UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

ESTUDIO DE LA INTERACCIÓN PATÓGENO-HOSPEDERO EN UN MODELO MURINO DE TUBERCULOSIS PULMONAR MEDIANTE EL ANÁLISIS DEL TRANSCRIPTOMA Y EL SECRETOMA.

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Doctora en Ciencias

PRESENTA:
M.C. MARÍA FERNANDA CORNEJO GRANADOS

TUTOR PRINCIPAL

DR. ADRIÁN OCHOA LEYVA
Instituto de Biotecnología. Universidad Nacional Autónoma de México

MIEMBROS DEL COMITÉ TUTOR

DR. ROGELIO HERNÁNDEZ PANDO
Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán

DR. ALEJANDRO GARCIARUBIO GRANADOS
Instituto de Biotecnología. Universidad Nacional Autónoma de México

Cuernavaca, Morelos. Septiembre, 2021

El presente trabajo se realizó bajo la dirección del **Dr. Adrián Ochoa Leyva** en el **Departamento de Microbiología Molecular** del **Instituto de Biotecnología** de la **Universidad Nacional Autónoma de México.**

Para su realización, contó con el financiamiento del Consejo Nacional de Ciencia y Tecnología (CONACyT) con la beca de doctorado (CVU 443238) y el programa Ciencia Frontera 2019-263986. Así mismo contó con financiamiento de la Universidad Nacional Autónoma de México con el programa DGAPA PAPIIT UNAM IN215520 y del Programa de Apoyo a los Estudios de Posgrado (PAEP).

Miembros del Jurado

PRESIDENTE
Dr. Francisco Xavier Soberón Mainero

SECRETARIO
Dr. Fidel Alejandro Sánchez Flores

VOCAL
Dra. Clara Inés Espitia Pinzón

SUPLENTE
Dra. Bertha Josefina Espinoza Gutiérrez

SUPLENTE
Dra. Blanca Itzel Taboada Ramírez

Agradecimientos

Al Dr. Adrián Ochoa Leyva por ser guía y amigo durante estos años. Gracias por el apoyo, la confianza y la oportunidad de participar en los proyectos de investigación del laboratorio. Su ayuda ha sido esencial en estos años de formación.

Al Dr. Rogelio Hernández Pando por todo el apoyo, confianza y consejos desde la maestría. Al Dr. Alejandro Garciarubio por todos los comentarios y contribuciones al proyecto durante los tutorales.

Al Técnico Académico Biól. Filiberto Sánchez por sus invaluables contribuciones al proyecto. Al Técnico Académico M. en T.I. Juan Manuel Hurtado Ramírez por toda su ayuda en el área informática del trabajo. A los Doctores Dulce Mata, Brenda Marquina y Jorge Barrios por su ayuda y amistad durante estos años.

A todos mis compañeros del laboratorio 22 por su compañía, enseñanzas y ayuda durante estos años. A Ere Ángeles y Valeria Yescas por su disposición y ayuda en el área administrativa para que este proyecto saliera adelante.

A todos los miembros del jurado por sus invaluables comentarios y correcciones a la tesis. No sólo enriquecieron el trabajo, también contribuyeron a mi aprendizaje y formación.

A mis papás. Sin ustedes no lo podria haber logrado.

A mi hermana. Eres un motor enorme en mi vida.

Índice

Resumen 1

I. Introducción. 1

Variabilidad genética y su asociación con la virulencia de las diferentes cepas. 2

Expresión génica y su asociación con la virulencia de las diferentes cepas. 3

Modelo murino para el estudio de la tuberculosis pulmonar. 4

Importancia de las proteínas secretadas en la infección por *M. tuberculosis*. 5

II. Antecedentes. 1

III. Hipótesis. 1

IV. Objetivos 1

Objetivo General 1

Objetivos Específicos 1

V. Metodología 2

V.I Secuenciación y ensamble de la cepa infecciosa de *M. tuberculosis.* 2

V.II Transcriptoma de *M. tuberculosis* 3

1. Cultivos de bacteria. 3

2. Infección de ratones. 3

3. Extracción de ARN y uso de sondas de captura de ribosomales para obtener el transcriptoma de *M. tuberculosis.* 4

4. Análisis de datos del transcriptoma de *M. tuberculosis*. 7

V.III Transcriptoma de ratón. 8

1. Extracción de ARN. 8

2. Construcción de librerías de ARN para secuenciación. 8

3. Análisis de datos del transcriptoma de ratón. 9

V.IV Secretoma de M. tuberculosis y M. abscessus. 9

1. Aislados clínicos. 9

2. Secuenciación y ensamble de los genomas. 10

3. Predicción del secretoma. 11

4. Anotación funcional y análisis comparativo del secretoma. 11

5. Cálculo de la Abundancia de Regiones Antigénicas. 12

6. Construcción del secretoma experimental de *M. tuberculosis.* 12

VI. Resultados 14

VI.I RNA-seq para determinar el transcriptoma de *M. tuberculosis* en un modelo *in vivo* de tuberculosis pulmonar. 14

1. Armado *de novo* del genoma de *M. tuberculosis*. 14

2. Estrategias experimentales para obtener el transcriptoma de *M. tuberculosis* en un modelo *in vivo* al día 21 post infección. 14
3. Transcriptoma de *M. tuberculosis* en un modelo *in vivo* al día 21 post infección. 19
4. Comparación del secretoma y transcriptoma de *M. tuberculosis* al día 21 post-infección. ..25
5. Transcriptoma de ratón en un modelo in vivo al día 21 post infección. 26
VI.II Predicción del secretoma de *M. tuberculosis* y *M. abscessus*. 28
VI.III Implementación del servidor web Secret-AAR para la búsqueda de homología entre proteínas del secretoma de micobacterias y para determinar la densidad antigénica de proteínas. 32
1. Análisis del secretoma....................32
2. Determinación de la Abundancia de Regiones Antigénicas (AAR). 32
VII. Discusión....................34
VIII. Conclusiones....................40
IX. Perspectivas....................40
IX. Referencias 41
X. Anexos 47

Índice de Figuras

Figura 1. Desarrollo clínico de la infección por Mtb en humanos....................1

Figura 2. Clasificación de Mtb en seis linajes. 3

Figura 3. Importancia del secretoma en la interacción patógeno-hospedero. 6

Figura 4. Diseño experimental general para obtener el transcriptoma de *M. tuberculosis* y su hospedero a partir de pulmones de ratón infectados. 5

Figura 5. Genoma ensamblado de Mtb....................15

Figura 6. Estrategias experimentales para obtener el perfil de expresión de Mtb (E) y de ratón (R) a partir de pulmones infectados. 16

Figura 7. Tinción Zihel-Neelsen del homogenado de pulmón antes y después de la lisis celular diferencial.. 17

Figura 8. Procedimiento experimental detallado de la Estrategia 3 (E3) para obtener el transcriptoma de Mtb. 18

Figura 9. Estadística de uso del servidor Secret-AAR 33

Índice de Tablas

Tabla I. Top 20 dominios y familias de InterPro significativamente sobrerrepresentadas en los 529 genes expresados de Mtb. 20

Tabla II. Top 10 de términos GO por categoría significativamente enriquecidos en los 529 genes más expresados de Mtb durante la infección in vivo. 21

Tabla III. Vías KEGG más representadas en los 529 genes expresados de Mtb. 22

Tabla IV. Top 20 de genes con respecto a los 529 genes expresados de Mtb. 23

Tabla V. Top 20 de genes con respecto a los 101 genes expresados de Mtb. 25

Tabla VI. Secretomas predichos de las cepas de referencia de *M. tuberculosis* y *M. abscessus.* ...29

Resumen

Las técnicas de secuenciación masiva como el RNA-seq para estudiar la infección por *M. tuberculosis* en un modelo *in vivo* presentan retos como i) elegir un sistema de lisis para enriquecer el ARN del patógeno, y ii) lograr una depleción de transcritos ribosomales eficiente. Por ello, algunos autores han usado la citómetria de flujo para separar las células infectadas o han aumentado la profundidad de secuenciación de las muestras. Sin embargo, estas opciones conllevan gastos adicionales que pueden no estár disponibles para todos los grupos de investigación. En esta tesis presentamos un protocolo con lisis celular diferencial y sondas de captura de ARN ribosomal que permitió observar 702 genes totales de *M. tuberculosis* y tres ARNs no codificantes (MTS2823, RF00023 y RF00010) expresados al día 21 post-infección en un modelo murino. Entre los genes más expresados observamos la transposasa para IS1081 (Rv2512c) y ocho genes para proteínas de la familia PE-PGRS. Además, 16.07% de los genes expresados codificaban para proteínas secretadas como PE68, lppN y lpqH.

Asimismo presentamos una estrategia bioinformática para predecir el secretoma y la abundancia de regiones antigénicas de diferentes genomas de *M. tuberculosis* y *M. abscessus.* Observamos que para *M. tuberculosis* el secretoma representaba el ~12% de las proteínas totales, mientras que para *M. abscessus* representaba el ~18%. Además, los secretomas de *M. tuberculosis* con genotipo Beijing y de *M. abscessus* con morfotipo rugoso mostraron una mayor abundancia de regiones antigénicas. Finalmente, comparamos el secretoma de *M. tuberculosis* contra 338 proteínas reportadas experimentalmente como secretadas y observamos que el ~70% de estas proteínas estaba incluido en el secretoma predicho, mientras que dos publicaciones previas sólo coincidían en un ~34% y ~41%.

Todos los secretomas predichos están disponibles en el servidor web Secret-AAR, a través de dos herramientas: i) identificación (vía BLASTp) si una proteína de interés pertenece a un secretoma predicho o experimental, y ii) predicción de la abundancia de regiones antigénicas de cualquier secuencia de aminoácidos.

En conclusión, la estrategia experimental mostró eficacia para estudiar los genes y ARN no codificantes de *M. tuberculosis* más expresados durante la infección *in vivo*. Y la estrategia bioinformática permitió predecir de manera sistemática el secretoma y el potencial antigénico de *M. tuberculosis* y *M. abscessus*. Ambas estrategias, pueden ser utilizadas para el estudio de otras cepas y otros días de la infección.

I. Introducción.

La tuberculosis (Tb) es una enfermedad infecciosa considerada una de las diez principales causas de muerte alrededor del mundo (1).

Esta enfermedad es causada por el bacilo intracelular *Mycobacterium tuberculosis (*Mtb) el cual ingresa a través de la vía respiratoria por la inhalación de aerosoles provenientes de un individuo infectado. Una vez dentro del sistema respiratorio, se establece preferentemente en los pulmones (tuberculosis pulmonar) o puede diseminarse a otros órganos del cuerpo (tuberculosis extra pulmonar) (Figura 1) (2).

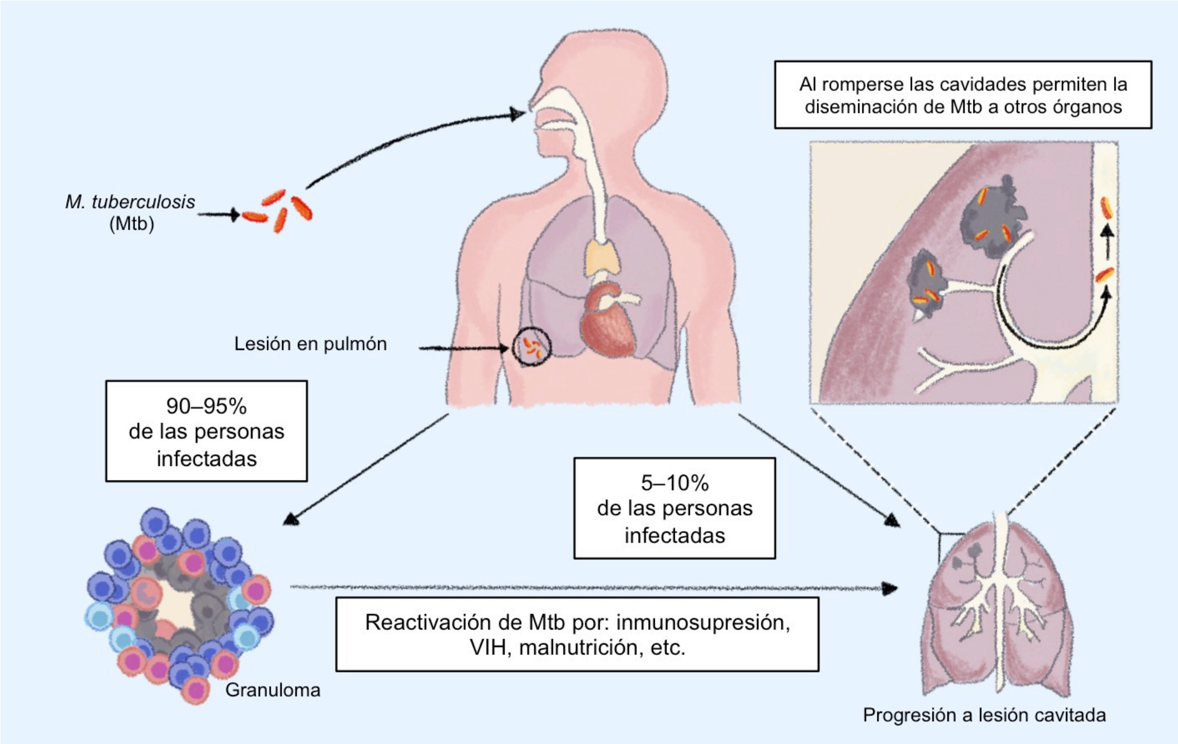

M. tuberculosis
(Mtb)
Al romperse las cavidades permiten la diseminación de Mtb a otros órganos
Lesión en pulmón
90–95% de las personas infectadas
5–10% de las personas infectadas
Reactivación de Mtb por: inmunosupresión, VIH, malnutrición, etc.
Granuloma
Progresión a lesión cavitada

Figura 1. Desarrollo clínico de la infección por Mtb en humanos. El bacilo ingresa por vía aérea a la persona sana estableciéndose principalmente en los pulmones. De allí, el 90-95% de las personas contienen a la bacteria dentro de los granulomas manteniendo la infección latente, la cual, se puede reactivar en condiciones de inmunosupresión, infecciones por VIH, malnutrición, etc. Por otro lado, el 5-10% de las personas desarrollan desde el inicio una infección activa, la cual puede causar lesiones cavitadas en los pulmones y la diseminación de la bacteria a otros órganos. (modificada de Rook et al. (3))

Una vez que el bacilo entra en contacto con el organismo hospedero, se pueden presentar varios escenarios. Entre el 5-10% de las personas presentan una enfermedad activa y transmisible dentro de los dos primeros años de la infección, la cual puede presentar de síntomas leves a muy severos. Por otro lado, el 90-95% de las personas infectadas mantienen al bacilo latente, conteniéndolo en una cápsula de células del sistema inmune llamada granuloma. Este estado latente, el cual se estima que afecta a cerca de un cuarto de la población mundial (1, 4) puede permanecer durante toda la vida de la persona o reactivarse si la eficiencia del sistema inmune disminuye por desnutrición, medicamentos, infección por VIH u otras enfermedades (1, 2, 5) (Figura 1).

Actualmente existen fármacos efectivos contra la Tb como la isoniazida, rifampicina, etambutol y pirazinamida considerados antibióticos de primera línea, y levofloxacin, bedaquiline y delamanid como fármacos de segunda línea. Sin embargo, estos medicamentos necesitan administrarse por largos períodos (6 meses o más) y conllevan varios efectos secundarios como toxicidad al hígado, lo que ha provocado la falta de adherencia a los tratamientos y ha favorecido el desarrollo de cepas resistentes. Por ejemplo, en el 2019 se calculó que el 3.3% de los casos nuevos y el 17.7% de las recaídas de Tb a nivel mundial tenían resistencia a algún medicamento (1, 2).

Variabilidad genética y su asociación con la virulencia de las diferentes cepas.

Tradicionalmente, los diferentes escenarios clínicos de la tuberculosis habían sido atribuidos a factores ambientales y del hospedero (6). Sin embargo, con la publicación del genoma de Mtb (7) y el desarrollo de diferentes técnicas de genotipificación se reveló la variabilidad genética que existe entre las cepas, estableciéndose una clasificación en diferentes linajes nombrados de acuerdo a su prevalencia alrededor del mundo (8) (Figura 2).

Posteriormente, en los estudios con células *in vitro* y modelos animales se observaron diferencias en el desarrollo de la enfermedad entre cepas de diferentes linajes (9-13) e incluso entre cepas del mismo linaje (14). Sin embargo, al intentar relacionar la virulencia de las cepas con características del hospedero como su capacidad de respuesta inmune, hay observaciones contradictorias (10, 13, 15).

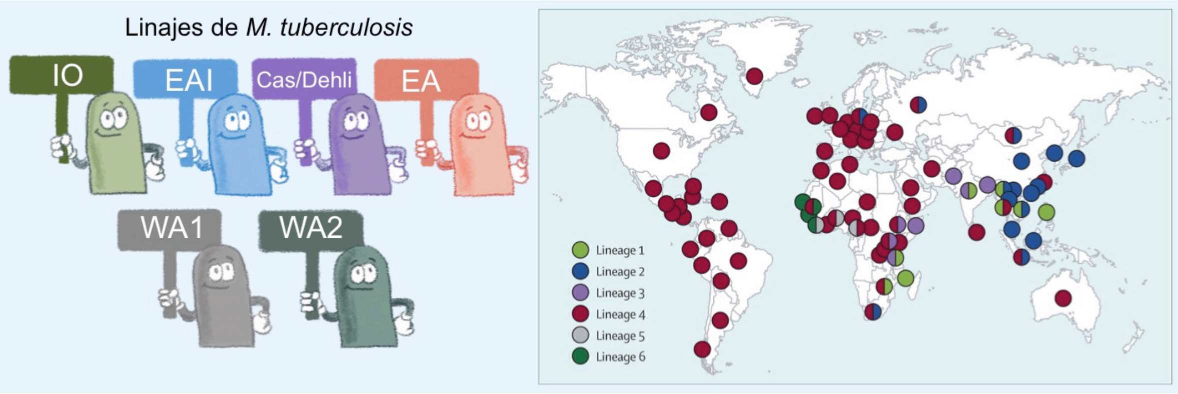


Figura 2. Clasificación de Mtb en seis linajes. La clasificación se ha realizado de acuerdo a características genéticas y nombrados por su distribución alrededor del mundo (mapa tomado de Gagneux et al. (16))

Después, estudios dirigidos empezaron a asociar mutaciones en los genomas de Mtb con las diferencias fenotípicas de las diversas cepas. Por ejemplo, se observó que mutaciones no sinónimas en el sistema PhoPR alteraban la síntesis de lípidos de membrana y la secreción del antígeno ESAT-6, lo que resulta en cepas menos virulentas (17). Además, se encontró que los polimorfismos de nucleótido único (SNPs) sinónimos río arriba del gen DosR generaban sitios alternos de inicio de la transcripción en cepas del linaje 2, lo que parece aumentar su virulencia (18).

Por otro lado, con estudios masivos de genómica comparativa se ha observado que entre dos cepas de Mtb existe una diferencia promedio de 1,200 SNPs (0.03% del genoma), mientras que entre dos especies del género Mycobacterium, por ejemplo, Mtb y *M. canetti* hay una diferencia del 2.7% de su genoma (19, 20).

Expresión génica y su asociación con la virulencia de las diferentes cepas.

Diversos estudios de expresión génica han utilizado microarreglos para identificar los genes sobre expresados de Mtb en macrófagos vs cultivos líquidos (21-23), o genes inducidos en macrófagos inmediatamente después de fagocitar a la bacteria (24-26). Sin embargo, estas aproximaciones utilizando modelos *in vitro* y cultivos líquidos no consideran muchos de los factores que participan en una infección *in vivo*.

Actualmente, sólo hay dos estudios sobre la expresión génica de Mtb durante una infección *in vivo.* En el estudio de Talaat et al. (22) utilizaron microarreglos para comparar la expresión génica de la cepa H37Rv en pulmones de ratones BALB/c, ratones con inmunodeficiencia combinada severa (SCID), y cultivos líquidos en cuatro tiempos de la infección. Y en el estudio de Pisu et al. (24) utilizaron la citometría de flujo, para aislar macrófagos alveolares e intersticiales de ratones infectados con la cepa Erdman y compararon la expresión génica de Mtb y su hospedero entre los dos tipos celulares utilizando RNA-seq.

Modelo murino para el estudio de la tuberculosis pulmonar.

Existen diversos estudios en modelos animales para el estudio de la tuberculosis pulmonar. Uno de ellos, es el desarrollado por el Dr. Rogelio Hernández-Pando en el Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán en la Ciudad de México (27), en el cual se puede observar la evolución de la enfermedad en dos fases discernibles, la fase temprana (aguda) y la fase crónica (progresiva).

Brevemente, el modelo se basa en ratones machos BALB/c de 6-8 semanas de edad (22 g de peso aproximadamente), a los cuales se les introducen bacilos vivos de Mtb por vía intratraqueal a una dosis aproximada de 250,000 bacterias suspendidas en 100 µL de solución salina.

