum* CSUR P1344**	Proteína hipotética	CQD16798.1
***M. gordonae* CTRI 14-8773**	Proteína hipotética	WP_055576679.1
***M. gilvum* PYR-GCK**	Proteína hipotética	ABP47663.1
***M. gilvum* DSM 45189 / LMG 24558 / Spyr1**	Proteína hipotética	ADU01179.1
***M. smegmatis* MKD8**	Proteína hipotética	WP_003894879.1
***M. avium* subsp. *avium* 2285 R**	Proteína hipotética	EUA33059.1
M. chelonae	Proteína hipotética	WP_064408866.1
***M. chelonae* 203**	Proteína hipotética	OHT78135.1

Tabla suplementaria 8. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la toxina zeta de la cepa *Mycobacterium* MHSD3. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***M. parascrofulaceum* ATCC BAA-614**	Proteína hipotética	EFG79138.1
***Mycobacterium* sp. KMS**	ATPasa	WP_011560054.1
***M. europaeum* CSUR P1344**	Toxina zeta	CQD16793.1
***Mycobacterium* sp. JS623**	Proteína hipotética	AGB26847.1
***M. orygis* 112400015**	Proteína hipotética	EMT37384.1
***M. bovis* BCG / Pasteur 1173P2**	Proteína hipotética	CAL70389.1
***M. tuberculosis* CDC 1551 / Oshkosh**	Proteína hipotética	WP_003401859.1
***M. bovis* ATCC BAA-935 / AF2122/97**	Proteína hipotética	CDO41611.1
***M. tuberculosis* ATCC 35801 / TMC 107 / Erdman**	Proteína hipotética	WP_003401859.1
***M. tuberculosis* ATCC 25177 / H37Ra**	Proteína hipotética	WP_003401859.1
***M. africanum* GM041182**	Proteína hipotética	WP_003401859.1
***M. caprae* MB2**	Proteína hipotética	WP_003401859.1
***M. bovis* 1595**	ATPasa	AKQ99931.1
***M. tuberculosis* ATCC 25618 / H37Rv**	Proteína hipotética	CCP43096.1
***M. tuberculosis* MT43**	Proteína hipotética	WP_003401859.1
***M. tuberculosis* C**	Proteína hipotética	EAY58796.1
***M. tuberculosis* BTB05-013**	Proteína hipotética	KCN22861.1
***M. tuberculosis* MAL010121**	Proteína hipotética	WP_003401859
***M. tuberculosis* OFXR-27**	Proteína hipotética	WP_003401859
***M. tuberculosis* M1034**	Proteína hipotética	WP_003401859
***M. tuberculosis* MD15956**	Proteína hipotética	KAQ14452.1
***M. tuberculosis* W-148**	Proteína hipotética	WP_003401859
***M. bovis* BCG 26**	Proteína hipotética	AMC48904.1
***M. canettii* CIPT 140010059**	ATPasa	WP_014000190
***M. mageritense* DSM 44476 = CIP 104973**	ATPasa	WP_036442085
M. tuberculosis	Proteína ZTL	COV58376.1
***M. chimaera* MCIMRL6**	ATPasa	WP_054585353
***Mycobacterium* sp. 05-1390**	ATPasa	WP_014711570
***Mycobacterium* sp. TKK-01-0059**	ATPasa	WP_014711570
***Mycobacterium* sp. MOTT36Y**	ATPasa	WP_014711570
***M. phlei* RIVM601174**	ATPasa	WP_003888374
***M. tuberculosis* TKK-01-0051**	ATPasa	WP_044485837
***M. vaccae* ATCC 25954**	ATPasa	WP_003929790
***M. colombiense* CECT 3035^{T}**	ATPasa	WP_007770932

Continuación de la Tabla suplementaria 8. Números de acceso de las secuencias de proteínas utilizadas en la construcción del dendograma de la toxina zeta de la cepa *Mycobacterium* MHSD3. Se indica la cepa de procedencia y la anotación con la que cada proteína aparece en las bases de datos.

Cepa	Anotación	Número de acceso
***M. conceptionense* MLE**	ATPasa	KMV19162.1
***M. senegalense* CK1**	ATPasa	WP_019346252
***Rhodococcus opacus* R7 plasmid pPDG1**	ATPasa	AII10527.1
***M. farcinogenes* DSM 43637**	ATPasa	WP_036387328
***M. vulneris* ACS5020**	ATPasa	OCB45067.1
***M. gilvum* DSM 45189 / LMG 24558 / Spyr1**	ATPasa	WP_013472824
M. neworleansense	Toxina zeta	CRZ14295.1
***M. gilvum* PYR-GCK**	ATPasa	WP_011896020
***M. smegmatis* ATCC 700084 / mc2155**	ATPasa	YP_887740.1
***Mycobacterium*. sp. MAC_080597_8934**	ATPasa	WP_003876749
***M. paratuberculosis* ATCC BAA-968 / K-10**	ATPasa	AAS04093.1
***M. avium subsp. hominissuis* 101**	ATPasa	WP_003876749.1
***M. avium* subsp. *avium* 2285 R**	ATPasa	WP_003876749.1
***M. avium* XTB13-223**	ATPasa	WP_003876749.1
***M. avium* subsp. *hominissuis* TH135**	ATPasa	WP_003876749.1
***M. avium* subsp. *avium* 11-4751**	ATPasa	ETB21123.1
***M. chelonae* 1558**	ATPasa	WP_070917639.1
***M. chelonae* 15517**	ATPasa	WP_070917639.1
***M. chelonae* 15518**	ATPasa	WP_070917639.1
***M. chelonae* 203**	Proteína hipotética	WP_070921490.1
***Mycobacterium* sp. QIA-37**	ATPasa	WP_064408867.1

Tabla suplementaria 9. Aminoácidos del centro activo de las proteínas con dominio PIN utilizadas. Asp: aspartato, Glut: glutamina, Gly: glicina, Asn: asparagina.