Durante las primeras tres semanas de infección se observa la fase temprana, caracterizada por una alta producción de citocinas proinflamatorias tipo Th1 (e.g. IL-2, INFγ, IL-12) que junto con niveles altos de iNOS activan a los macrófagos que fagocitan a la bacteria y limitan el desarrollo de la infección (27, 28). En el tejido pulmonar, se observa infiltrado inflamatorio formado principalmente por macrófagos y linfocitos, y hacia el día 21 post-infección, se observan los granulomas completamente formados por el aglomerado de células del sistema inmune que rodean a las bacterias dentro de los macrófagos (27, 28).

Después de las tres semanas, la expresión de las citocinas tipo Th1 decae y aumenta la cantidad de IL-4 y de factor de crecimiento transformante beta (TGFß), además comienza un aumento gradual de la neumonía en el tejido pulmonar que domina sobre los granulomas. El daño tisular junto con la alta carga bacteriana, son lo que eventualmente lleva a la muerte de los ratones (12, 27).

El curso de la enfermedad en este modelo, presenta varias similitudes con el desarrollo clásico en humanos. Por ejemplo, se basa en una vía de infección aérea, que es la vía natural de entrada del bacilo, hay un control inicial de la infección por macrófagos activados y la formación de granulomas, se presenta un claro perfil pro-inflamatorio (citocinas Th1) en la fase temprana y una progresión hacia neumonía con un perfil antiinflamatorio en la fase crónica (27).

Importancia de las proteínas secretadas en la infección por *M. tuberculosis*.

El secretoma, es el conjunto de proteínas excretadas o secretadas por una célula. Estas participan en procesos celulares importantes como la adhesión, comunicación, migración e invasión de tejidos (Figura 3) (29).

Para patógenos intracelulares como *M. tuberculosis*, el secretoma juega un papel importante para su sobrevivencia y proliferación dentro de los macrófagos*,* además de ser antígenos clave para la interacción con el hospedero y el desarrollo de la respuesta inmune protectora (30-33). Por ejemplo, antígenos muy estudiados en Mtb como el Ag85 es secretado por vía clásica y estimula las células T del hospedero (34, 35), además los sistemas de secreción tipo VII (ESX1-ESX5) son grupos de genes que codifican para proteínas de la familia PE y PPE y se encargan de la secreción de proteínas altamente antigénicas como ESAT-6 y CFP-10, las cuales carecen de péptido señal (36) y están ampliamente distribuidos dentro del género Mycobacterium, tanto en especies de crecimiento lento como Mtb como de crecimiento rápido como Mabs (37, 38).

Dada su exposición al medio extracelular, las proteínas secretadas se consideran importantes para la virulencia en las diferentes cepas de Mycobacterium (33, 39). Por ejemplo, la comparación entre las proteínas de filtrados de cultivo celular mostró que cepas hipervirulentas de Mtb tienen una mayor expresión de proteínas involucradas en el metabolismo de lípidos (FadA3, FbpB y EchA3), adaptación y detoxificación (GroEL2, SodB y HspX), y formación de la pared celular (LprA, Tig y EsxB) (40) que las cepas hipovirulentas. Así mismo, se han observado diferencias en las proteínas secretadas entre cepas resistentes y sensibles a antibióticos (41). Sin embargo, dada la complejidad de la pared celular de las micobacterias, el aislamiento y estudio del secretoma es limitado y aún no es claro cuáles son las diferencias en las proteínas secretadas que podrían favorecer el desarrollo de los diferentes escenarios durante la infección *in vivo*.

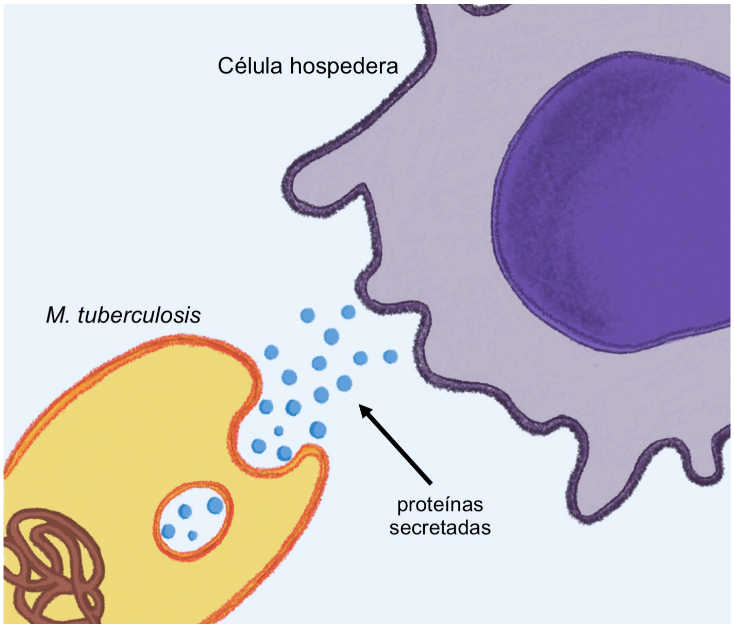


Figura 3. Importancia del secretoma en la interacción patógeno-hospedero. Proteínas secretadas de Mtb como Ag85, ESAT-6 y CFP10 son altamente antigénicas y estimulan la respuesta del sistema inmune en el hospedero.

En el presente trabajo describimos una estrategia experimental utilizando lisis celular diferencial y sondas de captura de ribosomales para obtener el transcriptoma de Mtb y su hospedero al día 21 post infección, el cual, representa el punto máximo de expresión de la respuesta inmune del hospedero. Así mismo, planteamos una estrategia bioinformática que utiliza cuatro predictores especializados en identificar proteínas secretadas por diferentes vías. Con ella, predijimos el secretoma de la cepa de referencia de Mtb H37Rv y dos aislados clínicos del genotipo Beijing, el cual es de gran interés por su alta virulencia y transmisibilidad alrededor del mundo (12-14). También analizamos el genoma de referencia de Mabs ATCC19977*,* y 15 aislados clínicos que incluyen a las tres subespecies de Mabs (*M. abscessus* supsp. *abscessus*, *M. abscessus* subsp. *masiliense* y *M. abscessus* subsp. *bolletii*). De todos los secretomas obtenidos se analizó su perfil funcional y el potencial antigénico.

II. Antecedentes.

El análisis transcriptómico de Mtb, usando técnicas de secuenciación masiva como el RNA-seq en modelos *in vivo* conlleva varias limitaciones, i) seleccionar un sistema de lisis efectivo que permita separar las células del hospedero y del patógeno (42, 43); ii) obtener ARN suficiente y de alta calidad (21, 42) y iii) lograr una disminución eficiente del ARN ribosomal (21, 42).

Para afrontar estas limitaciones, algunos autores han optado por estudiar cultivos de bacterias, células infectadas *in vitro* (22, 24, 44), usar marcas fluorescentes en las bacterias para facilitar la separación de las células infectadas (42, 45), amplificar selectivamente el ARN o ADNc de Mtb o aumentar la cantidad de bacterias en la infección para extraer suficiente ARN bacteriano (46). Sin embargo, estas estrategias no consideran la enorme cantidad de factores involucrados en un modelo *in vivo* o tienen el potencial de alterar el curso típico de la infección modificando el perfil de expresión génica.

A la fecha, sólo se han reportado dos estudios que buscan describir el transcriptoma de Mtb durante una infección *in vivo* en un modelo murino. En el estudio de Talaat et al.(22), extrajeron el ARN total de un grupo de pulmones de ratones infectados y usaron microarreglos para comparar la expresión de genes de Mtb H37Rv durante la infección en pulmones de ratones BALB/c, ratones con inmunodeficiencia combinada severa (SCID), y cultivos líquidos. Observaron que al día 21 post-infección hay un mayor número de genes sobre expresados exclusivamente en los ratones, tanto BALB/c como SCID, sugiriendo que hay una respuesta al sistema inmune del hospedero. Entre ellos, estaban genes de metabolismo y regulación de hierro, metabolismo y degradación de lípidos. En general, esta estrategia permitió describir la expresión génica de Mtb tanto en ratones como en cultivo, sin embargo algunas de las desventajas es que los microarreglos requieren de una gran cantidad de ARN (10μg) y tienen un rango dinámico limitado para discernir entre diferentes niveles de expresión (47).

Por otro lado, en el estudio de Pisu et al. (24) utilizaron citometría de flujo para aislar macrófagos alveolares e intersticiales de ratones infectados con la cepa Erdman y compararon la expresión génica de Mtb y su hospedero entre los dos tipos celulares utilizando RNA-seq. Observaron que los macrófagos alveolares expresaban genes que promovían un ambiente más permisivo para la multiplicación de las micobacterias. Por ejemplo, estos macrófagos mostraron una alta expresión de genes para la regulación de estrés oxidante, biosíntesis de ácidos grasos y colesterol, y fosforilación oxidativa; mientras que las micobacterias sobre expresaron genes de crecimiento, división celular y remodelación de pared celular. En contraste, los macrófagos intersticiales mostraron genes de respuesta inflamatoria que favorecen un ambiente más hostil, y las micobacterias expresaban genes de respuesta a estrés como el operón dosR.

Si bien marcar a las bacteria y aislar diferentes tipos de células infectadas permite enfocar el estudio de la infección, esta metodología requiere de equipos y técnicas especializadas como la citometría de flujo que no están disponibles para todos los grupos de investigación. Además, en este estudio los autores realizaron varias rondas de secuenciación para alcanzar un millón de lecturas de Mtb (alrededor de 100 millones de secuencias totales por muestra), lo que conlleva gastos adicionales en servicios de secuenciación.

Por otro lado, dada la importancia de las proteínas secretadas, actualmente hay varios estudios que buscan determinar de manera experimental el secretoma de micobacterias patógenas para el humano como Mtb y Mabs utilizando técnicas tradicionales como geles de electroforesis en dos dimensiones, cromatografía líquida y espectrometría de masas (33, 48). Sin embargo, la complejidad de la pared celular de las micobacterias, compuesta por grandes cantidades de ácidos micólicos y peptidoglicanos ha dificultado estos análisis (48).

Una manera de sobrepasar estas limitaciones es utilizando estrategias bioinformáticas que permitan predecir sistemáticamente el secretoma a partir de genomas secuenciados. En este sentido, hay dos estudios que predicen el secretoma de Mtb de manera bioinformática. El primero de Vizcaíno et al. (2010) (49) utiliza de manera independiente cinco predictores de localización celular para determinar las proteínas superficiales y secretadas de Mtb. Esto resultó en 825 proteínas secretadas, 8 de las cuales fueron analizadas experimentalmente para comprobar su secreción. En el segundo estudio, Roy et al. (2013) (50) utilizaron cuatro predictores para determinar las proteínas de Mtb secretadas por la vía clásica, lo que resultó en 267 proteínas, 46 de las cuales tienen confirmación experimental de ser secretadas en otro estudio previo. Sin embargo, estos estudios no hacen una anotación funcional, ni analizan el potencial antigénico de los secretomas. Tampoco hay estudios que aborden el tema del secretoma y su antigenicidad para otras micobacterias como Mabs.

III. Hipótesis.

M. tuberculosis expresa una alta proporción de genes para proteínas secretadas involucradas en la interacción con su hospedero al día 21 post-infección. Y dadas las limitaciones técnicas que existen, el uso de la lisis celular diferencial y sondas biotiniladas de captura de ribosomales permitirán observar los genes más expresados de *M. tuberculosis* en un modelo murino *in vivo* al día 21 post-infección.

IV. Objetivos

Objetivo General

Obtener el transcriptoma y el secretoma de *M. tuberculosis* para explorar los mecanismos de infección en un modelo *in vivo* de tuberculosis pulmonar.

Objetivos Específicos

1. Obtener el genoma de la cepa infecciosa de *M. tuberculosis*.

2. Obtener el transcriptoma de *M. tuberculosis* y su hospedero durante la infección *in vivo* al día 21 post-infección en un modelo murino de tuberculosis pulmonar.

3. Determinar si los genes observados en el transcriptoma corresponden a proteínas secretadas y explorar su posible papel en la interacción patógeno-hospedero.

4. Determinar y caracterizar de manera bioinformática, el secretoma de dos de las especies de micobacterias más relevantes para la salud humana, *M. tuberculosis* H37Rv y *M. abscessus* ATCC19977, y compararlo con el secretoma de aislados clínicos de ambas especies.

5. Predecir la antigénicidad de los secretomas obtenidos y asociarla con su virulencia.

V. Metodología

V.I Secuenciación y ensamble de la cepa infecciosa de *M. tuberculosis.*

Se extrajo el ADN total de Mtb con el kit Quick-DNA Fecal/Soil Microbe Miniprep (Zymo Research, CA, EUA, Cat. D6010) a partir de una de las alícuotas del cultivo utilizado para infectar a los ratones y siguiendo las recomendaciones del fabricante. La cantidad y la calidad del ADN extraido se determinó con gel de agarosa y fluorómetro Qubit (Invitrogen, CA, EUA, Cat. Q32851) respectivamente.

Después, se contruyó una librería para secuenciación con el Kit Nextera XT DNA Library Preparation Kit (Illumina, CA, EUA, Cat. FC-131-1024) siguiendo las recomendaciones del fabricante y seleccionando un tamaño de inserto de 400-600 pb. La librería final se cuantificó con fluorómetro Qubit y la distribución de tamaños se determinó con bioanalizador para ADN. Esta librería se secuenció en una celda MiSeq v2 (Illumina Cat. MS-102-2003) de 300 ciclos en formato pareado en el Instituto Nacional de Medicina Genómics (INMEGEN) en la Ciudad de México.

El total de lecturas crudas se filtraron con el programa FASTX-toolkit (v0.0.13) estableciendo una calidad mínima >Q20 y removiendo los adaptadores. Después, se utilizó el programa SPADES (v3.15.2) (51) y MEDUSA (52) para construir el ensamble de novo. Adicionalmente se realizó un anális para evaluar el "completness" del genoma en el web server gVolante (https://gvolante.riken.jp/index.html) (53). Brevemente este servidor integra los métodos de anális de otras dos plataformas, CEGMA (Core Eukaryotic Genes Mapping Approach) y BUSCO (Benchmarking Universal Single Copy Orthologs) y compara el ensamble de interés con bases de genes ortólogos. Finalmente se asigna un score dependiendo si el ensamble cuenta con los genes ortólogos, los cuales pueden caracterizarse como completo, fragmentado, duplicado o faltante.

Para la identificación de los genes, se utilizó Glimmer (v3.02) (54) dentro del programa Blast2Go (v5.2) (55) con los parámetros para Genes Procariotes. Una vez identificados los genes, se buscaron homólogos en la base de datos no-redundante (nr) con BLASTX, estableciendo un límite de E-value de 1.0 E^{-3}. Particularmente, los genes identificados como miembros de la familia PPE-PGRS fueron posteriormente analizados con blast en la base de datos de Mycobrowser (56) para tener una descripción más específica de esos genes. Además, todos los genes se asociaron a familias protéicas con InterProScan y se mapearon contra términos de Gene Ontology

(GO) utilizando los siguientes parámetros: E-value-hi-filer: 1.0E^{-3}; annotation cut-off: 55 and GO weight: 5.

Finalmente, las regiones de ARN no codificante se analizaron con Infernal (v1.1.3) (57) comparando contra la base de datos de Rfam 14.1. Las coordenadas de ARN no codificante identificados por el programa se agregaron al resto de las anotaciones en el archivo .gff.

V.II Transcriptoma de *M. tuberculosis*

1. Cultivos de bacteria.

La cepa de referencia de Mtb se cultivó a 37 ºC en medio líquido Middlebrook 7H9 (Millipore, MA, EUA, Cat. M0178) enriquecido con Suplemento ADC (albúmina, dextrosa y catalasa) (Millipore, MA, EUA, Cat. M0553). La densidad óptica (DO) del cultivo se monitoreó semanalmente, y la pureza se comprobó con una tinción Zihel-Neelsen. Los bacilos se recuperaron cuando el cultivo alcanzó la fase de crecimiento logarítmica media (DO 0.6). Para esto, el cultivo pasó por tres ciclos de lavado y centrifugación de 5 min a 3000 rpm con solución salina amortiguada por fosfatos (PBS) y tween 80 al 0.05%. Las bacterias recuperadas se resuspendieron en solución salina estéril y se almacenaron a -80ºC en alícuotas de 1 mL.

Para determinar la concentración de bacterias, se realizaron diluciones seriadas de tres alícotas independientes y se sembraron 10 µL de cada dilución en cajas de Petri con medio sólido Middlebrook 7H10 (Millipore, MA, EUA, Cat. M0303). Todas las cajas se incubaron a 37 ºC y 5% de $CO_{2.}$ En el día 14 de incubación, se contaron las colonias crecidas en cada dilución y se calculó el número de Unidades Formadoras de Colonias (UFCs) por mililitro de alícuota. Posteriormente, a partir de la concentración de bacterias en cada alícuota, se prepararon dosis individuales de 2.5x10^{5} UFCs en 100 µL de solución salina para cada ratón.

2. Infección de ratones.

El modelo experimental de tuberculosis pulmonar utilizado en este trabajo ha sido descrito previamente (27, 28). Brevemente, ratones BALB/c machos de 6-8 semanas de edad (22g de peso aproximandamente) y libres de patógenos fueron anestesiados con 100 µL de vapor de sevofluorano (Sevorane®) (Abbvie, IL, EUA) dentro de una cámara de acrílico e inoculados intratraquealmente con $2.5x10^5$ UFCs de Mtb H37Rv utilizando una cánula de calibre 22Gx1.0" y punta roma de 1.25 mm. Una vez inoculados los ratones se mantuvieron en posición vertical hasta que pasara la anestesia y finalmente se colocaron aleatoriamente en grupos de 10 en cajas con microaisladores.

Veintiún días después de la infección, los ratones fueron sacrificados por exanguinación. De cada ratón se recuperaron ambos pulmones con material quirúrgico estéril, cada pulmón se colocó en tubos de microcentrífuga con 1 mL de RNAlater® (Sigma-Aldrich, MO, EUA, Cat. R0901) y se congeló inmediatamente en nitrógeno líquido.

Todos los procedimientos de infección y sacrificio se realizaron en una campana de bioseguridad clase III. Los ratones fueron proveídos por el Departamento de Investigación Experimental y Bioterio del Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán y su uso fue aprobado por el comité de ética del mismo instituto (Comisión de Investigación en Animales CINVA 1329, aprovado el 23 de febrero de 2015).

3. Extracción de ARN y uso de sondas de captura de ribosomales para obtener el transcriptoma de *M. tuberculosis.*

Se utilizaron ambos pulmones del mismo ratón para obtener el ARN tanto de Mtb como del ratón. Para Mtb, se establecieron tres estrategias para enriquecer el ARN de bacteria a partir de los pulmones infectados. Dos estrategias se aplicaron durante la extracción de ARN y la tercera, después de construir las librerías de secuenciación (Figura 4).

3.1 Estrategias 1 y 2.

Extracción de ARN. Tomamos los pulmones del lado izquierdo de siete ratones y los pulverizamos utilizando morteros y pistilos estériles congelados con nitrógeno líquido. El polvo congelado de cada pulmón se colocó en tubos para microcentrífuga fríos y se asignaron al azar a una de las dos estrategias para la extracción de ARN (1 pulmón para la Estrategia 1 y 6 pulmones para la Estrategia 2).

i) Estrategia 1 Extracción directa de ARN total: El polvo congelado de un pulmón se trató directamente con el kit Quick RNA mini prep (Zymo Research CA, USA; Cat. R1055) siguiendo las recomendaciones del fabricante incluyendo la digestión con DNAsa I para remover el ADN contaminante. Finalmente se cuantificó el ARN resultante con un fluorómetro Qubit y se revisó la integridad con un bioanalizador Agilent RNA 6000 Pico (Agilent Technologies, CA, EUA, Cat. 5067-4626).

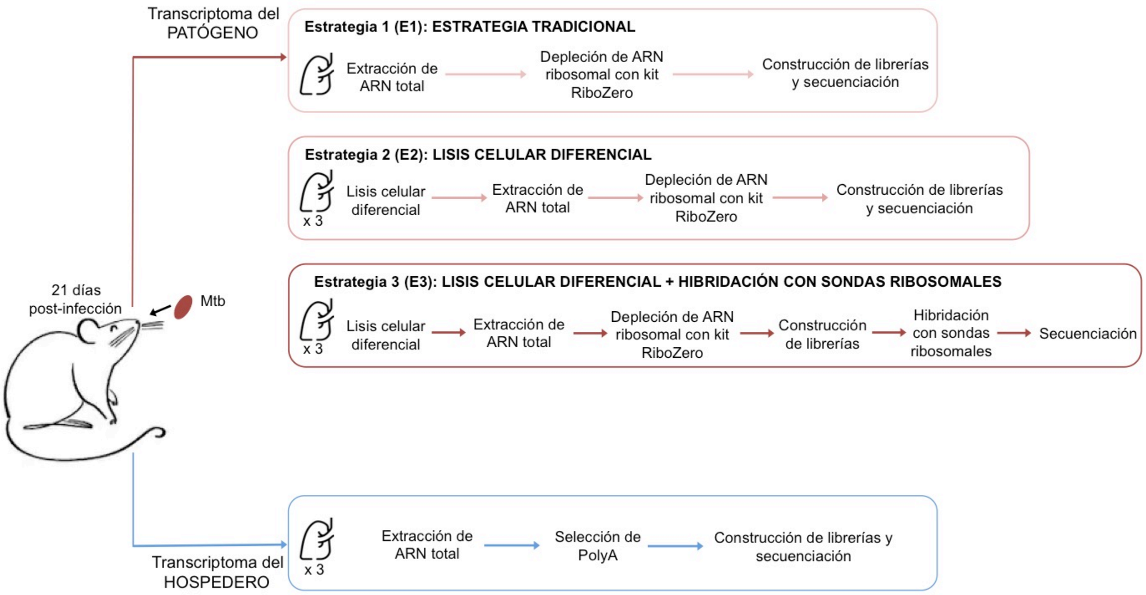

Transcriptoma del PATÓGENO
Estrategia 1 (E1): ESTRATEGIA TRADICIONAL
Extracción de ARN total
Depleción de ARN ribosomal con kit RiboZero
Construcción de librerías y secuenciación
Estrategia 2 (E2): LISIS CELULAR DIFERENCIAL
x 3
Lisis celular diferencial
Extracción de ARN total
Depleción de ARN ribosomal con kit RiboZero
Construcción de librerías y secuenciación
Estrategia 3 (E3): LISIS CELULAR DIFERENCIAL + HIBRIDACIÓN CON SONDAS RIBOSOMALES
x 3
Lisis celular diferencial
Extracción de ARN total
Depleción de ARN ribosomal con kit RiboZero
Construcción de librerías
Hibridación con sondas ribosomales
Secuenciación
21 días post-infección
Mtb
x 3
Extracción de ARN total
Selección de PolyA
Construcción de librerías y secuenciación
Transcriptoma del HOSPEDERO

Figura 4. Diseño experimental general para obtener el transcriptoma de *M. tuberculosis* y su hospedero a partir de pulmones de ratón infectados.

ii) Estrategia 2. Extracción de ARN con lisis diferencial y centrifugación: Con esta estrategia se trataron seis pulmones de manera independiente. Al polvo congelado de cada pulmón se le agregaron 600 µL de buffer RLT (Qiagen, Hilden Alemania, Cat.79216) con β mercaptoetanol frío y 1 µL de inhibidor de RNasas murino (New England BioLabs, MA, EUA, Cat. M0314L). Se homogenizó con la pipeta y se centrifugó a 14,000 rpm/4 ºC por 5 min. Después, se descartó el sobrenadante y conservamos el pellet que se formó en el fondo del tubo. Este proceso permitió concentrar a las bacterias en el pellet, ya que un buffer de lisis suave como el RLT rompe la mayoría de las células de ratón, pero deja intactas las células de bacteria gracias a su pared celular más gruesa.

Inmediatamente después, extrajimos el ARN del pellet con el kit Quick RNA mini prep (Zymo Research CA, USA; Cat. R1055) siguiendo las recomendaciones del fabricante incluyendo la digestión con DNAsa I para remover el DNA contaminante. Finalmente se cuantificó el ARN resultante con un fluorómetro Qubit y se revisó la integridad con un bioanalizador Agilent RNA 6000 Pico. (Agilent Technologies, CA, EUA, Cat. 5067-1513)

Construcción de librerías de ARN para secuenciación. Se preparó una librería a partir del ARN obtenido de cada pulmón. Una librería con el ARN de la Estrategia 1 y seis librerías con el ARN extraído con la Estrategia 2.

Primero, todo el ARN extraído de cada pulmón fue tratado con el kit Ribo-Zero rRNA Epidemiology Removal (Illumina, CA, USA; Cat. MRZE706) siguiendo las recomendaciones del fabricante. Después, se tomaron 700 ng del ARN depletado y se procesaron con el kit NEBNext Ultra RNA Library Prep Kit para Illumina (New England BioLabs, MA, USA; Cat. E7530S) ajustando la limpieza para seleccionar insertos entre 400-600 pb y un PCR de enriquecimiento de 15 ciclos. La cantidad y calidad de todas las librerías se determinó con un fluorómetro Qubit y un bioanalizador High-Sensitivity DNA respectivamente.

3.2 Estrategias 3.

Construcción de sondas para captura de ribosomales**.** Se amplificaron las regiones de ARN ribosomal del genoma de Mtb y se purificaron los amplicones con perlas AMPure XP (Beckman Coulter, CA, EUA, Cat A63881) . Después, los amplicones se fragmentaron a un tamaño promedio de 100-300 bp en un instrumento Covaris. Los fragmentos resultantes se procesaron con el kit para librerías NEBNext Fast DNA Library Prep Set for Ion Torrent (New England BioLabs, MA, EUA; Cat. E6270S) siguiendo las recomendaciones del fabricante. Finalmente se amplificaron las sondas con oligos biotinilados y se purificaron con perlas AMPure XP (Beckman Coulter, CA, EUA, Cat A63881).

Hibridación sustractiva con sondas ribosomales. Esta estrategia está basada en la metodología previamente publicada por Carpenter et al. (58). Tres de la librerías construídas con el ARN extraído con la Estrategía 2 se hibridaron de manera independiente durante 72 horas con una rampa de temperaturas de 95-65 ºC con las sondas de Mtb para reducir la presencia de secuencias ribosomales. Después de la hibridación, se utilizaron perlas magnéticas cubiertas con estreptavidina (Dyabeads MyOne Streptavidin C1; Invitrogen, CA, EUA; Cat. 65001) y soluciones de lavado a temperaturas entre 65-99 ºC para separar el ARN ribosomal capturado. Finalmente, la fracción de la librería que no fue capturada se cuantificó con fluorómetro Qubit y se revisó la calidad con un bioanalizador para ADN. Es importante mencionar que la hibridación se realiza sobre las librerías por la inestabilidad de las moléculas de ARN, una hibridación DNA-DNA permite una captura más eficiente. Todas las librerías finales se secuenciaron en una celda NextSeq 500 Mid Output v2 (Illumina Cat. FC-404-2001) de 150 ciclos en formato pareado en el Instituto Nacional de Medicina Genómics (INMEGEN) en la Ciudad de México.

4. Análisis de datos del transcriptoma de *M. tuberculosis*.

Las lecturas crudas se filtraron con el programa FASTX-toolkit (v0.0.13) (http://hannonlab.cshl.edu/fastx_toolkit/) estableciendo una calidad mínima >Q20 y removiendo los adaptadores. Después, se construyó un archivo multifasta para disminuir alineamientos espurios y separar las lecturas de ratón y de bacteria como lo recomienda el estudio de Avraham *et al* (42). Este archivo multifasta contenía el genoma de referencia de ratón GRCm38.p6 (GenBank GCA_000001635.8), las secuencias ribosomales de Mtb y el ensamble de la cepa infecciosa de Mtb.

Las lecturas limpias se mapearon con SMALT (v0.7.6) (https://www.sanger.ac.uk/tool/smalt-0/), ajustando parámetros estrictos para sólo considerar como positivos las lecturas que mapearon con una cobertura mínima ≥80%. Las lecturas mapeadas a las secuencias ribosomales, al genoma de ratón y al ensamble de Mtb se contaron de manera independiente con al suite de Samtools (v1.3.1) (59). Brevemente el archivo con el mapeo se separó en archivos independientes de acuerdo al nombre de los genomas y posteriormente se contaron las lecturas mapeadas en cada archivo. Finamente, para determinar el número de lecturas que mapearon a cada uno de los genes y ARN no codificante de Mtb, se utilizó el programa Bedtools (v2.26.0) (60) para intersectar el archivo .bam del mapeo con el archivo .gff que contenía todas las anotaciones y sus coordenadas.

El número de lecturas mpaeadas a cada gen se normalizó calculando las lecturas por kilobase por millón de lecturas mapeadas (RPKM por sus siglas en inglés) siguiendo la fórmula RPKM = Número de lecturas mapeadas / (longitud del gen/1000 * Lecturas totales/1,000,000) La normalización es necesaria para disminuir el efecto de la profundidad de secuenciación dispareja entre las muestras.

Finalmente, se calcularon las correlaciones de Pearson entre cada réplica y entre las diferentes estrategias con el programa GraphPad Prism 6 y se utilizó Blast2Go para el análisis de enriquecimiento de términos InterPro y GO con una prueba exacta de Fisher tomando un valor de p significativo <0.05 y usando el ensamble de Mtb como referencia. Brevemente, este análisis toma las anotaciones funcionales (GO e InterPro) de cada uno de los genes del genoma y lo establece como "set de referencia", después compara contra las anotaciones en la lista de genes de interés (test set). La comparación se hace para cada término por separado y finalmente hace una corrección por análisis múltiple (61). Adicionalmente se utilizó la base de datos de la Enciclopedia de Genes y Genomas de Kyoto (KEGG) en el servidor KAAS (62) para determinar las vías metabólicas más abundantes, utilizando el método del bi-directional best-hit (BBH) y *Mycobacterium tuberculosis* como el set de genes de referencia.

V.III Transcriptoma de ratón.

1. Extracción de ARN.

Se utilizaron los mismos ratones usados para la Estrategia 3 del transcriptoma de Mtb. Primero, se pulverizó el pulmón derecho completo de los tres ratones independientes utilizando mortero y pistilo estériles, congelados con nitrógeno líquido. El homogenado resultante de cada pulmón se procesó con el kit Quick RNA miniprep (Zymo Research CA, USA; Cat. R1055) siguiendo las recomendaciones del fabricante, incluyendo la digestión con DNAsa I para remover el ADN contaminante. Después de la extracción, se cuantificó y se revisió la calidad del ARN obtenido con fluorómetro Qubit y con bioanalizador Agilent RNA 6000 Pico.

2. Construcción de librerías de ARN para secuenciación.

Se preparó una librería de secuenciación por cada pulmón. Primero, se tomaron 700 ng del ARN total de cada pulmón y se seleccionaron las secuencias poliadeniladas con el kit NEBNext Poly(A) mRNA Magnetic Isolation (New England BioLabs, MA, USA; Cat. E7490) siguiendo las recomendaciones del fabricante. Después, el ARN poliadenilado se utilizó para construir la librería con el kit NEBNext Ultra RNA Library Prep Kit for Illumina (New England BioLabs, MA, USA; Cat. E7530S), ajustando la selección de tamaños para insertos entre 400-600 pb y un PCR de enriquecimiento de 15 ciclos. Al final, se determinó la cantidad y la calidad de las librerías con un fluorómetro Qubit y un bioanalizador para ADN. Todas las librerías se secuenciaron en una celda NextSeq 500 Mid Output v2 (Illumina Cat. FC-404-2001) de 150 ciclos en formato pareado en el Instituto Nacional de Medicina Genómics (INMEGEN) en la Ciudad de México.

3. Análisis de datos del transcriptoma de ratón.

Se utilizó el programa FASTX-toolkit (v0.0.13) para filtrar las lecturas con una calidad mínima >Q20 y se removieron los adaptadores con Trim_Galore (v0.4.2) (https://github.com/FelixKrueger/TrimGalore).

Las lecturas limpias se alinearon al genoma de referencia de ratón *Mus musculus* GRCm38.p6 (GenBank GCA_000001635.8) usando Bowtie2 (v2.3.5) (63) con los parámetros default y se obtuvo la cuenta de lecturas mapeadas a cada gen con HTSeq (v0.6.1) (64). Después, se normalizó la expresión de cada gen calculando el RPKM.

Adicionalmente, se utilizó Blast2Go para un análisis de enriquecimiento de términos GO con una prueba exacta de Fisher tomando un valor de *p* significativo <0.05 y usando el genoma de *Mus musculus* como referencia. También se utilizó la base de datos KEGG para determinar las vías metabólicas más abundantes, utilizando el método BBH y *Mus musculus* como el set de genes de referencia.

V.IV Secretoma de M. tuberculosis y M. abscessus.

1. Aislados clínicos.

Para Mtb se seleccionaron dos aislados clínicos del genotipo Beijing del laboratorio del Dr. Rogelio Hernández Pando. Y para Mabs seleccionamos 15 aislados clínicos que incluyeron miembros de las tres subespecies (*M. abscessus* supsp. *abscessus*, (MAB$_A$) n =7; *M. abscessus* supsp. *masiliense* (MAB$_M$) n=4; y *M. abscessus* supsp. *bolleti* (MAB$_B$) n=4) y los dos morfotipos de colonias (rugoso (R) n=6, y liso (S) n=8) (no determinado, n=1). Las cepas fueron aisladas de diferentes fuentes biológicas representando tanto infecciones pulmonares como extra pulmonares. Estos aislados fueron obtenidos por el Dr. Florian P. Maurer del Centro Nacional de Referencia para Micobacterias en Borstel, Alemania.

2. Secuenciación y ensamble de los genomas.

La extracción de ADN genómico y la secuenciación de los aislados de Mtb se realizó previamente en el Instituto Nacional de Medicina Genómica (INMEGEN) en la Ciudad de México por M.C. Vito A. Cantú–Robles, otro miembro del laboratorio. Brevemente, el ADN genómico de los aislados de Mtb se realizó a partir de cultivos líquidos en fase logarítmica utilizando el kit Quick gDNA MiniPrep (Zymo Research CA, USA; Cat. D3006), siguiendo las recomendaciones del fabricante. Posteriormente las librerías se construyeron con el kit NebNext DNA Library Prep (New England BioLabs, MA, USA; Cat. E6040S) y se secuenciaron en la plataforma GAIIx de Illumina a una longitud de 72 pb.

Las secuencias crudas se procesaron con el programa FastX-Toolkit (v0.0.13) removiendo los adaptadores y las lecturas con una calidad menor a Q20. Finalmente, las secuencias limpias se utilizaron para los armados de novo con el programa Velvet (v1.2.10) (65). Los ensambles finales se analizaron con RAST (66) para obtener todos los marcos de lectura abiertos (ORFs). Del mismo modo se obtuvieron los ORFs del genoma de referencia de Mtb depositado en GenBank NC_000962.3.

Por otro lado, el ADN de los 15 aislados clínicos de Mabs se extrajo y se secuenció en el Centro Nacional de Referencia para Micobacterias en Borstel, Alemania por el grupo del Dr. Florian P. Maurer. Las secuencias resultantes se procesaron con Trimmomatic (v0.40) (67), eliminando adaptadores y limpiando las lecturas con una ventana de 20 pb y una calidad mínima >30. Las lecturas limpias se usaron para construir los armados de novo con SPADES (v3.15.2) utilizando los parámetros default.

Finalmente, cada ensamble y el genoma de referencia de *M. abscessus* ATCC19977 (GenBank CU4588961) también se analizaron con RAST para obtener los ORFs.

3. Predicción del secretoma.

Esta estrategia bioinformática para predecir el secretoma de Mtb y Mabs se publicó previamente en Cornejo-Granados et al., 2017 (68) y Cornejo-Granados et al. 2021, (69). Brevemente, cada uno de los proteomas de Mtb y de Mabs obtenidos con RAST se analizaron con seis programas en total. Primero, de manera independiente usamos cuatro programas dedicados a predecir proteínas secretadas por diferentes vías. SignalP 4.1 (70) predice la presencia de péptidos señales que favorezcan la secreción de la proteína por la vía clásica (Sec-dependiente); SecretomeP 2.0 (71) predice la secreción de proteínas por la vía no clásica; TatP1.0 (72) determina la secreción de proteínas vía Tat (twin-arginine translocation), y LipoP1.0 (73) predice dominios de lipoproteínas.

Posteriormente juntamos todas las proteínas positivas a cada uno de los programas y las analizamos con dos programas más. TMHMM2.0 (74), el cual identifica la localización y orientación de hélices transmembranales. En este caso, todas las proteínas predichas con cero motivos transmembranales, fueron asignadas directamente como parte del secretoma mientras que el resto (≥1 motivos transmembranales) se analizaron con Phobius (75), el cual identifica péptidos señal en la región N-terminal que podrían haberse clasificado erróneamente como motivos transmembranales. En este caso, todas las proteínas predichas con un péptido señal por Phobius se agregaron a la lista del secretoma.

Finalmente, las proteínas caracterizadas como no secretadas se clasificaron como transmembranales (TM) si contaban con dominios transmembranales según TMHMM2.0, o en caso contrario, se clasificaban como intracelulares (incell).

4. Anotación funcional y análisis comparativo del secretoma.

Utilizamos Blast2GO (v5.2) para anotar cada secretoma de acuerdo a su homología contra la base de datos no redundante utilizando BLASTp con un E-value límite de 0.001. Así mismo, hicimos análisis de enriquecimiento de términos de ontología genética (GO) y familias protéicas de InterPro considerando sólo los términos enriquecidos con un valor de p ≤ 0.05.

Adicionalmente utilizamos el servidor KAAS para comparar los secretomas contra la Enciclopedia de Genes y Genomas de Kyoto (KEGG) utilizando el método bi-directional best hit (BBH) y conocer las vías metabólicas más representadas. Finalmente las comparaciones entre los secretomas de las diferentes cepas para establecer las proteínas compartidas y únicas se realizaron con alineamientos bi-direccionales con BLASTp (E-value 0.001).

5. Cálculo de la Abundancia de Regiones Antigénicas.

La Abundancia de Regiones Antigénicas (AAR) es una medida para predecir la densidad antigénica de una proteína y se obtiene dividiendo la longitud de la secuencia primaria de la proteína entre el número de regiones antigénicas. Por lo tanto, proteínas con valores de AAR bajos tendrán una densidad antigénica mayor que proteínas con valores de AAR altos (76).

Para calcular el número de regiones antigénicas utilizamos el programa Bepipred1.0 (77), estableciendo el score límite sugerido por los autores de 0.35 que otorga una especificidad del 75%. Después, utilizando scripts desarrollados en casa contamos el número de regiones antigénicas con una longitud mínima de 7 aminoácidos. Y finalmente, dividimos la longitud de la proteína entre el número de regiones antigénicas. Posteriormente, utilizamos el test estadístico de Mann-Whitney para establecer si la diferencia entre los valores de AAR de los secretomas era significativa.

6. Construcción del secretoma experimental de *M. tuberculosis.*

Para validar nuestra estrategia bioinformática comparamos el secretoma predicho de Mtb contra un secretoma validado experimentalmente. Para construirlo, realizamos una búsqueda sistemática en NCBI de los artículos que reportaran proteínas secretadas analizadas con técnicas experimentales. Para la búsqueda se usaron 10 términos en inglés: secreted tuberculosis proteins, membrane tuberculosis proteins, proteome analysis tuberculosis, secretome analysis tuberculosis, secretome database tuberculosis, culture filtrate proteins tuberculosis, exported proteins tuberculosis, secreted antigens tuberculosis, surface proteins tuberculosis, culture supernatant tuberculosis. Se excluyeron capítulos de libro, reportes de conferencias y estudios en organismos diferentes a Mtb. La búsqueda resultó en 338 proteínas previamente reportadas como secretadas. Después realizamos un BLASTp para determinar cuántas de las 338 proteínas tenían un homólogo en el secretoma predicho.

VI. Resultados

VI.I RNA-seq para determinar el transcriptoma de *M. tuberculosis* en un modelo *in vivo* de tuberculosis pulmonar.

1. Armado *de novo* del genoma de *M. tuberculosis*.

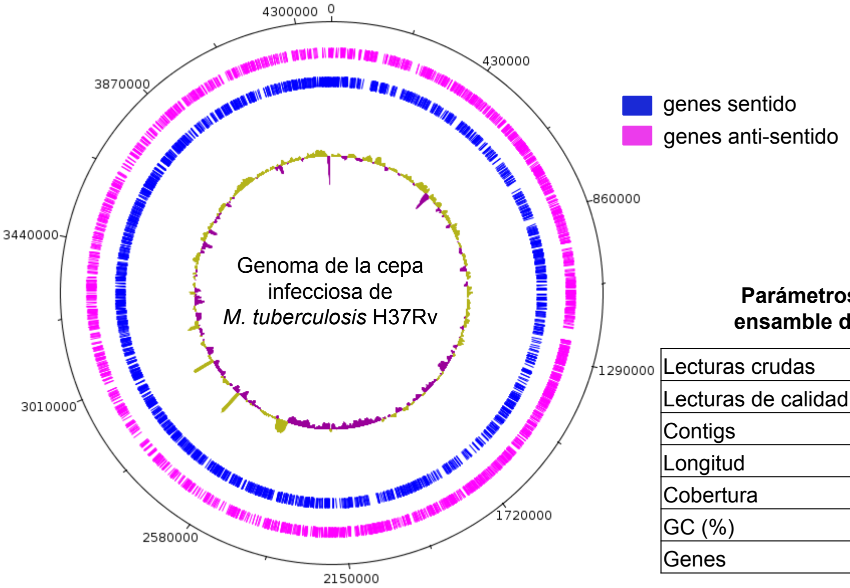
Para realizar el análisis de los datos de RNA-seq, primero ensamblamos y anotamos el genoma completo de la cepa de Mtb que se utilizó para infectar a los ratones (Figura 5). De la secuenciación, se obtuvieron 14,421,346 lecturas totales con 150 pb de longitud promedio de las cuales, el 98.1% tuvo una calidad ≥Q20. Las lecturas limpias, se ensamblaron con SPADES y MEDUSA obteniendo un sólo cromosoma de 4,392,486 pb de longitud, con una cobertura promedio de 788x y un contenido de GC del 65.57%. Adicionalmente, el análisis con gVolante mostró un "completness" del 97.5%, con 39 genes ortólogos core completos y 1 core faltante. Posteriormente, con Glimmer se identificaron 4,234 genes, 99.24% de los cuales se les asignó un homólogo con Blast comparando con la base de datos no-redundante del NCBI (Figura 6).