Cepa	Proteína	Aminoácidos del centro activo
***Pyrobaculum aerophilum* ATCC 51768**	Proteína con dominio PIN	Asp8, Glut38, Asp92, Asp110
***Archaeoglobus fulgidus* DSM 4304**	Proteína con dominio PIN	Asp26, Glut57, Asp115, Asp133
***M. tuberculosis* H37Rv**	VapC5	Asp26, Glut57, Asp115, Asp133
***M. llatzerense* MG13^{T}**	VapC27	Asp10, Glut46, Asp103, Asp121
***M. llatzerense* MG13^{T}**	VapC28	Asp4, Glut40, Asp100, Gly118
CR-UIB1	VapC5	Asp6, Glut31, Asp92, Asn110

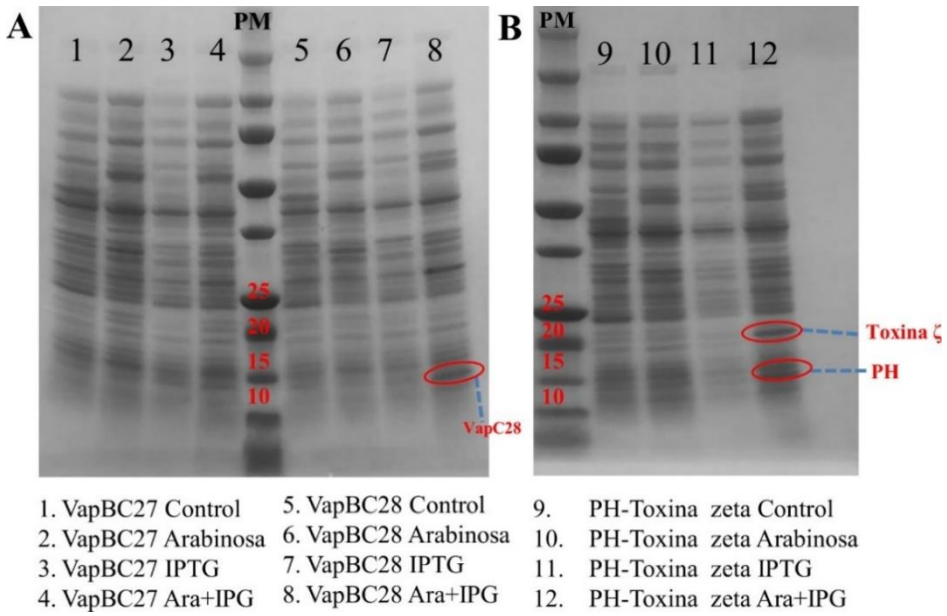


Figura suplementaria 1. Patrón de bandas proteicas obtenidos para los sistemas VapBC27 y VapBC28 (A) y el sistema HP y Toxina zeta (B) . Se destacan en rojo los pesos moleculares (en kDa) referentes a la zona donde se enmarcan las proteínas de interés. Se destacan también las bandas potencialmente representativas de las toxinas VapC28 y zeta, así como la proteína hipotética relacionada con esta última**.**

Figura suplementaria 2. Alineamiento obtenido a partir de la superposición de las estructuras terciarias. Entre paréntesis se indican el número de aminoácidos que no se superponen entre los distintos bloques alineados. Se resaltan aquellos aminoácidos conservados del centro activo en amarillo, y en rojo los aminoácidos que corresponden a la cuarta posición del mismo, no conservados en las toxinas VapC28 y VapC5.

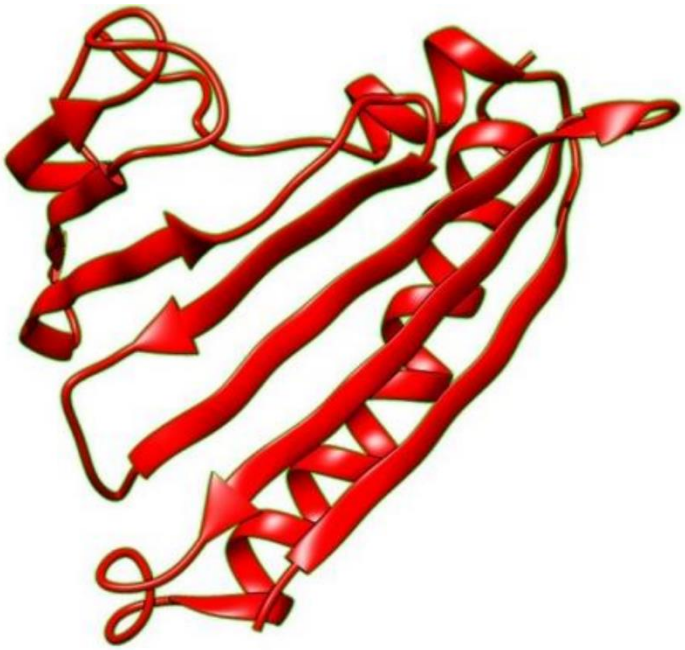


Figura suplementaria 3. Predicción estructural realizada por I-TASSER a partir de la secuencia de aminoácidos de la hipotética toxina MT0934 del par MT0933-34 de *M. llatzerense* MG13^{T}.