Es importante mencionar que para un primer análisis se utilizó el genoma de referencia de H37Rv (GCF000195955.2). Sin embargo, en este caso observábamos la expresión de 65 genes menos que al usar el armado *de novo*.

Al analizar el promedio de identidad de nucleótidos (ANI) entre el genoma de referencia y el armado obtuvimos un porcentaje de identidad del 99.98% y una cobertura de alineamiento del 100%, lo que nos indica que ambos genomas pertenecen a la misma especie. Probablemente la diferencia en el número de genes observados se debe a que hay variaciones a nivel de nucleótido y los parámetros estrictos de mapeo que utilizamos no permitieron alinear más lecturas al genoma de referencia. Por todo lo anterior, decidimos usar el armado para el análisis final.

2. Estrategias experimentales para obtener el transcriptoma de *M. tuberculosis* en un modelo *in vivo* al día 21 post infección.

Se probaron tres estrategias experimentales (Figura 4) que buscaban enriquecer el ARN de Mtb obtenido de pulmones infectados para poder observar el mayor número de genes expresados de la bacteria durante la infección en el pulmón.



Parámetros estadísticos del ensamble de *M. tuberculosis*.

Lecturas crudas	14,421,346
Lecturas de calidad (>Q20)	14,147,340
Contigs	1
Longitud	4,392,486 bp
Cobertura	788X
GC (%)	65.57
Genes	4,234

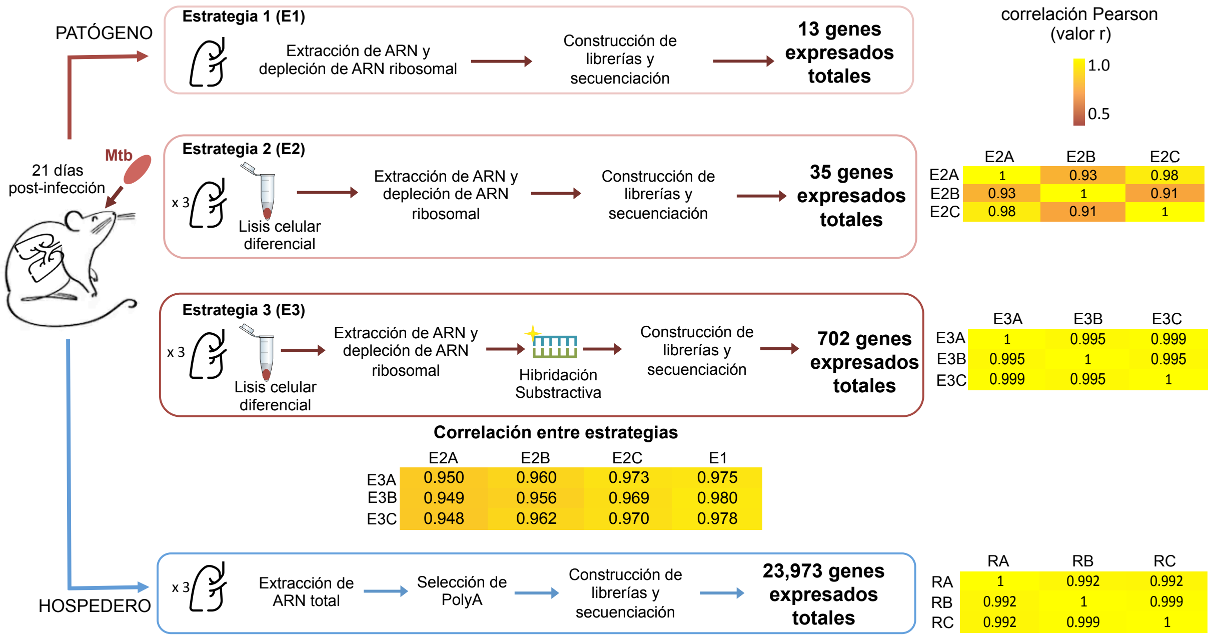
Figura 5. Genoma ensamblado de Mtb. Las regiones con una cobertura mayor o menor al promedio (788x) se muestran en el círculo interno de color amarillo y morado respectivamente, mientras que los círculos azul y rosa representan los genes “sentido” y “anti-sentido” respectivamente. Los principales parámetros estadísticos del armado se muestran en la tabla de la derecha.

Estrategia 1: Estrategía estándar para RNA-seq.

Como una primera aproximación, para conocer la cantidad de secuencias de Mtb que se obtienen con la estrategia estándar de RNA-seq, secuenciamos el ARN total extraído de un pulmón infectado (Estrategia 1 (E1), Figura 6). Para esta primera estrategia, sólo se construyó una librería porque el objetivo era tener una idea de qué tan baja era la proporción de secuencias de Mtb extrayendo el ARN total de un pulmón infectado. Encontramos que sólo el 1.70% de las secuencias correspondían a Mtb (Tabla Anexa I), representando la expresión de sólo 13 genes, mientras que el resto eran secuencias de ratón.

Estrategia 2: Enriquecimiento de ARN bacteriano con lisis celular diferencial.

Al observar la proporción tan baja de secuencias de Mtb, estandarizamos un protocolo con lisis celular diferencial para concentrar las micobacterias presentes en los pulmones infectados (Estrategia 2 (E2), Figura 6). Para ello, seleccionamos una primer solución para lisar las células pulmonares de ratón manteniendo íntegras las células de la bacteria, aprovechando que estas últimas tienen una pared celular más gruesa y resistente que las células de ratón.


PATÓGENO
Estrategia 1 (E1)
Extracción de ARN y depleción de ARN ribosomal
Construcción de librerías y secuenciación
13 genes expresados totales
correlación Pearson (valor r)
1.0
0.5
Mtb
21 días post-infección
Estrategia 2 (E2)
x 3
Lisis celular diferencial
Extracción de ARN y depleción de ARN ribosomal
Construcción de librerías y secuenciación
35 genes expresados totales
Estrategia 3 (E3)
x 3
Lisis celular diferencial
Extracción de ARN y depleción de ARN ribosomal
Hibridación Substractiva
Construcción de librerías y secuenciación
702 genes expresados totales
Correlación entre estrategias
HOSPEDERO
x 3
Extracción de ARN total
Selección de PolyA
Construcción de librerías y secuenciación
23,973 genes expresados totales

	E2A	E2B	E2C
E2A	1	0.93	0.98
E2B	0.93	1	0.91
E2C	0.98	0.91	1

	E3A	E3B	E3C
E3A	1	0.995	0.999
E3B	0.995	1	0.995
E3C	0.999	0.995	1

	E2A	E2B	E2C	E1
E3A	0.950	0.960	0.973	0.975
E3B	0.949	0.956	0.969	0.980
E3C	0.948	0.962	0.970	0.978

	RA	RB	RC
RA	1	0.992	0.992
RB	0.992	1	0.999
RC	0.992	0.999	1

Figura 6. Estrategias experimentales para obtener el perfil de expresión de Mtb (E) y de ratón (R) a partir de pulmones infectados. El valor de R2 de las correlaciones de Pearson con la abundancia de expresión de genes entre réplicas biológicas (A, B and C) y estrategias (E) se muestran en mapas de calor amarillos.

Después de centrifugar el homogenado de pulmón retiramos el sobrenadante y con una tinción Zihel-Neelsen corroboramos la concentración de bacterias en el pelletdespués de la lisis celular diferencial (Figura 7). Finalmente, extrajimos el ARN total del pellet concentrado de bacterias, obteniendo valores promedio de integridad de ARN (RIN) de 7.

Los datos de secuenciación de tres réplicas biológicas tratadas con la lisis celular diferencial confirmaron un aumento del 3.4 veces en promedio en la proporción promedio de secuencias de Mtb, de 1.70% en E1 a 5.83% en E2 (Tabla Anexa I). Así mismo, el protocolo mostró una buena reproducibilidad con valores de correlación de Pearson >0.9 (Figura 7).

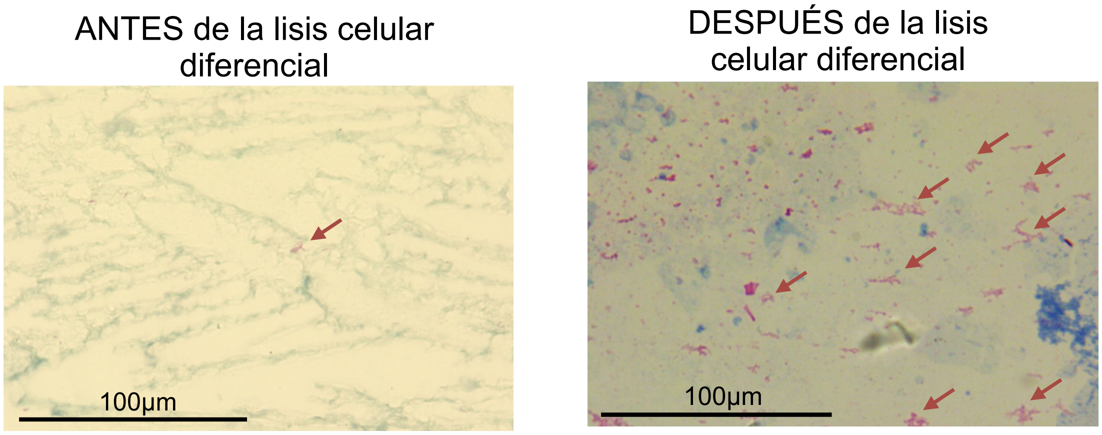

ANTES de la lisis celular diferencial
100µm
DESPUÉS de la lisis celular diferencial
100µm

Figura 7. Tinción Zihel-Neelsen del homogenado de pulmón antes y después de la lisis celular diferencial. Las flechas rojas muestran los cúmulos de bacilos observados en cada caso.

Sin embargo, a pesar de haber hecho una primera eliminación de ARN ribosomal con un kit comercial, el ~97% de las secuencias de Mtb aún mapearon a transcritos ribosomales, y sólo el 3% de lecturas restante representaba la expresión de 35 genes.

Estrategia 3: Hibridación con sondas ribosomales de Mtb.

Con el fin de reducir la cantidad de ARN ribosomal, diseñamos sondas biotiniladas para hibridar y sustraer los transcritos ribosomales de Mtb presentes en tres pulmones infectados (Estrategia (E3), Figura 6 y 8).

Primero, construimos tres librerías de secuenciación a partir del ARN extraído de tres pellets de pulmón independientes utilizando el protocolo de lisis celular diferencial. Después hibridamos cada una de las librerías con las sondas biotiniladas y secuenciamos la porción de la librería que no fue capturada (Figura 8).

En este caso, los datos de secuenciación mostraron una reducción de 1.3 veces en los transcritos ribosomales de Mtb, de 97.07% (E2) a 72.51% (E3) (Tabla Anexa I), mientras que las secuencias remanentes permitieron detectar la expresión de 702 genes de Mtb. Además, las tres muestras mostraron una buena reproducibilidad entre sí con valores de correlación de Pearson >0.9 (Figura 6).

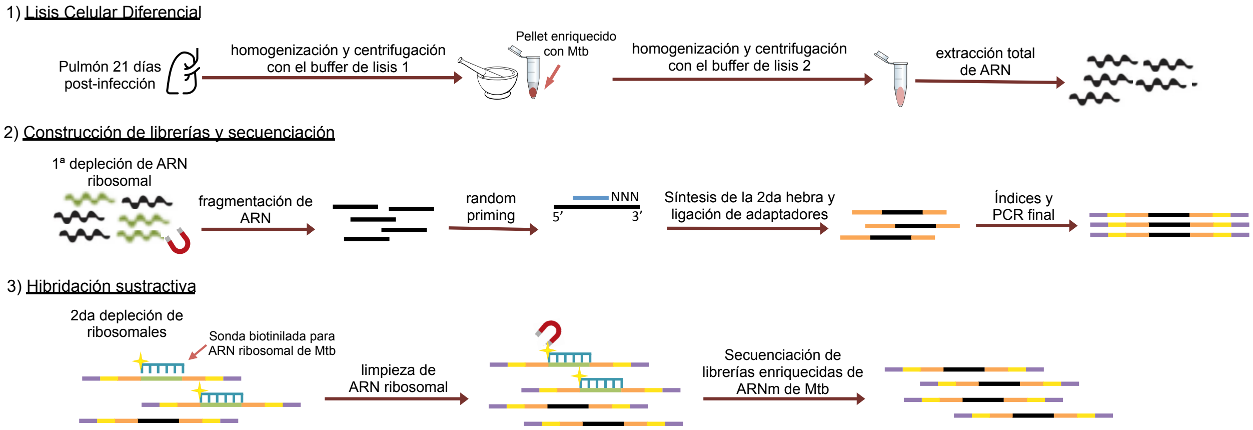

1) Lisis Celular Diferencial
Pulmón 21 días post-infección
homogenización y centrifugación con el buffer de lisis 1
Pellet enriquecido con Mtb
homogenización y centrifugación con el buffer de lisis 2
extracción total de ARN
2) Construcción de librerías y secuenciación
1ª depleción de ARN ribosomal
fragmentación de ARN
random priming
NNN
5′
3′
Síntesis de la 2da hebra y ligación de adaptadores
Índices y PCR final
3) Hibridación sustractiva
2da depleción de ribosomales
Sonda biotinilada para ARN ribosomal de Mtb
limpieza de ARN ribosomal
Secuenciación de librerías enriquecidas de ARNm de Mtb

Figura 8. Procedimiento experimental detallado de la Estrategia 3 (E3) para obtener el transcriptoma de Mtb.

Estos resultados demostraron que la Estrategia 3 fue eficiente para depletar ARN ribosomal de Mtb y produjo un enriquecimiento de ~20 veces en el número de genes observados en comparación con E2 (sólo lisis celular diferencial).

Es importante notar que la profundidad de secuenciación en las diferentes estrategias no es comparable, E2 tiene en promedio 14,343,220 lecturas limpias, mientras que E3 tiene 50,211,880 lecturas. Para evaluar si vemos más genes en E3 por un efecto de la profundidad de secuenciación, tomamos 14 millones de lecturas al azar de cada una de las muestras de E3 y mapeamos a los diferentes genomas. Como resultado observamos 358 genes de Mtb, 10 veces más que los genes de E2, lo cual sugiere que el uso de sondas en E3 ayuda a observar más genes expresados.

Finalmente para descartar un sesgo en la expresión de genes observada debido al uso de la lisis celular diferencial (E2) y la hibridación con sondas (E3), los genes observados en E1 contra los observados en E2 y E3, y observamos que todos los genes de E1 se observan también en E2 y E3. Sin embargo, en E1 hay un promedio de 16 lecturas mapeadas a cada gen, mientras que en E2 y E3 hay 24 y 49 lecturas respectivamente. Adicionalmente comparamos la abundancia de expresión de los genes de Mtb normalizada por RPKM para ambas estrategias (E2 y E3) contra la abundancia obtenida para E1 (estrategia tradicional). Y observamos una correlación de Pearson con r>0.9 entre las réplicas y entre las estrategias (Figura 7). Estas correlaciones altas mostraron que la Estrategia 3 (lisis celular diferencial e hibridación con sondas) generó resultados reproducibles sin sesgo en la expresión de genes.

Ahora, utilizando los datos de la Estrategia 3 realizamos un análisis más detallado de los genes expresados de Mtb.

3. Transcriptoma de *M. tuberculosis* en un modelo *in vivo* al día 21 post infección.

De acuerdo a los datos de secuenciación de E3, observamos que el 61.59% de las secuencias totales de Mtb mapearon a 702 genes, lo que corresponde al 16.58% del total de los genes en el genoma de Mtb; mientras que el restante 38.41% de las secuencias mapeó a 30 ARNs no codificantes o regiones intergénicas. En un primer análisis sólo consideramos 529 genes que tuvieron ≥2 lecturas en al menos una muestra de pulmón y un RPKM promedio ≥1 y así evitar genes que sólo tuvieran una lectura.

Para obtener un descripción general de las funciones de los genes expresados de Mtb, determinamos los dominios y familias de InterPro y los términos de Gene Ontology (GO) significativamente sobrerrepresentados (p<0.05) en los 529 genes. (Tablas I y II, Tablas Anexas 2 y 3). En este análisis observamos en promedio un enriquecimiento significativo (p<0.05) de 5 genes para cada uno de los ocho dominios proteicos de InterPro relacionados con el metabolismo de ácidos grasos y 4 para las dos familias que representan subunidades del sistema VII de secreción (Tabla I y Tabla Anexa 2). Por otro lado, el análisis de términos GO mostró un enriquecimiento significativo (p<0.05) de 35 genes en promedio para cada término GO en la categoría de Proceso Biológico, 37 genes para Componente Celular y 58 genes para Función Molecular. Estos términos estaban relacionados con células T, co-estimulación de linfocitos, nitrato reductasa, complejos de receptores alfa-beta en células T, receptores endoteliales, unión a zimógenos y actividad de receptores tipo scavenger (Tabla II y Tabla Anexa 3).

Adicionalmente, el mapeo de los 529 genes expresados a vías metabólicas KEGG mostró diversas vías con ≥20% representado (mínimo 10 genes de la vía expresados), como la vía del metabolismo del glioxilato y dicarboxilato, ciclo del citrato, biosíntesis de ácidos grasos, quorum sensing y sistemas bacterianos de secreción (Tabla III y Tabla Anexa 4).

Tabla I. Top 20 dominios y familias de InterPro significativamente sobrerrepresentadas en los 529 genes expresados de Mtb.

InterPro ID	Categoría	Descripción	P-Value
IPR014043	dominio	Acil transferasa	6.267E-07
IPR009081	dominio	Dominio APC de unión a fosfopanteteína	1.438E-06
IPR014030	dominio	Beta-cetoacilsintetasa, N-terminal	2.747E-05
IPR014031	dominio	Beta-cetoacilsintetasa, C-terminal	4.885E-05
IPR013968	dominio	Policétido sintasa, dominio cetoreductasa	1.624E-04
IPR023836	familia	EccCa-like, Actinobacteria	1.087E-03
IPR023837	familia	EccCb-like, Actinobacteria	1.087E-03
IPR032821	dominio	Cetoacil-sintetasa, extensión C-terminal	1.190E-03
IPR006091	dominio	Acyl-CoA oxidasa/deshidrogenasa, dominio central	1.363E-03
IPR020807	dominio	Policétido synthase, dominio dehidratasa	1.757E-03
IPR009075	dominio	Acil-CoA deshidrogenasa/oxidasa C-terminal	1.995E-03
IPR002543	dominio	dominio FtsK	6.180E-03
IPR003029	dominio	dominio S1	7.039E-03
IPR003495	dominio	CobW/HypB/UreG, dominio de unión a nucleótido	7.039E-03
IPR023753	dominio	FAD/NAD	1.360E-02
IPR000788	dominio	subunidad larga de ribonucleótido reductasa, C-terminal	1.558E-02
IPR001030	dominio	subunidad larga de aconitasa/3-isopropilmalato deshidratasa	1.558E-02
IPR002300	dominio	Aminoacil-tRNA sintetasa	1.558E-02
IPR003714	dominio	proteína PhoH-like	1.558E-02
IPR004100	dominio	ATPasa, complejo F1/V1/A1, subunidad alfa/beta, dominio N-terminal	1.558E-02

Tabla II. Top 10 de términos GO por categoría significativamente enriquecidos en los 529 genes más expresados de Mtb durante la infección in vivo.

GO ID	Nombre GO	P-Value
A. Proceso Biológico		
GO:0031295	Coestimulación de células T	1.765E-05
GO:0031294	Coestimulación de linfocitos	1.765E-05
GO:0009081	Metabolismo de aminoácidos con cadenas ramificadas	4.007E-05
GO:0009083	Catabolismo de aminoácidos con cadenas ramificadas	1.102E-04
GO:0006549	Metabolismo de isoleucina	1.197E-04
GO:0000288	Catabolismo de ARN mensajero	2.663E-04
GO:0006573	Metabolismo de valina	3.513E-04
GO:0060184	Cambio de ciclo celular	4.878E-04
GO:0051728	Cambio de mitosis a meiosis	4.878E-04
GO:0006574	Catabolismo de valina	5.108E-04
B. Componente Celular		
GO:0005643	Poro nuclear	4.20E-03
GO:0000932	P-body	5.39E-03
GO:0009986	Superficie celular	6.00E-03
GO:0098978	Sinapsis glutamaérgica	6.18E-03
GO:1990527	Complejo Tec1p-Ste12p-Dig1p	7.04E-03
GO:0110165	Estructura anatómica celular	1.23E-02
GO:1990526	Complejo Ste12p-Dig1p-Dig2p	1.60E-02
GO:0030496	Midbody	1.81E-02
GO:0009325	Complejo de nitrato reductasa	1.83E-02
GO:0016020	Membrana	2.56E-02
C. Función Molecular		
GO:0004085	Actividad deshidrogenasa butyryl-CoA	1.62E-04
GO:0005488	Unión	6.96E-04
GO:0052890	Actividad de oxidoreductasa en el grupo CH-CH con flavina como aceptor	7.83E-04
GO:0043168	Unión de aniones	1.16E-03
GO:0003955	Actividad de deshidrogenasa NAD(P)H	1.94E-03
GO:0017056	Componente de poro nuclear	2.87E-03
GO:0004962	Actividad de receptor a endotelina	2.94E-03
GO:0036094	Unión de moléculas pequeñas	3.15E-03
GO:1901363	Unión de compuesto heterocíclico	4.09E-03
GO:0097159	Unión de compuesto orgánico cíclico	4.10E-03

Tabla III. Vías KEGG más representadas en los 529 genes expresados de Mtb.

KEGG ID	Vía	%
03020	ARN polimerasa	75.00
00630	Metabolismo de Glioxilato y dicarboxilato	34.15
03018	Degradación de ARN	33.33
00562	Metabolismo de inositol	33.33
00910	Metabolismo de nitrógeno	31.82
05152	Tuberculosis	30.77
00680	Metabolismo de metano	28.57
00290	Biosíntesis de valina, leucina e isoleucina	28.57
00430	Metabolismo de taurina e hipotaurina	28.57
00020	Ciclo del citrato	28.13
00730	Metabolismo de tiamina	27.27
00983	Metabolismo de fármacos	27.27
00061	Biosíntesis de ácidos grasos	26.67
02024	Quorum sensing	25.00
00010	Glicólisis / Gluconeogénesis	24.24
00270	Metabolismo de cisteína y metionina	24.24
00260	Metabolismo de glicina, serina y treonina	24.00
03060	Exporte de proteínas	23.53
01053	Biosíntesis de sideróforo en péptidos no ribosomales	22.22
03070	Sistema de secreción bacterianos	21.43

En general, el gen más expresado de Mtb fue la transposasa para la secuencia de inserción IS1081 (Rv2512c) (Tabla IV), mientras que la transposasa para IS6110, una de las más estudiadas en Mtb se observó alrededor del lugar 250 de los 529 genes expresados.

Notablemente, los siguientes ocho genes más expresados fueron miembros de la familia PE-PGRS (Tabla IV), los cuales son importantes antígenos de superficie, que tienen una clara interacción con el sistema inmune del hospedero (78-80). La presencia abundante de genes PE-PGRS coincidió con la expresión de subunidades de los sistemas de secreción necesarios para su transporte, especialmente ESX-1, dentro de los primeros cien genes más expresados (Tabla Anexa 3), ESX-3 y ESX-5 dentro de los primeros ciento cincuenta genes más expresados (Tabla Anexa 3) (81).

Tabla IV. Top 20 de genes con respecto a los 529 genes expresados de Mtb.

Ranking de expresión	Gen ID	Descripción	RPKM promedio
1	Rv2512c	Transposasa para la secuencia de inserción IS1081	6.84E+05
2	Rv1067c	PE_PGRS19 PE-PGR	3.45E+05
3	Rv3512	PE_PGRS56 PE-PGRS	1.91E+05
4	Rv1067c	PE_PGRS19 PE-PGR	1.65E+05
5	Rv3344c	PE_PGRS49 PE-PGR	1.63E+05
6	Rv0279c	PE_PGRS4 PE-PGRS	1.10E+05
7	Rv3512	PE_PGRS56 PE-PGRS	5.01E+04
8	Rv0746	PE_PGRS9 PE-PGRS	4.35E+04
9	Rv3345c	PE_PGRS50 PE-PGR	1.94E+04
10	Rv0105c	uracil-DNA glicosilasa	7.61E+03
11	Rv2840c	proteína con dominio DUF448	6.94E+03
12	Rv0440	chaperona GroEL	5.71E+03
13	Rv3620c	peptidasa M22	4.88E+03
14	Rv0454	proteína hipotética	4.58E+03
15	Rv0967	regulador de la transcripción	4.12E+03
16	Rv2424c	transposasa	4.05E+03
17	Rv1228	lipoproteína lpqX	3.66E+03
18	Rv2424c	proteína no caracterizada	3.11E+03
19	Rv3053c	Chain A, Glutaredoxina Nrdh	3.09E+03
20	Rv0637	Subunidad HadC (3R)-hidroxiacil-ACP deshidratasa	2.96E+03

En relación con el 38.41% de las secuencias mapeadas a ARNs regiones intergénicas, estas estaban distribuidas en 30 localizaciones, tres de las cuales fueron caracterizadas como ARN no codificantes y representaban el 18.04% de las secuencias mapeadas estas regiones. Las 30 regiones estaban localizadas cerca de 37 genes que observamos en la lista de los 529 genes expresados. Adicionalmente, las regiones identificadas como ARN no codificantes, incluyeron el ARN pequeño ncRv0036, mejor conocido como MTS2823, el cual concentró 235 lecturas (42.34%); el ARN de transferencia de mensajero RF00023 con 167 lecturas (30.09%), y la ribozima RF0010 con 153 lecturas (27.57%). Sin embargo, los genes inmediatos a estos ARNnc no muestran una alta expresión.

Dado que el análisis que realizamos consideró los 529 genes con ≥2 lecturas en al menos una muestra de pulmón y un RPKM promedio ≥1, ahora hicimos un corte más estricto tomando 101 genes que tuvieron al menos 1 lectura en dos muestras y analizamos si nuestras observaciones sobre el perfil de expresión génica de Mtb cambiaban (Tabla Anexa 6). Con este corte, se eliminaron los genes que estuvieran en un sólo ratón.

El análisis de enriquecimiento de términos GO con los 101 genes más expresados mostró términos relacionados con regulación de la transcripción, transporte de aminoácidos y síntesis de componentes de la pared bacteriana como los términos más significativamente enriquecidos (p ≤0.05). Por otro lado, términos relacionados con la patogénesis como coestimulación de células T y linfocitos, señalización de receptor de antígenos, síntesis de ácidos grasos y transporte de hierro, que en el primer análisis considerando los 529 estaban en los primeros lugares de la tabla aparecieron ahora con un valor de p más alto pero aún con significancia estadística (p ≥0.05). Adicionalmente, el análisis de vías metabólicas KEGG mostró tres vías cubiertas en >10% por los 101 genes más expresados, Tuberculosis (23.08%), degradación de ARN (16.67%) y biosíntesis del grupo sideróforo en péptidos no ribosomales (11.11%).

Por otro lado, la lista de los 20 genes más expresados con este corte se mantiene muy similar a la lista presentada en el primer análisis (Tabla IV y V), sugiriendo que el análisis con el corte más laxo es robusto. En este nuevo top 20, sólo hay 8 genes de los lugares del 13-20 que cambian. Ahora observamos genes que antes estaban debajo del lugar 20, por ejemplo, al gen para PPE68 (Rv3873), el factor sigE de la ARN polimerasa (Rv1221), la proteína alfa cristalina (Rv0251c), una bactoferrina (Rv1876), dos proteinas del sistema VII de secreción (Rv3615c y Rv0287), una proteína ribosomal (Rv2442c) y una proteína de fusión (Rv0350) (Tabla V).

Tabla V. Top 20 de genes con respecto a los 101 genes expresados de Mtb.

Ranking de expresión	Gen ID	Descripción	RPKM promedio
1	Rv2512c	Transposasa para la secuencia de inserción IS1081	6.84E+05
2	Rv1067c	PE_PGRS19 PE-PGR	3.45E+05
3	Rv3512	PE_PGRS56 PE-PGRS	1.91E+05
4	Rv1067c	PE_PGRS19 PE-PGR	1.65E+05
5	Rv3344c	PE_PGRS49 PE-PGR	1.63E+05
6	Rv0279c	PE_PGRS4 PE-PGRS	1.10E+05
7	Rv3512	PE_PGRS56 PE-PGRS	5.01E+04
8	Rv0746	PE_PGRS9 PE-PGRS	4.35E+04
9	Rv3345c	PE_PGRS50 PE-PGR	1.94E+04
10	Rv0440	chaperona GroEL	5.71E+03
11	Rv3620c	peptidasa M22	4.88E+03
12	Rv2424c	proteína no caracterizada	3.11E+03
13	Rv3873	PPE68	2.02E+03
14	Rv1221	factor sigE de ARN polimerasa	1.97E+03
15	Rv0251c	Hsp20/proteína de la familia alfa cristalina	1.63E+03
16	Rv1876	bacterioferritina	1.61E+03
17	Rv3615c	formador de filamento del sistema de secreción tipo VII	1.57E+03
18	Rv0287	proteína del sistema de secreción tipo VII EsxS	1.37E+03
19	Rv2442c	proteína ribosomal L21 de la subunidad 50S	1.30E+03
20	Rv0350	proteína de fusion	1.04E+03

4. Comparación del secretoma y transcriptoma de *M. tuberculosis* al día 21 post-infección.

Comparamos los 529 genes más expresados con el secretoma predicho de Mtb previamente publicado (68), y observamos que 16.07% (85 genes) eran potencialmente secretados (Tabla Anexa 5), un valor 4% más alto que el porcentaje de proteínas secretadas presentes en el genoma completo de Mtb. Es decir, de acuerdo a la predicción del secretoma publicada previamente (68), en el genoma completo de Mtb observamos que el 12% corresponden a genes del secretoma. Si tomamos 529 genes de Mtb al azar esperaríamos que el ~12% (63 genes) pertenecieran al secretoma, sin embargo, observamos que en los 529 genes expresados hay un porcentaje mayor (~16%), lo que sugiere, un ligero aumento en el número de genes del secretoma expresados durante esta etapa de la infección. Adicionalmente para explorar si esta observación era azarosa, formamos 100 grupos de 529 genes elegidos al azar de los 4234 genes totales de Mtb y los comparamos con el secretoma. Observamos que 41 de los 100 grupos tenía más de 85 genes que alineaban al secretoma (E-value <0.001 y cobertura >70%). Esto sugiere que el aumento de genes del secretoma que observamos tiene una probabilidad del 41% de ser un evento al azar.

Es importante mencionar que para entender la relevancia de estas observaciones en el contexto de la infección, es necesario comparar con días antes y después del día 21 pos-infección.

Adicionalmente, el 28.24% (24 proteínas) de las 85 proteínas secretadas son de la familia PE-PGRS, incluyendo la proteína PPE68, la cual se sabe es muy eficiente para estimular la respuesta inmune del hospedero y puede llegar a causar daño inflamatorio en los tejidos (82, 83) (Tabla Anexa 5). Es importante notar que aproximadamente el 10% del genoma de Mtb codifica para proteínas de la familia PE-PGRS (84) y que en nuestro caso, de los 529 genes expresados, sólo el 5.1% corresponde a estas proteínas. Esto puede deberse a que la mayoría de estas proteínas están expresadas por debajo de nuestros límites de detección o a que la presencia de estas proteínas toma mayor relevancia en etapas más avanzadas de la infección.

5. Transcriptoma de ratón en un modelo in vivo al día 21 post infección.

Para analizar la expresión génica del hospedero, secuenciamos el ARN poliadenilado de los mismos ratones infectados usados para el transcriptoma de la bacteria (Figura 4 y 6). Después del mapeo y normalización de la abundancia de expresión por RPKM, observamos una buena reproducibilidad biológica con una correlación de Pearson promedio de r=0.9946 (Figura 6) entre las tres réplicas biológicas. Posteriormente, de los 23,973 genes totales expresados, seleccionamos los 15,677 genes con al menos diez lecturas en dos de las muestras para estudios posteriores.

El análisis de términos GO, considerando el 5% de los genes más expresados mostró un enriquecimiento significativo (p<0.05) de términos relacionados con inflamación, inmunidad adaptativa, péptidos de unión a antígeno, actividad de quimiocinas, unión a receptores de células T, procesamiento y presentación de antígenos, regulación de la diferenciación de leucocitos, producción de citocinas, respuesta celular a lipopolisacáridos y a moléculas de origen bacteriano (Tabla Anexa 7). Así mismo, el análisis de vías metabólicas KEGG mostró procesamiento y presentación de antígeno, fagolisosoma, señalización de IL-17 y TNF, y señalización de receptores Toll como algunas de las vías más representadas (Tabla Anexa 8).

De los 20 genes más expresados, observamos dos proteínas surfactantes (Sftpc y Sftpa) que participan en la modulación de la fagocitosis, tres genes de histocompatibilidad tipo II, el microARN 6381.

En la sección de Anexos se encuentra el artículo correspondiente a esta sección.

VI.II Predicción del secretoma de *M. tuberculosis* y *M. abscessus*.

Dada la importancia del secretoma para la sobrevivencia e interacción de Mtb con su hospedero y por ser una fuente importante de proteínas inmunogénicas que pueden ser candidatas a vacunas, blancos terapéuticos y de diagnóstico, en esta sección mostramos una estrategia bioinformática que permitií predecir sistemáticamente el secretoma a partir de genomas secuenciados.

Esta estrategia la utilizamos para obtener el secretoma de: Mtb H37Rv, *M. abscessus* (Mabs) ATCC 19977 utilizando los genomas depositados en NCBI, y 17 aislados clínicos de Mtb y Mabs, los cuales secuenciamos y ensamblamos para este estudio. En la Tabla VI se enlistan todos los genomas analizados.

Los ensambles finales de los dos aislados de Mtb Beijing resultaron en 151 contigs para el aislado 46 y 144 contigs para el aislado 48, con una profundidad de secuenciación aproximada de 71-72x y un promedio de 3,722 proteínas. Por otro lado, para los quince aislados de Mabs obtuvimos ensambles con 38-78 contigs (media= 58 contigs), con una cobertura entre 217-368x (media=310x) y un promedio de 5,082 proteínas.

Al analizar el promedio de identidad de nucleótidos (ANI) entre los aislados y las cepas de referencia, observamos que los dos aislados de Mtb Beijing tenían una identidad promedio de 99.9% con una cobertura del 99.7% al genoma de referencia. Mientras que los 15 aislados de Mabs tenían una identidad promedio de 98.1% con una cobertura del 91.2% al genoma de referencia. Lo que sugiere que las diferencias entre los aislados y los genomas de referencia correspondientes son muy pocas.

Una característica interesante entre los secretomas predichos de Mtb y de Mabs fue que para Mtb H37Rv y los dos aislados clínicos Beijing obtuvimos un promedio de 540 proteínas secretadas, lo que representa ~12% del genoma, mientras que para Mabs y los 15 aislados clínicos obtuvimos un promedio de 939 proteínas secretadas que representa el ~18% del genoma (Tabla VI). Además, al comparar sólo los secretomas de las cepas de referencia, observamos que hay 337 proteínas compartidas entre Mtb H37Rv y Mabs ATCC19977.

En general, la anotación de los secretomas con términos GO y familias de InterPro, mostró términos esperados para proteínas secretadas en todos los secretomas, como “región extracelular”, “periferia celular” y “estructura de encapsulamiento externa”.

En contraste, observamos que los dominios de PPE y PGRS en el secretoma de Mtb representaban un ~12%, mientras que para Mabs sólo representaban un 0.3%. Así mismo, el análisis de vías KEGG mostró una abundancia alta de la vía “Quorum sensing” para el secretoma de Mabs, mientras que en Mtb, esta vía no está presente. Esto puede estar relacionado a la habilidad de Mabs para formar biofilms, lo cual contribuye a su tolerancia a antibióticos (85, 86). Si bien las micobacterias en general tienen la capacidad de formar biofilms, varios estudios han mostrado que su proceso de formación y composición varía dependiendo de la especie (87). Por ejemplo, en el caso de las micobacterias no tuberculosas como Mabs, la formación de biofilms se debe a los glicopeptidolípidos en la superficie de la bacteria (85). Este biofilm, les da la capacidad de adherirse a superficies y biomateriales, lo que les permite sobrevivir más tiempo en diferentes ambientes, tanto dentro como fuera del hospedero (85, 87).

Tabla VI. Secretomas predichos de las cepas de referencia de *M. tuberculosis* y *M. abscessus.*

Cepa	Id Genoma	Origen	Morfotipo	Proteínas totales	Proteínas secretadas	% proteínas secretadas
M. tuberculosis H37Rv	referencia (GenBank AL123456.3)	-	-	4,337	548	12.64
M. abscessus subsp. abscessus	referencia ATCC19977 (GenBank CU458896.1)	-	-	4,942	886	17.93
M. tuberculosis Beijing	aislado 46	esputo	-	3,702	553	14.94
	aislado 48	esputo	-	3,743	519	13.87
M. abscessus subsp. abscessus	4549-15	esputo	rugoso	5,105	929	18.2
	11351-15	esputo	rugoso	5,138	966	18.8
	8844-15	piel	liso	4,854	956	19.7
	3563-15	esputo	liso	5,239	968	18.48
	12389-15	esputo	liso	5,276	990	18.76
	2677-16	esputo	liso	4,900	919	18.76
	2572-17	tejido (implante mamario)	NA	4,847	874	18.03
M. abscessus subsp. massiliense	14479-15	esputo	rugoso	5,120	962	18.79
	10896-16	esputo	rugoso	5,109	950	18.59
	10003-15	esputo	liso	4,835	891	18.43
	16155-15	esputo	liso	4,884	898	18.39
M. abscessus subsp. bolletii	11702-16	esputo	rugoso	5,079	931	18.33
	713-16	nodo linfático	rugoso	5,456	1,037	19.01
	7742-15	sangre	liso	4,913	885	18.01
	13116-16	nodo linfático	liso	5,305	990	18.66

Adicionalmente, dado que el secretoma se considera una fuente importante de proteínas antigénicas, determinamos y comparamos la Abundancia de Regiones Antigénicas (AAR) de las proteínas secretadas y las no secretadas. Esta métrica, determina la densidad antigénica de una secuencia de aminoácidos dividiendo la longitud de la secuencia entre el número de regiones antigénicas (76), por lo tanto, valores bajos de AAR indican una mayor densidad antigénica y viceversa.

En general, observamos que los valores predichos de abundancia antigénica de las proteínas secretadas fueron significativamente ($p<0.001$) más bajos que para las proteínas no secretadas, mientras que cuando separamos las proteínas no secretadas en intracelulares y transmembranales, las proteínas intracelulares mostraron ser más antigénicas que las transmembranales y que las secretadas. Además, observamos que los secretomas de los aislados clínicos Beijing de Mtb mostraron la mayor antigenicidad (AAR promedio 37.53) de todos los genomas analizados (AAR promedio 40.9), lo cual puede estar relacionado con la alta virulencia que este genotipo ha mostrado tanto en humanos como en modelos animales (12-14).

Por otro lado, al comparar los secretomas de las cepas de Mabs que producen colonias con un morfotipo rugoso (R) contra las cepas de morfotipo liso (L), observamos secretomas más grandes en las cepas R de los cuales, cerca del ~90% está compartido con las cepas L. Por otro lado, las proteínas secretadas únicas de las cepas R mostraron mayor densidad antigénica que las proteínas únicas para las cepas L.

Finalmente, para evaluar la validez de la estrategia bioinformática comparamos los secretomas predichos de Mtb contra 338 proteínas reportadas experimentalmente como secretadas y observamos que ~70% del secretoma experimental estaba incluido en los secretomas predichos. El 30% restante no fue identificado por ninguno de los programas (SignalP, SecretomeP, LipoP o TatP). Es posible que estas proteínas no tengan los motivos más comunes en su secuencia para indicar la secreción y por ello fueron negativos a la búsqueda bioinformática. Para corroborar que esta coincidencia no fuera al azar construímos 1000 secretomas con proteínas aleatorias del mismo tamaño de los secretomas predichos (548, 553 y 519 proteínas) y comparamos cada uno con el secretoma experimental. En este caso, observamos una coincidencia máxima del 40%, indicando que la coincidencia de 70% con los secretomas predichos no fue al azar. Adicionalmente comparamos los secretomas de Mtb de dos estudios similares previos (49, 50) y observamos una coindicencia del 34.32% con el estudio de Roy et al. (2013) y del 41.42% con el estudio de Vizcaíno et al. (2010), lo que indica que nuestra estrategia bioinformática hace una predicción más completa de las proteínas secretadas de Mtb. La diferencia en los porcentajes de coincidencia puede deberse a que la estrategia bioinformática usada por los tres estudios es diferente. En el estudio de Vizcaíno et al. (49), usan los mismos programas (SignalP, SecretomeP, LipoP y TatP) para un primer filtrado de las proteínas. Después, usan TMHMM y tres predictores generales de localización subcelular (Gpos-PLoc, PSORTb v2.0.4 y PA-SUB v2.5) como filtros para seleccionar ocho proteínas candidatas a un análisis experimental de secreción y antigenicidad. Al final observaron que de las ocho proteínas candidatas, seis mostraban evidencia experimental de ser secretadas. El objetivo principal de este estudio no era reportar el secretoma completo de Mtb, sino evaluar la eficiencia de los programas para predecir proteínas secretadas y proponer candidatos a vacunas.

Por otro lado, el trabajo de Roy et al.(50), utiliza SignalP como programa de partida para filtrar todas las proteínas de Mtb. Y posteriormente, toma las proteínas que contenían un péptido señal (positivas a SignalP) y las analiza de manera independiente con TMHMM, Pred-Lipo y TatFind. Finalmente compararon sus resultados con el estudio de Leversen et al. 2009, cuyos resultados están incluídos en nuestro secretoma experimental, y observaron una coincidencia del 81% en las proteínas predichas con péptido señal. Notablemente, esta estrategia deja fuera a las proteínas secretadas que no tienen un péptido señal.

Lamentablemente, los estudios experimentales sobre secretomas de Mabs son limitados y no fue posible hacer una comparación similar con los genomas de Mabs.

Los resultados de esta sección se publicaron en dos artículos científicos anexos al final de la tesis, en los cuáles se pueden consultar los resultados a profundidad.

VI.III Implementación del servidor web Secret-AAR para la búsqueda de homología entre proteínas del secretoma de micobacterias y para determinar la densidad antigénica de proteínas.

Implementamos el servidor web de uso público Secret-AAR con dos objetivos principales: i) identificar (vía BLASTp) si una proteína de interés pertenece al secretoma predicho o experimental de Mtb o Mabs analizados descritos en la sección anterior, y ii) determinar la abundancia de regiones antigénicas (AAR) de cualquier secuencia de aminoácidos.

Para usar Secret-AAR (http://microbiomics.ibt.unam.mx/tools/aar/index.php) el usuario debe elegir una de las dos herramientas disponibles (análisis del secretoma o cálculo de AAR) y someter la secuencia de aminoácidos de interés en formato fasta directo en la caja de texto o como archivo de texto.

1. Análisis del secretoma.

Con esta herramienta, el usuario puede explorar si su proteína de interés forma parte de los secretomas predichos o experimentales de Mtb o Mabs disponibles en el sitio. Para esto, Secret-AAR toma la secuencia de aminoácidos del usuario y la compara vía BLASTp contra los secretomas disponibles. En total hay 22 secretomas de micobacterias disponibles: 4 cepas de referencia y 2 aislados clínicos de Mtb, y 1 cepa de referencia y 15 aislados clínicos de Mabs.

De esta lista, el usuario puede elegir uno o más secretomas al mismo tiempo para realizar el análisis. El servidor soporta hasta 1,000 secuencias por cada análisis con una longitud máxima de 5,000 aminoácidos por secuencia.

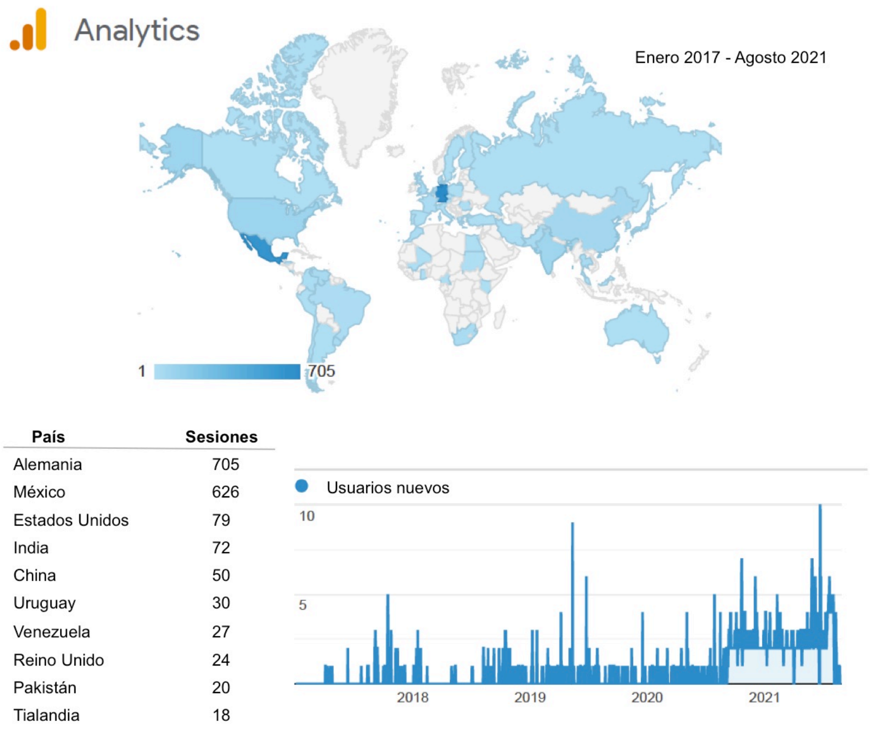
Los resultados contienen todos los datos del BLASTp (identidad, cobertura, E-value) para todos los hits contra la base de datos elegida. Así mismo, si la secuencia tiene homología contra proteínas en los secretomas experimentales, como parte del resultado se muestra el DOI del artículo donde la proteína fue reportada originalmente.

2. Determinación de la Abundancia de Regiones Antigénicas (AAR).

Para realizar el cálculo de AAR, el servidor predice las regiones antigénicas con los algoritmos de Bepipred1.0 (77) estableciendo el score límite sugerido por los autores de 0.35 que otorga una especificidad del 75% (77). Después, utilizando scripts desarrollados en casa, cuenta el número de regiones antigénicas con una longitud mínima de 6 aminoácidos. Y finalmente, divide la longitud de la proteína entre el número de regiones antigénicas.

Como resultado, el servidor muestra una tabla con el nombre y la longitud de la proteína "query", el número de regiones antigénicas y el valor de AAR. Así mismo, cada secuencia tiene la opción de mostrar la salida original de BepiPred para visualizar cada una de las regiones antigénicas identificadas. El servidor soporta hasta 1,000 secuencias por cada análisis con una longitud máxima de 5,000 aminoácidos por secuencia.

Desde su lanzamiento en 2017 el servidor ha recibido 1,651 visitas de 10 países diferentes, siendo Alemania y México los dos países con más visitas registradas (Figura 9). El sitio está en constante actualización y es posible que haya más secretomas disponibles en el futuro. Anexo al final de esta tesis se encuentra la publicación correspondiente donde se pueden consultar mayores detalles sobre este servidor.



País	Sesiones
Alemania	705
México	626
Estados Unidos	79
India	72
China	50
Uruguay	30
Venezuela	27
Reino Unido	24
Pakistán	20
Tialandia	18

Figura 9. Estadística de uso del servidor Secret-AAR de Enero 2017-Agosto 2021 (Obtenido de Google Analitycs https://analytics.google.com/).

VII. Discusión

La estrategia diseñada para obtener el transcriptoma de *M. tuberculosis* (Mtb) y su hospedero a partir de un modelo *in vivo* utilizando secuenciación masiva (RNA-seq) buscaba sobrepasar dos de las limitaciones técnicas más importantes: i) obtener suficiente ARN de bacteria de buena calidad y ii) disminuir de manera eficiente el ARN ribosomal para poder observar la expresión del mayor número posible de genes de Mtb.

Para superar la primera limitación, establecimos un protocolo de lisis celular diferencial utilizando dos soluciones de lisis. Primero, una solución de lisis suave, permitió romper las células pulmonares de ratón dejando intacta a la micobacteria gracias a su compleja pared celular externa rica en lípidos (88). Después, con una solución de lisis más fuerte, logramos romper la pared celular de Mtb y extraer el ARN total manteniendo una buena calidad.

Es importante mencionar que estudios previos de Mtb habían aplicado protocolos similares de lisis diferencial en cultivos de células *in vitro* y en tejidos infectados, juntando varias muestras en una sola extracción y posteriormente analizando la expresión con microarreglos. Sin embargo, con nuestro protocolo pudimos obtener ARN de calidad y cantidad suficientes para realizar RNA-seq de pulmones de ratón individuales, ofreciendo la posibilidad de estudiar las respuestas individuales a la infección por Mtb y de utilizar un menor número de ratones para los experimentos.

Los datos de secuenciación del ARN enriquecido en micobacteria mostraron que a pesar de haber usado un kit para eliminación de ribosomales, el 97% de las secuencias de Mtb pertenecían a transcritos ribosomales. Algunos estudios han reportado que el uso de kits para depleción remueven un 70-85% de ARN ribosomal en cultivos líquidos de Mycobacterium (46), sin embargo, nuestras muestras derivan de tejido infectado y tienen una mezcla de células de ratón y micobacteria, lo que podría haber disminuido la eficiencia en la depleción.

El uso de sondas biotiniladas específicas para ARN ribosomal de Mtb disminuyó un 24% los transcritos ribosomales y aumentó el número de genes observados, de 13 genes siguiendo la estrategia tradicional para RNA-seq a 702. Como se mencionó anteriormente, aunque la profundidad de secuenciación en las diferentes estrategias no es comparable, el ejercicio de subsampleó mostró que una aún una menor profundidad de secuenciación en E3 muestra 10 veces más genes que E2.

La caracterización funcional de los 529 genes más expresados de Mtb mostró un perfil de producción de energía y biosíntesis de macromoléculas, el cual, es consistente con un crecimiento exponencial de la micobacteria, además, la presencia de vías como el glioxilato sugieren que la bacteria está en un ambiente de estrés y utiliza lípidos como fuente de carbono para su metabolismo (44). Por otro lado, la expresión de genes relacionados con la estimulación de células T y el reclutamiento de células inflamatorias, coincide con el momento de la infección que analizamos (día 21), donde se lleva a cabo la formación de granulomas en el pulmón (27, 28). Esta respuesta de células T se ha reportado que en modelos de ratón inicia desde el día 14 post-infección (89).

El gen de Mtb más expresado fue una transposasa para la secuencia de inserción IS1081, un elemento génico que tradicionalmente se ha usado como marca para la genotipificación de cepas de *M. bovis* and *M. tuberculosis* y que no tiene reportes previos que la relacionen con mecanismos de infección (90). Caso contrario a la secuencia de inserción IS6110, que además de ser un marcador molecular (91) ha sido sugerida como promotor de la expresión de genes vecinos afectando la adaptación de la micobacteria (91) Sin embargo, la transposasa para IS6110 fue observada en los últimos lugares de nuestra lista de genes expresados. Además, de acuerdo a la anotación del genoma de la cepa infecciosa de Mtb observamos 33 regiones para transposasas e IS6110, pero sólo 6 regiones para transposasas de IS1081, lo que sugiere que la alta expresión de la transposasa para IS1081 que observamos no está relacionada con un alto número de copias del gen, y más bien podría jugar un papel regulatorio durante la infección por Mtb. Es importante notar que estamos trabajando con un ensamble que no se puede considerar como un genoma de referencia, por lo que estas observaciones sólo son basadas en la anotación inicial de los genes.

La importancia de las secuencias de inserción en la fisiología de Mtb se ha estudiado sobre todo para IS6110. Algunos autores proponen que puede afectar la expresión de genes recorriendo el marco de lectura para la transcripción o modificando la estructura secundaria del ARN mensajero impidiendo la traducción (92). Además se observó una actividad alta de la transposasa de IS6110 en ratones con infecciones crónicas con Mtb H37Rv y en cultivos líquidos deficientes en nutrientes (92). Esto sugiere que la transposición de IS6110 es dinámica y responde a condiciones de estrés como las que se dan durante la infección in vivo. Por lo tanto, dada la alta expresión que observamos de la transposasa IS1081, sería importante estudiar si esta región juega también un papel en el desarrollo de Mtb en diferentes ambientes.

Adicionalmente como parte de los genes más expresados, observamos genes de la familia PE-PGRS. En particular, PGRS49 y PGRS50 que han sido sugeridos como fuertes candidatos a vacunas (93), y PE-PGRS9 y PE-PGRS53, cuya presencia había sido previamente relacionada con etapas tardías de la enfermedad (78, 94). De los estudios transcriptómicos previos, sólo el estudio de Talaat et al. (22) reportó ocho genes de la familia PGRS con expresión en ratones BALB/c, sin embargo se ubicaron por debajo del lugar 30 de expresión. Esta familia de proteínas representa aproximadamente un 10% de los genes de Mtb (84, 95) pero su papel biológico aún no está completamente determinado. Muchas de ellas son proteínas de membrana que representan una fuente de variación antigénica, y otras son proteínas secretadas que interactúan con el hospedero despertando la respuesta inmune (78, 80, 82). Sin embargo, hasta la fecha no se sabe qué miembros específicos de esta familia son esenciales en la infección in vivo o si determinan la virulencia de las diferentes cepas. Hasta nuestro conocimiento, este es el primer estudio de transcriptoma que resalta el papel de esta familia como los genes más expresados durante la infección de Mtb *in vivo*.

Por otro lado, al comparar los 529 genes más expresados con el secretoma de Mtb, encontramos un ligero incremento en la proporción de proteínas secretadas (16.07%) a comparación con el porcentaje de proteínas secretadas presentes en el genoma (~12%) (68). Y notablemente, este porcentaje aumentó al 27.72% cuando analizamos los 101 genes más expresados. Será importante comparar estos porcentajes en estudios donde se analicen diferentes cepas o diferentes tiempos de la infección.

El 16.07% de genes que codifican proteínas secretadas representan 85 genes, entre los cuales, encontramos siete lipoproteínas, entre ellas lppN (Rv2270) y LpqH (Rv3763), las cuales se ha sugerido que ayudan a la internalización de la bacteria en los macrófagos. En particular, LpqH ha sido ampliamente estudiada y se considera una proteína importante en la interacción con las células del hospedero (96). Algunos de sus efectos favorecen al hospedero como la inducción de IL-12 que aumenta la respuesta inmune, mientras que otros favorecen a la bacteria como inhibir la señalización de INF gama, disminuir la expresión de MHC-II y el procesamiento de antígenos (97-99). En particular se ha observado que la producción de IL-12 es importante para iniciar la respuesta inmune celular que controla a Mtb. Participa en la estimulación de infocitos T y en regular la migración de células dendríticas (100). Y a su vez, se ha observado un aumento en la producción de IL-12 cuando hay una sobreexpresión de la liporoteína LpqH (97). Por otro lado, se sabe que la expresión de INF gama en el hospedero es esencial para la defensa contra Mtb. Esta citocina participa en la activación de macrófagos y regula la expresión de MHC-II. Sin embargo, se ha observado que cuando los macrófagos se exponen a un lisado de bacteria, disminuye la expresión de MHC-II y la producción de INF-gama. Posterirmente, al caracterizar los componentes del lisado que tenían este efecto, se observó que uno de los más importantes era la lipoproteína LpqH (99), favoreciendo entonces, que la bacteria persista dentro de las células.

Interesantemente, ~28% de las 85 proteínas secretadas expresadas son de la familia PE-PGRS. Entre ellas observamos a la PPE68, la cual es codificada en la región de diferenciación RD1 y por lo tanto está ausente en cepas avirulentas de Mtb como BCG (101). Además, se considera un inmunomodulador muy eficiente para estimular la respuesta inmune del hospedero pudiendo incluso causar daño inflamatorio en los tejidos. Interesantemente, aunque la proteína PPE68 no se considera esencial para la vida de Mtb, algunos estudios han reportado que alteraciones en su secuencia afectan la virulencia de la cepa (101). Esta proteína ocupa el lugar 25 de los 529 genes más expresados.

Una de las ventajas de usar RNA-seq es la posibilidad de analizar la expresión de regiones no codificantes. En nuestro análisis, el 38.41% de las secuencias de Mtb mapearon a regiones intergénicas. Algunos estudios reportan que la expresión de regiones intergénicas en cultivos líquidos en fase logarítmica está alrededor del 28%, mientras que en cultivos en fase estacionaria sube al 58% (102), sugiriendo que la expresión de las regiones intergénicas es variable a lo largo del desarrollo de Mtb en diferentes ambientes.

En particular, el ARN pequeño MTS2823 fue el más expresado de los tres observados, acumulando el 42.34% de las lecturas. La expresión de este ARN pequeño se ha encontrado acumulada en cultivos en fase estacionaria (102), pulmones de ratón con tuberculosis crónica (102), e incluso en micobacterias latentes (103). Algunos autores sugieren que al menos en cultivos, su función es regular a la baja la expresión de genes durante la fase exponencial de crecimiento (102), sin embargo, su mecanismo aún no ha sido definido. Nuestros resultados sugieren que MTS2328 también tiene un papel importante durante etapas tempranas de la infección y sería importante comparar si su expresión continua aumentando en etapas crónicas como un mecanismo de regulación que favorece la persistencia de la infección. .

En relación al hospedero, predomina la expressión de genes de citocinas pro-inflamatorias lo cual es esperado para las etapas tempranas de la infección por Mtb especialmente en este modelo experimental murino (28) y es consistente con otros estudios (104). Así mismo, observamos varios genes para la co-estimulación de linfocitos y células T, reclutamiento celular y procesamiento de antígenos. Este, es un proceso que se ha reportado inicia cerca del día 14-21 post-infección y se expande hacia las etapas crónicas de la enfermedad (105). Sobre todo, algunos autores sugieren que la presencia de células T en etapas tempranas de la infección está relacionado con la formación de granulomas (106), donde los macrófagos alveolares y las células epiteliales que contienen a la micobacteria presentan antígenos y activan a las células T para producir una variedad de citocinas y quimiocinas para continuar el reclutamiento celular y el mantenimiento del granuloma (106). Hay que notar que esta caracterización se basa en un solo día post-infección y coincide con algunos estudios previos en etapas similares, sin embargo, para entender la relevancia de estas funciones es necesario comparar con otros días más tempranos y tardíos de la infección.

Además observamos las proteínas surfactantes Sftpc y Sftpa dentro de los 20 genes más expresados. Particularmente, Sftpa se conoce que actúa como opsonina para favorecer la internalización de MTb y otros patógenos a los macrófagos alveolares (107-109), aunque interesantemente, esta vía de internalización favorece la disminución de las especies reactivas de nitrógeno (109), sugiriendo que la unión a Sftpa es un mecanismo que favorece la supervivencia de Mtb dentro del macrófago.

Finalmente, observamos la expresión del microRNA mi6381 en el lugar veinte de los genes más expresados. Hasta ahora, la expresión de este microRNA no ha sido relacionada con enfermedades pulmonares de ningún tipo. Varios estudios *in vitro* sugieren que los microRNAs como mir155, mir135b, o mir146a juegan un papel importante durante la infección por Mtb (110-114). Particularmente, mir155 se encontró sobre expresado en células mononucleares de sangre pacientes con tuberculosis en comparación con pacientes sanos (115). Sin embargo, no observamos a mir155 en nuestros resultados, sugiriendo que diferentes miRNAs pueden estar activos en diferentes tiempos de la infección o en diferentes tejidos.

La estrategia experimental propuesta para obtener los genes expresados de Mtb mostró ser reproducible y permitió observar el conjunto de genes más expresados durante la infección por Mtb. Pudimos observar la expresión de proteínas antigénicas (PGRS) y secretadas (lipoproteínas y PGRS) que juegan un papel importante en la interacción con el hospedero.

Aunque reconocemos que la cantidad de genes analizada (529 genes) es limitada, el principal objetivo era demostrar la reproducibilidad y eficacia de la metodología para enriquecer el ARN bacteriano y estudiar la respuesta de Mtb durante la infección *in vivo*. Consideramos que esta estrategia es una buena alternativa para analizar los genes más expresados de Mtb utilizando RNA-seq cuando no se tiene la posibilidad de una mayor profundidad de secuenciación o cuando no se cuenta con equipo adicional para separar las células infectadas y enriquecer el ARN de la bacteria (24).

El estudio del secretoma es esencial para entender la interacción patógeno-hospedero en cualquier infección, sin embargo, los estudios que buscan describir el secretoma completo de micobacterias como Mtb y Mabs son limitados. En ese sentido, la estrategia bioinformática que detallamos ayudaó a predecir de manera sistemática el secretoma de los genomas de Mtb y Mabs.

Lo primero que observamos, fue que el secretoma de Mtb, tanto en la cepa de referencia como en los aislados clínicos, representa ~12% de las proteínas totales del genoma, mientras que para Mabs, este porcentaje sube al ~18%. Esta diferencia, no es de sorprender, ya que a la fecha hay cerca de 200 especies de micobacterias, entre las cuales, varios estudios sugieren que se comparten muy pocos genes (~1000 genes) (116, 117). Además, las micobacterias no tuberculosas como Mabs se considera que representan un estado evolutivo más antiguo con genomas más grandes (117, 118). Por otro lado, la presencia de más proteínas secretadas en Mabs pudiera estar relacionada con su habilidad para causar un mayor espectro de enfermedades y de adaptarse a un mayor número de ambientes (119-121).

Además, observamos que los dominios PPE y PGRS en el secretoma de Mtb representaban un ~12%, mientras que para Mabs sólo representaban un 0.3%. Esta diferencia puede relacionarse con los sistemas de secreción presentes en cada especie. En *M. tuberculosis*, los operones ESX1, 2, 3 y 5 contienen genes para proteínas PE-PGRS, de los cuales, se ha observado que ESX-5 transporta proteínas importantes para la virulencia (122). En contraste, Mabs sólo tiene dos sistemas de secreción (ESX-3 y ESX-4), de los cuales sólo el operón de ESX-3 tiene genes para PE/PGRS (123).

Por otro lado, al comparar los secretomas de las cepas de Mabs que producen colonias con un fenotipo rugoso (R) contra las cepas de fenotipo liso (L), observamos secretomas más grandes en las cepas R de los cuales, cerca del ~90% está compartido con las cepas L. Además, las proteínas secretadas únicas de las cepas R mostraron mayor densidad antigénica que las proteínas únicas para las cepas L. El alto porcentaje de proteínas secretadas compartidas no es de sorprender, ya que cepas con morfotipo R pueden surgir a partir de cepas con morfotipo L. Este fenómeno se ha descrito tanto en pacientes como en ratones donde después de una infección inicial con una cepa de morfotipo R, se aislaron cepas de morfotipo L que mostraron una habilidad disminuida de producir glicopeptidolípidos (GPLs) y una virulencia atenuada (124-127).

Es importante mencionar que si bien el análisis de identidad de nucleótidos sugiere que los aislados son muy similares a las cepas de referencia, no podemos descartar que las diferencias entre los secretomas se deban a variaciones a nivel genómico

Finalmente el desarrollo de una estrategia bioinformática que permita predecir de manera sistemática el secretoma de cualquier genoma disponible de micobacterias como Mtb y Mabs facilitará el análisis y la comparación con datos de expresión genética para conocer qué proteínas secretadas son importantes en diferentes puntos de la infección y con diferentes cepas. Estrategias como esta pueden ayudar a guiar el diseño de experimentos y priorizar la búsqueda de blancos terapéuticos o moléculas de diagnóstico.

VIII. Conclusiones

El uso de la lisis celular diferencial y las sondas de captura de ARN ribosomal permitió describir los genes y regiones intergénicas más expresadas de Mtb y su hospedero en un modelo *in vivo*.

Por otro lado, la estrategia bioinformática nos permitió predecir de manera sistemática el secretoma completo de *M. tuberculosis* y *M. abscessus,* así como evaluar su potencial antigénico. Ambas estrategias, pueden ser utilizadas para estudiar la expresión génica o para conocer el secretoma de otras cepas de *M. tuberculosis.*

IX. Perspectivas

Actualmente, tanto el protocolo experimental, como la estrategia para el análisis de los datos se están utilizando para estudios de comparación utilizando diferentes cepas de Mtb a diferentes tiempos post-infección.

Así mismo, se planea optimizar el uso de las sondas para captura de ribosomales en días post-infección donde el número de bacterias es significativamente mayor al analizado en este trabajo.

Por otro lado, el sitio web SecretAAR está en actualización continua y se continúa agregando secretomas nuevos a las bases de datos.

IX. Referencias

1. W. H. Organization (2020) Global Tuberculosis Report 2020. in *Licence: CC BY-NC-SA 3.0 IGO* (Geneva).
2. J. L. Flynn, J. Chan, Immunology of tuberculosis. *Annu Rev Immunol* **19**, 93-129 (2001).
3. G. A. Rook, K. Dheda, A. Zumla, Immune responses to tuberculosis in developing countries: implications for new vaccines. *Nat Rev Immunol* **5**, 661-667 (2005).
4. R. M. Houben, P. J. Dodd, The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS Med* **13**, e1002152 (2016).
5. M. Coscolla, S. Gagneux, Does M. tuberculosis genomic diversity explain disease diversity? *Drug Discov Today Dis Mech* **7**, e43-e59 (2010).
6. I. Comas, S. Gagneux, The past and future of tuberculosis research. *PLoS Pathog* **5**, e1000600 (2009).
7. S. T. Cole *et al.*, Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537-544 (1998).
8. M. Coscolla, S. Gagneux, Consequences of genomic diversity in Mycobacterium tuberculosis. *Semin Immunol* **26**, 431-444 (2014).
9. F. Gehre *et al.*, Deciphering the growth behaviour of Mycobacterium africanum. *PLoS Negl Trop Dis* **7**, e2220 (2013).
10. R. Sarkar, L. Lenders, K. A. Wilkinson, R. J. Wilkinson, M. P. Nicol, Modern lineages of Mycobacterium tuberculosis exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages. *PLoS One* **7**, e43170 (2012).
11. N. Krishnan *et al.*, Mycobacterium tuberculosis lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. *PLoS One* **6**, e23870 (2011).
12. B. Lopez *et al.*, A marked difference in pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes. *Clin Exp Immunol* **133**, 30-37 (2003).
13. N. Reiling *et al.*, Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice. *mBio* **4** (2013).
14. M. Kato-Maeda *et al.*, Beijing sublineages of Mycobacterium tuberculosis differ in pathogenicity in the guinea pig. *Clin Vaccine Immunol* **19**, 1227-1237 (2012).
15. A. van Laarhoven *et al.*, Low induction of proinflammatory cytokines parallels evolutionary success of modern strains within the Mycobacterium tuberculosis Beijing genotype. *Infect Immun* **81**, 3750-3756 (2013).
16. S. Gagneux, P. M. Small, Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *Lancet Infect Dis* **7**, 328-337 (2007).
17. J. Gonzalo-Asensio *et al.*, Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl Acad Sci U S A* **111**, 11491-11496 (2014).
18. G. Rose *et al.*, Mapping of genotype-phenotype diversity among clinical isolates of mycobacterium tuberculosis by sequence-based transcriptional profiling. *Genome Biol Evol* **5**, 1849-1862 (2013).
19. P. Supply *et al.*, Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. *Nat Genet* **45**, 172-179 (2013).

20. I. Comas *et al.*, Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* **45**, 1176-1182 (2013).
21. D. Schnappinger *et al.*, Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages: Insights into the Phagosomal Environment. *J Exp Med* **198**, 693-704 (2003).
22. A. M. Talaat, R. Lyons, S. T. Howard, S. A. Johnston, The temporal expression profile of Mycobacterium tuberculosis infection in mice. *Proc Natl Acad Sci U S A* **101**, 4602-4607 (2004).
23. G. Lamichhane, S. Tyagi, W. R. Bishai, Designer arrays for defined mutant analysis to detect genes essential for survival of Mycobacterium tuberculosis in mouse lungs. *Infect Immun* **73**, 2533-2540 (2005).
24. D. Pisu, L. Huang, J. K. Grenier, D. G. Russell, Dual RNA-Seq of Mtb-Infected Macrophages In Vivo Reveals Ontologically Distinct Host-Pathogen Interactions. *Cell Rep* **30**, 335-350 e334 (2020).
25. R. A. Rienksma *et al.*, Comprehensive insights into transcriptional adaptation of intracellular mycobacteria by microbe-enriched dual RNA sequencing. *BMC Genomics* **16**, 34 (2015).
26. S. Homolka, S. Niemann, D. G. Russell, K. H. Rohde, Functional genetic diversity among Mycobacterium tuberculosis complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. *PLoS Pathog* **6**, e1000988 (2010).
27. R. Hernandez-Pando *et al.*, Correlation between the kinetics of Th1, Th2 cells and pathology in a murine model of experimental pulmonary tuberculosis. *Immunology* **89**, 26-33 (1996).
28. R. Hernandez-Pando *et al.*, Analysis of the local kinetics and localization of interleukin-1 alpha, tumour necrosis factor-alpha and transforming growth factor-beta, during the course of experimental pulmonary tuberculosis. *Immunology* **90**, 607-617 (1997).
29. H. Tjalsma *et al.*, Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome. *Microbiol Mol Biol Rev* **68**, 207-233 (2004).
30. W. L. Beatty, D. G. Russell, Identification of mycobacterial surface proteins released into subcellular compartments of infected macrophages. *Infect Immun* **68**, 6997-7002 (2000).
31. J. Pieters, J. Gatfield, Hijacking the host: survival of pathogenic mycobacteria inside macrophages. *Trends Microbiol* **10**, 142-146 (2002).
32. I. Smith, Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. *Clin Microbiol Rev* **16**, 463-496 (2003).
33. H. Malen, F. S. Berven, K. E. Fladmark, H. G. Wiker, Comprehensive analysis of exported proteins from Mycobacterium tuberculosis H37Rv. *Proteomics* **7**, 1702-1718 (2007).
34. T. Fifis, C. Costopoulos, A. J. Radford, A. Bacic, P. R. Wood, Purification and characterization of major antigens from a Mycobacterium bovis culture filtrate. *Infect Immun* **59**, 800-807 (1991).
35. M. A. Horwitz, B. W. Lee, B. J. Dillon, G. Harth, Protective immunity against tuberculosis induced by vaccination with major extracellular proteins of Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* **92**, 1530-1534 (1995).
36. A. M. Abdallah *et al.*, Type VII secretion--mycobacteria show the way. *Nat Rev Microbiol* **5**, 883-891 (2007).
37. R. Nessar, E. Cambau, J. M. Reyrat, A. Murray, B. Gicquel, Mycobacterium abscessus: a new antibiotic nightmare. *J Antimicrob Chemother* **67**, 810-818 (2012).

38. M. R. Lee *et al.*, Mycobacterium abscessus Complex Infections in Humans. *Emerg Infect Dis* **21**, 1638-1646 (2015).
39. J. Zheng *et al.*, Analysis of the secretome and identification of novel constituents from culture filtrate of bacillus Calmette-Guerin using high-resolution mass spectrometry. *Mol Cell Proteomics* **12**, 2081-2095 (2013).
40. F. Vargas-Romero *et al.*, Secretome profile analysis of hypervirulent Mycobacterium tuberculosis CPT31 reveals increased production of EsxB and proteins involved in adaptation to intracellular lifestyle. *Pathog Dis* **74** (2016).
41. C. Putim *et al.*, Secretome profile analysis of multidrug-resistant, monodrug-resistant and drug-susceptible Mycobacterium tuberculosis. *Arch Microbiol* **200**, 299-309 (2018).
42. R. Avraham *et al.*, A highly multiplexed and sensitive RNA-seq protocol for simultaneous analysis of host and pathogen transcriptomes. *Nat Protoc* **11**, 1477-1491 (2016).
43. A. J. Westermann, S. A. Gorski, J. Vogel, Dual RNA-seq of pathogen and host. *Nat Rev Microbiol* **10**, 618-630 (2012).
44. K. H. Rohde, R. B. Abramovitch, D. G. Russell, Mycobacterium tuberculosis invasion of macrophages: linking bacterial gene expression to environmental cues. *Cell Host Microbe* **2**, 352-364 (2007).
45. S. J. Waddell, K. Laing, C. Senner, P. D. Butcher, Microarray analysis of defined Mycobacterium tuberculosis populations using RNA amplification strategies. *BMC Genomics* **9**, 94 (2008).
46. S. Wang *et al.*, Revealing of Mycobacterium marinum transcriptome by RNA-seq. *PLoS One* **8**, e75828 (2013).
47. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63 (2009).
48. F. Zhou *et al.*, Protein array identification of protein markers for serodiagnosis of Mycobacterium tuberculosis infection. *Sci Rep* **5**, 15349 (2015).
49. C. Vizcaino *et al.*, Computational prediction and experimental assessment of secreted/surface proteins from Mycobacterium tuberculosis H37Rv. *PLoS Comput Biol* **6**, e1000824 (2010).
50. A. Roy, S. Bhattacharya, A. K. Bothra, A. Sen, A database for Mycobacterium secretome analysis: 'MycoSec' to accelerate global health research. *OMICS* **17**, 502-509 (2013).
51. A. Bankevich *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477 (2012).
52. E. Bosi *et al.*, MeDuSa: a multi-draft based scaffolder. *Bioinformatics* **31**, 2443-2451 (2015).
53. O. Nishimura, Y. Hara, S. Kuraku, gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**, 3635-3637 (2017).
54. A. L. Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636-4641 (1999).
55. A. Conesa, S. Gotz, Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* **2008**, 619832 (2008).
56. A. Kapopoulou, J. M. Lew, S. T. Cole, The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb)* **91**, 8-13 (2011).
57. E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).

58. M. L. Carpenter *et al.*, Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am J Hum Genet* **93**, 852-864 (2013).
59. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
60. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
61. F. Al-Shahrour, R. Diaz-Uriarte, J. Dopazo, FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578-580 (2004).
62. Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, M. Kanehisa, KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35**, W182-185 (2007).
63. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
64. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
65. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).
66. R. K. Aziz *et al.*, The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
67. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
68. F. Cornejo-Granados *et al.*, Secretome Prediction of Two M. tuberculosis Clinical Isolates Reveals Their High Antigenic Density and Potential Drug Targets. *Front Microbiol* **8**, 128 (2017).
69. F. Cornejo-Granados *et al.*, Secretome characterization of clinical isolates from the Mycobacterium abscessus complex provides insight into antigenic differences. *BMC Genomics* **22**, 385 (2021).
70. J. D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795 (2004).
71. J. D. Bendtsen, L. Kiemer, A. Fausboll, S. Brunak, Non-classical protein secretion in bacteria. *BMC Microbiol* **5**, 58 (2005).
72. J. D. Bendtsen, H. Nielsen, D. Widdick, T. Palmer, S. Brunak, Prediction of twin-arginine signal peptides. *BMC Bioinformatics* **6**, 167 (2005).
73. A. S. Juncker *et al.*, Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* **12**, 1652-1662 (2003).
74. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580 (2001).
75. L. Kall, A. Krogh, E. L. Sonnhammer, Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res* **35**, W429-432 (2007).
76. S. Gomez *et al.*, Genome analysis of Excretory/Secretory proteins in Taenia solium reveals their Abundance of Antigenic Regions (AAR). *Sci Rep* **5**, 9683 (2015).
77. J. E. Larsen, O. Lund, M. Nielsen, Improved method for predicting linear B-cell epitopes. *Immunome Res* **2**, 2 (2006).
78. G. Delogu *et al.*, PE_PGRS proteins are differentially expressed by Mycobacterium tuberculosis in host tissues. *Microbes Infect* **8**, 2061-2067 (2006).
79. M. J. Brennan, G. Delogu, The PE multigene family: a 'molecular mantra' for mycobacteria. *Trends Microbiol* **10**, 246-249 (2002).

80. S. Banu *et al.*, Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens? *Mol Microbiol* **44**, 9-19 (2002).

81. W. Bitter, E. N. Houben, J. Luirink, B. J. Appelmelk, Type VII secretion in mycobacteria: classification in line with cell envelope structure. *Trends Microbiol* **17**, 337-338 (2009).

82. D. Bottai, R. Brosch, Mycobacterial PE, PPE and ESX clusters: novel insights into the secretion of these most unusual protein families. *Mol Microbiol* **73**, 325-328 (2009).

83. F. Sayes *et al.*, Strong immunogenicity and cross-reactivity of Mycobacterium tuberculosis ESX-5 type VII secretion: encoded PE-PPE proteins predicts vaccine potential. *Cell Host Microbe* **11**, 352-363 (2012).

84. S. Kohli *et al.*, Comparative genomic and proteomic analyses of PE/PPE multigene family of Mycobacterium tuberculosis H(3)(7)Rv and H(3)(7)Ra reveal novel and interesting differences with implications in virulence. *Nucleic Acids Res* **40**, 7113-7122 (2012).

85. I. M. Orme, D. J. Ordway, Host response to nontuberculous mycobacterial infections of current clinical importance. *Infect Immun* **82**, 3516-3522 (2014).

86. G. Clary *et al.*, Mycobacterium abscessus Smooth and Rough Morphotypes Form Antimicrobial-Tolerant Biofilm Phenotypes but Are Killed by Acetic Acid. *Antimicrob Agents Chemother* **62** (2018).

87. J. Esteban, M. Garcia-Coca, Mycobacterium Biofilms. *Front Microbiol* **8**, 2651 (2017).

88. P. J. Brennan, Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis. *Tuberculosis (Edinb)* **83**, 91-97 (2003).

89. A. J. Wolf *et al.*, Initiation of the adaptive immune response to Mycobacterium tuberculosis depends on antigen production in the local lymph node, not the lungs. *J Exp Med* **205**, 105-115 (2008).

90. M. N. Nghiem, B. V. Nguyen, S. T. Nguyen, T. T. Vo, H. V. Nong, A Simple, Single Triplex PCR of IS6110, IS1081, and 23S Ribosomal DNA Targets, Developed for Rapid Detection and Discrimination of Mycobacterium from Clinical Samples. *J Microbiol Biotechnol* **25**, 745-752 (2015).

91. T. Roychowdhury, S. Mandal, A. Bhattacharya, Analysis of IS6110 insertion sites provide a glimpse into genome evolution of Mycobacterium tuberculosis. *Sci Rep* **5**, 12567 (2015).

92. J. Gonzalo-Asensio *et al.*, New insights into the transposition mechanisms of IS6110 and its dynamic distribution between Mycobacterium tuberculosis Complex lineages. *PLoS Genet* **14**, e1007282 (2018).

93. P. Bettencourt *et al.*, Identification of antigens presented by MHC for vaccines against tuberculosis. *NPJ Vaccines* **5**, 2 (2020).

94. N. A. Kruh, J. Troudt, A. Izzo, J. Prenni, K. M. Dobos, Portrait of a pathogen: the Mycobacterium tuberculosis proteome in vivo. *PLoS One* **5**, e13938 (2010).

95. S. L. Sampson, Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clin Dev Immunol* **2011**, 497203 (2011).

96. M. Ocampo, H. Curtidor, M. Vanegas, M. A. Patarroyo, M. E. Patarroyo, Specific interaction between Mycobacterium tuberculosis lipoprotein-derived peptides and target cells inhibits mycobacterial entry in vitro. *Chem Biol Drug Des* **84**, 626-641 (2014).

97. G. R. Stewart *et al.*, Effect of deletion or overexpression of the 19-kilodalton lipoprotein Rv3763 on the innate response to Mycobacterium tuberculosis. *Infect Immun* **73**, 6831-6837 (2005).

98. E. H. Noss *et al.*, Toll-like receptor 2-dependent inhibition of macrophage class II MHC expression and antigen processing by 19-kDa lipoprotein of Mycobacterium tuberculosis. *J Immunol* **167**, 910-918 (2001).

99. A. J. Gehring *et al.*, The Mycobacterium tuberculosis 19-kilodalton lipoprotein inhibits gamma interferon-regulated HLA-DR and Fc gamma R1 on human macrophages through Toll-like receptor 2. *Infect Immun* **71**, 4487-4497 (2003).

100. A. M. Cooper, A. Solache, S. A. Khader, Interleukin-12 and tuberculosis: an old story revisited. *Curr Opin Immunol* **19**, 441-447 (2007).

101. Y. Jiang *et al.*, Polymorphisms in the PE35 and PPE68 antigens in Mycobacterium tuberculosis strains may affect strain virulence and reflect ongoing immune evasion. *Mol Med Rep* **13**, 947-954 (2016).

102. K. B. Arnvig *et al.*, Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of Mycobacterium tuberculosis. *PLoS Pathog* **7**, e1002342 (2011).

103. D. V. Ignatov *et al.*, Dormant non-culturable Mycobacterium tuberculosis retains stable low-abundant mRNA. *BMC Genomics* **16**, 954 (2015).

104. G. Shepelkova *et al.*, Analysis of the lung transcriptome in Mycobacterium tuberculosis-infected mice reveals major differences in immune response pathways between TB-susceptible and resistant hosts. *Tuberculosis (Edinb)* **93**, 263-269 (2013).

105. S. Shafiani, G. Tucker-Heard, A. Kariyone, K. Takatsu, K. B. Urdahl, Pathogen-specific regulatory T cells delay the arrival of effector T cells in the lung during early tuberculosis. *J Exp Med* **207**, 1409-1420 (2010).

106. T. Ulrichs, S. H. Kaufmann, New insights into the function of granulomas in human tuberculosis. *J Pathol* **208**, 261-269 (2006).

107. R. Pasula *et al.*, Surfactant protein A (SP-A) mediates attachment of Mycobacterium tuberculosis to murine alveolar macrophages. *Am J Respir Cell Mol Biol* **17**, 209-217 (1997).

108. C. D. Gaynor, F. X. McCormack, D. R. Voelker, S. E. McGowan, L. S. Schlesinger, Pulmonary surfactant protein A mediates enhanced phagocytosis of Mycobacterium tuberculosis by a direct interaction with human macrophages. *J Immunol* **155**, 5343-5351 (1995).

109. V. L. Shepherd, J. P. Lopez, The role of surfactant-associated protein A in pulmonary host defense. *Immunol Res* **23**, 111-120 (2001).

110. T. Malardo *et al.*, MicroRNA expression signatures in lungs of mice infected with Mycobacterium tuberculosis. *Tuberculosis (Edinb)* **101**, 151-159 (2016).

111. D. S. Ghorpade, R. Leyland, M. Kurowska-Stolarska, S. A. Patil, K. N. Balaji, MicroRNA-155 is required for Mycobacterium bovis BCG-mediated apoptosis of macrophages. *Mol Cell Biol* **32**, 2239-2253 (2012).

112. M. V. Rajaram *et al.*, Mycobacterium tuberculosis lipomannan blocks TNF biosynthesis by regulating macrophage MAPK-activated protein kinase 2 (MK2) and microRNA miR-125b. *Proc Natl Acad Sci U S A* **108**, 17408-17413 (2011).

113. R. Kumar *et al.*, Identification of a novel role of ESAT-6-dependent miR-155 induction during infection of macrophages with Mycobacterium tuberculosis. *Cell Microbiol* **14**, 1620-1631 (2012).

114. J. Wang *et al.*, MicroRNA-155 promotes autophagy to eliminate intracellular mycobacteria by targeting Rheb. *PLoS Pathog* **9**, e1003697 (2013).

115. H. Iwai *et al.*, MicroRNA-155 knockout mice are susceptible to Mycobacterium tuberculosis infection. *Tuberculosis (Edinb)* **95**, 246-250 (2015).

116. E. Tortoli *et al.*, The new phylogeny of the genus Mycobacterium: The old and the news. *Infect Genet Evol* **56**, 19-25 (2017).

117. S. Malhotra, S. C. Vedithi, T. L. Blundell, Decoding the similarities and differences among mycobacterial species. *PLoS Negl Trop Dis* **11**, e0005883 (2017).

118. N. L. Bachmann *et al.*, Key Transitions in the Evolution of Rapid and Slow Growing Mycobacteria Identified by Comparative Genomics. *Front Microbiol* **10**, 3019 (2019).

119. F. Ripoll *et al.*, Non mycobacterial virulence genes in the genome of the emerging pathogen Mycobacterium abscessus. *PLoS One* **4**, e5660 (2009).

120. K. Ryan, T. F. Byrd, Mycobacterium abscessus: Shapeshifter of the Mycobacterial World. *Front Microbiol* **9**, 2642 (2018).

121. V. P. Waman *et al.*, Mycobacterial genomics and structural bioinformatics: opportunities and challenges in drug discovery. *Emerg Microbes Infect* **8**, 109-118 (2019).

122. L. Majlessi, R. Prados-Rosales, A. Casadevall, R. Brosch, Release of mycobacterial antigens. *Immunol Rev* **264**, 25-45 (2015).

123. E. Dumas *et al.*, Mycobacterial Pan-Genome Analysis Suggests Important Role of Plasmids in the Radiation of Type VII Secretion Systems. *Genome Biol Evol* **8**, 387-402 (2016).

124. S. T. Howard *et al.*, Spontaneous reversion of Mycobacterium abscessus from a smooth to a rough morphotype is associated with reduced expression of glycopeptidolipid and reacquisition of an invasive phenotype. *Microbiology (Reading)* **152**, 1581-1590 (2006).

125. A. V. Gutierrez, A. Viljoen, E. Ghigo, J. L. Herrmann, L. Kremer, Glycopeptidolipids, a Double-Edged Sword of the Mycobacterium abscessus Complex. *Front Microbiol* **9**, 1145 (2018).

126. E. Catherinot *et al.*, Hypervirulence of a rough variant of the Mycobacterium abscessus type strain. *Infection and Immunity* **75**, 1055-1058 (2007).

127. T. F. Byrd, C. R. Lyons, Preliminary characterization of a Mycobacterium abscessus mutant in human and murine models of infection. *Infection and Immunity* **67**, 4700-4707 (1999).

X. Anexos

Tabla Anexa 1. Lecturas mapeadas a los genomas de ratón y *M. tuberculosis* en las tres estrategias.

		Lecturas crudas	Lecturas limpias	Lecturas totales mapeadas al genoma de ratón	Lecturas totales mapeadas al genoma de Mtb	% lecturas mapeadas al genoma de Mtb con respecto a las lecturas limpias	PROMEDIO	Lecturas ribosomales de Mtb	% lecturas mapeadas transcritos ribosomales de Mtb con respecto a las lecturas mapeadas al genoma de Mtb	PROMEDIO
ESTRATEGIA 1	Extracción de ARN directa (estrategia tradicional)	7,672,944	7,585,764	4,599,634	128,784	1.70	1.70	124,446	96.63	96.63
ESTRATEGIA 2	Extracción de ARN con lisis celular diferencial	14,339,158	14,222,581	6,696,869	858,259	6.03	5.83	835,125	97.30	97.07
		17,385,954	17,281,608	7,894,465	1,059,150	6.13		1,026,435	96.91	
		11,610,750	11,525,470	5,556,047	612,649	5.32		594,224	96.99	
ESTRATEGIA 3	Lisis celular diferencial + Sondas biotiniladas	63,536,942	62,791,012	55,673,290	95,676	0.15	0.13	58,099	60.72	72.51
		40,357,826	39,855,911	34,996,425	59,092	0.15		55,234	93.47	
		46,740,872	46,156,549	41,138,380	34,725	0.08		21,996	63.34	

Tabla Anexa 2. Dominios y familias de InterPro significativamente sobrerrepresentadas en los 529 genes expresados de Mtb

InterPro ID	Categoría	Descripción	P-Value
IPR014043	domain	Acyl transferase	6.267E-07
IPR009081	domain	Phosphopantetheine binding ACP domain	1.438E-06
IPR014030	domain	Beta-ketoacyl synthase, N-terminal	2.747E-05
IPR014031	domain	Beta-ketoacyl synthase, C-terminal	4.885E-05
IPR013968	domain	Polyketide synthase, ketoreductase domain	1.624E-04
IPR023836	family	EccCa-like, Actinobacteria	1.087E-03
IPR023837	family	EccCb-like, Actinobacteria	1.087E-03
IPR032821	domain	Ketoacyl-synthetase, C-terminal extension	1.190E-03
IPR006091	domain	Acyl-CoA oxidase/dehydrogenase, central domain	1.363E-03
IPR020807	domain	Polyketide synthase, dehydratase domain	1.757E-03
IPR009075	domain	Acyl-CoA dehydrogenase/oxidase C-terminal	1.995E-03
IPR002543	domain	FtsK domain	6.180E-03
IPR003029	domain	S1 domain	7.039E-03
IPR003495	domain	CobW/HypB/UreG, nucleotide-binding domain	7.039E-03
IPR023753	domain	FAD/NAD(P	1.360E-02
IPR000788	domain	Ribonucleotide reductase large subunit, C-terminal	1.558E-02
IPR001030	domain	Aconitase/3-isopropylmalate dehydratase large subunit, alpha/beta/alpha domain	1.558E-02
IPR002300	domain	Aminoacyl-tRNA synthetase, class Ia	1.558E-02
IPR003714	domain	PhoH-like protein	1.558E-02
IPR004100	domain	ATPase, F1/V1/A1 complex, alpha/beta subunit, N-terminal domain	1.558E-02
IPR011115	domain	SecA DEAD-like, N-terminal	1.558E-02
IPR011116	domain	SecA Wing/Scaffold	1.558E-02
IPR011130	domain	SecA, preprotein cross-linking domain	1.558E-02
IPR013509	domain	Ribonucleotide reductase large subunit, N-terminal	1.558E-02
IPR014018	domain	SecA motor DEAD	1.558E-02
IPR023234	domain	NarG-like domain	1.558E-02
IPR025878	domain	Acetyl-CoA dehydrogenase-like C-terminal domain	1.558E-02
IPR000185	family	Protein translocase subunit SecA	1.558E-02
IPR001844	family	Chaperonin Cpn60	1.558E-02
IPR002423	family	Chaperonin Cpn60/TCP-1 family	1.558E-02
IPR003816	family	Nitrate reductase, gamma subunit	1.558E-02
IPR004392	family	Hydrogenase maturation factor HypB	1.558E-02
IPR005372	family	Uncharacterised protein family UPF0182	1.558E-02
IPR013538	family	Activator of Hsp90 ATPase homologue 1-like	1.558E-02
IPR013786	domain	Acyl-CoA dehydrogenase/oxidase, N-terminal	1.792E-02
IPR001584	domain	Integrase, catalytic core	1.811E-02
IPR013154	domain	Alcohol dehydrogenase, N-terminal	1.857E-02
IPR006656	domain	Molybdopterin oxidoreductase	2.900E-02
IPR003959	domain	ATPase, AAA-type, core	3.870E-02
IPR000194	domain	ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding domain	4.287E-02
IPR004675	domain	Alkylhydroperoxidase AhpD core	4.287E-02
IPR004843	domain	Calcineurin-like phosphoesterase domain, ApaH type	4.287E-02
IPR013155	domain	Methionyl/Valyl/Leucyl/Isoleucyl-tRNA synthetase, anticodon-binding	4.287E-02
IPR025948	domain	HTH-like domain	4.287E-02
IPR000537	family	UbiA prenyltransferase family	4.287E-02
IPR001714	family	Peptidase M24, methionine aminopeptidase	4.287E-02
IPR006254	family	Isocitrate lyase	4.287E-02
IPR013126	family	Heat shock protein 70 family	4.287E-02
IPR024520	family	Protein of unknown function DUF3558	4.287E-02
IPR038965	family	Transposase InsF-like	4.287E-02
IPR004707	family	Membrane transport protein MmpL family	4.611E-02

Tabla Anexa 3. Términos GO significativamente enriquecidos por categoría asociados a los 529 genes más expresados de Mtb durante la infección in vivo.

GO ID	Nombre GO	P-Value
A. Biological Process		
GO:0031295	T cell costimulation	1.765E-05
GO:0031294	lymphocyte costimulation	1.765E-05
GO:0009081	branched-chain amino acid metabolic process	4.007E-05
GO:0009083	branched-chain amino acid catabolic process	1.102E-04
GO:0006549	isoleucine metabolic process	1.197E-04
GO:0000288	nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay	2.663E-04
GO:0006573	valine metabolic process	3.513E-04
GO:0060184	cell cycle switching	4.878E-04
GO:0051728	cell cycle switching, mitotic to meiotic cell cycle	4.878E-04
GO:0006574	valine catabolic process	5.108E-04
GO:0006552	leucine catabolic process	5.108E-04
GO:0006550	isoleucine catabolic process	5.108E-04
GO:0006551	leucine metabolic process	6.270E-04
GO:0006939	smooth muscle contraction	6.963E-04
GO:0140013	meiotic nuclear division	7.369E-04
GO:1901565	organonitrogen compound catabolic process	1.030E-03
GO:0006631	fatty acid metabolic process	1.040E-03
GO:0031340	positive regulation of vesicle fusion	1.211E-03
GO:0031338	regulation of vesicle fusion	1.211E-03
GO:0044281	small molecule metabolic process	1.287E-03
B. Cellular Component		
GO:0005643	nuclear pore	4.20E-03
GO:0000932	P-body	5.39E-03
GO:0009986	cell surface	6.00E-03
GO:0098978	glutamatergic synapse	6.18E-03
GO:1990527	Tec1p-Ste12p-Dig1p complex	7.04E-03
GO:0110165	cellular anatomical entity	1.23E-02
GO:1990526	Ste12p-Dig1p-Dig2p complex	1.60E-02
GO:0030496	midbody	1.81E-02
GO:0009325	nitrate reductase complex	1.83E-02
GO:0016020	membrane	2.56E-02
GO:0031300	intrinsic component of organelle membrane	2.57E-02
GO:0031301	integral component of organelle membrane	2.57E-02
GO:0098982	GABA-ergic synapse	2.90E-02
GO:0062071	Pi Mi complex	2.90E-02
GO:0005863	striated muscle myosin thick filament	2.90E-02
GO:0030015	CCR4-NOT core complex	3.08E-02
GO:0030014	CCR4-NOT complex	3.08E-02
GO:0089713	Cbf1-Met4-Met28 complex	3.87E-02
GO:0042101	T cell receptor complex	4.29E-02
GO:0042105	alpha-beta T cell receptor complex	4.29E-02
C. Molecular Function		
GO:0004085	butyryl-CoA dehydrogenase activity	1.62E-04
GO:0005488	binding	6.96E-04
GO:0052890	oxidoreductase activity, acting on the CH-CH group of donors, with a flavin as acceptor	7.83E-04
GO:0043168	anion binding	1.16E-03
GO:0003955	NAD(P)H dehydrogenase (quinone) activity	1.94E-03
GO:0017056	structural constituent of nuclear pore	2.87E-03
GO:0004962	endothelin receptor activity	2.94E-03
GO:0036094	small molecule binding	3.15E-03
GO:1901363	heterocyclic compound binding	4.09E-03
GO:0097159	organic cyclic compound binding	4.10E-03
GO:0035375	zymogen binding	6.18E-03
GO:0097367	carbohydrate derivative binding	6.87E-03
GO:0030527	structural constituent of chromatin	7.04E-03
GO:0004590	orotidine-5'-phosphate decarboxylase activity	7.09E-03
GO:0005044	scavenger receptor activity	7.23E-03
GO:0038024	cargo receptor activity	7.23E-03
GO:1901265	nucleoside phosphate binding	7.82E-03
GO:0000166	nucleotide binding	7.82E-03
GO:0004032	alditol:NADP+ 1-oxidoreductase activity	1.11E-02
GO:0043167	ion binding	1.20E-02

Tabla Anexa 4. Vías KEGG más representadas en los 529 genes expresados de Mtb.

KEGG ID	Vía	%
03020	RNA polymerase	75.00
00630	Glyoxylate and dicarboxylate metabolism	34.15
03018	RNA degradation	33.33
00562	Inositol phosphate metabolism	33.33
00910	Nitrogen metabolism	31.82
05152	Tuberculosis	30.77
00680	Methane metabolism	28.57
00290	Valine, leucine and isoleucine biosynthesis	28.57
00430	Taurine and hypotaurine metabolism	28.57
00020	Citrate cycle (TCA cycle)	28.13
00730	Thiamine metabolism	27.27
00983	Drug metabolism - other enzymes	27.27
00061	Fatty acid biosynthesis	26.67
02024	Quorum sensing	25.00
00010	Glycolysis / Gluconeogenesis	24.24
00270	Cysteine and methionine metabolism	24.24
00260	Glycine, serine and threonine metabolism	24.00
03060	Protein export	23.53
01053	Biosynthesis of siderophore group nonribosomal peptides	22.22
03070	Bacterial secretion system	21.43
00572	Arabinogalactan biosynthesis - Mycobacterium	21.43
00760	Nicotinate and nicotinamide metabolism	21.05
00620	Pyruvate metabolism	20.83
02020	Two-component system	20.34
00660	C5-Branched dibasic acid metabolism	20.00
00350	Tyrosine metabolism	20.00
00190	Oxidative phosphorylation	18.75
00640	Propanoate metabolism	18.37
00072	Synthesis and degradation of ketone bodies	18.18
00521	Streptomycin biosynthesis	18.18
00625	Chloroalkane and chloroalkene degradation	18.18
00561	Glycerolipid metabolism	17.24
00860	Porphyrin and chlorophyll metabolism	17.14
00500	Starch and sucrose metabolism	16.67
00780	Biotin metabolism	16.67
00280	Valine, leucine and isoleucine degradation	15.79
00520	Amino sugar and nucleotide sugar metabolism	15.38
00240	Pyrimidine metabolism	15.38
00220	Arginine biosynthesis	15.00
00480	Glutathione metabolism	14.29
00650	Butanoate metabolism	14.04
03010	Ribosome	13.11
03410	Base excision repair	12.50
00071	Fatty acid degradation	12.00
00900	Terpenoid backbone biosynthesis	11.54
00230	Purine metabolism	11.11
00790	Folate biosynthesis	11.11
00770	Pantothenate and CoA biosynthesis	11.11
02010	ABC transporters	10.81
00360	Phenylalanine metabolism	10.81
00362	Benzoate degradation	10.00
00340	Histidine metabolism	10.00
00720	Carbon fixation pathways in prokaryotes	9.71
00920	Sulfur metabolism	9.52
00330	Arginine and proline metabolism	9.52
00310	Lysine degradation	9.30
03440	Homologous recombination	9.09
00380	Tryptophan metabolism	8.89
00970	Aminoacyl-tRNA biosynthesis	7.25
00903	Limonene and pinene degradation	7.14
00564	Glycerophospholipid metabolism	6.90
00410	beta-Alanine metabolism	6.45
04112	Cell cycle - Caulobacter	6.45
04212	Longevity regulating pathway - worm	5.36
00195	Photosynthesis	4.76
04931	Insulin resistance	4.05
04217	Necroptosis	2.08

Tabla Anexa 5. Genes expresados que pertenecen al secretoma de Mtb.

Gen ID	Descripción del gen
Rv3344c	PE_PGRS49 PE-PGR
Rv3512	PE_PGRS56 PE-PGRS
Rv0746	PE_PGRS9 PE-PGRS
Rv3345c	PE_PGRS50 PE-PGR
Rv0105c	uracil-DNA glycosylase
Rv2424c	Uncharacterised protein
Rv3053c	Chain A, Glutaredoxin Like Protein Nrdh
Rv3873	PPE68 PPE FAMILY
Rv0351	nucleotide exchange factor GrpE
Rv2270	lipoprotein lppN
Rv2784c	lipoprotein LppU
Rv0009	peptidyl-prolyl cis-trans isomerase
Rv0287	type VII secretion protein EsxS
Rv2442c	50S ribosomal protein L21
Rv3791	decaprenylphosphoryl-D-2-keto erythropentose reductase
Rv0830	SAM-dependent methyltransferase
Rv0281	SAM-dependent methyltransferase
Rv3035	PQQ enzyme repeat-containing protein
Rv0502	phospholipid/glycerol acyltransferase
Rv1087	PE_PGRS21 PE-PGRS
Rv2659c	integrase
Rv2518c	lipoprotein LppS
Rv3881c	type VII secretion system ESX-1 target EspB
Rv3685c	cytochrome P450
Rv0119	acyl-CoA synthetase
Rv1243c	PE_PGRS23 PE-PGRS
Rv2853	PE_PGRS48 PE-PGRS
Rv1266c	putative transmembrane serine/threonine-protein kinase H pknH
Rv0578c	PE_PGRS7 PE-PGRS
Rv0931c	serine/threonine protein kinase
Rv0932c	phosphate-binding protein
Rv3909	Conserved protein
Rv1664	polyketide synthase
Rv0598c	PIN domain-containing protein
Rv1917c	PPE34 PPE FAMILY
Rv3547	deazaflavin-dependent nitroreductase
Rv3418c	molecular chaperone GroES
Rv3034c	acetyltransferase
Rv1174c	hemophore-related protein
Rv0164	cyclase
Rv1274	lipoprotein lprB
Rv3350c	PPE56 PPE FAMILY
Rv2861c	type I methionyl aminopeptidase
Rv2544	lipoprotein lppB
Rv3344c	PE_PGRS49 PE-PGR
Rv3036c	DUF3298 domain-containing protein
Rv1714	oxidoreductase
Rv2585c	protein translocase subunit SecF

Rv1468c	PE_PGRS29 PE-PGR
Rv3878	secretion protein EspJ
Rv3502c	3-oxoacyl-ACP reductase
Rv1172c	PE12 PE FAMILY P
Rv1076	lipase lipU
Rv1768	PE_PGRS31 PE-PGRS
Rv0280	PPE3 PPE FAMILY P
Rv1971	MCE family protein
Rv1493	methylmalonyl-CoA mutase
Rv1702c	HNH endonuclease
Rv3106	NADPH:adrenodoxin oxidoreductase fprA (NADPH-ferredoxin reductase)
Rv2791c	transposase
Rv2577	purple acid phosphatase-related protein
Rv3763	lipoprotein LpqH
Rv2930	fatty-acid-CoA ligase fadD26
Rv2970c	lipase/esterase LIPN
Rv1275	DUF3558 domain-containing protein
Rv2220	glutamine synthetase 1
Rv3590c	PE_PGRS58 PE-PGR
Rv1984c	cutinase cfp21
Rv3202c	ATP-dependent DNA helicase
Rv0305c	PPE6 PPE FAMILY
Rv0278c	PE_PGRS3 PE-PGRS
Rv2587c	protein translocase subunit SecD
Rv0833	PE_PGRS13 PE-PGRS
Rv2905	lipoprotein LppW
Rv3634c	NAD-dependent epimerase/dehydratase family protein
Rv0755c	PPE12 PPE FAMILY
Rv3136	PPE51 PPE FAMILY
Rv0399c	D-alanyl-D-alanine carboxypeptidase
Rv2280	FAD-binding oxidoreductase
Rv2351c	Phospholipase C
Rv2264c	Conserved protein of uncharacterised function
Rv3159c	PPE53 PPE FAMILY
Rv3667	acetyl-coenzyme A synthetase
Rv0355c	PPE8 PPE FAMILY
Rv0101	non-ribosomal peptide synthetase

Tabla Anexa 6. 101 genes de M. tuberculosis expresados al día 21 post-infección con al menos 1 read en dos muestras

Ranking de expresión	Nombre del gen	Gen ID	Descripción del gen	RPKM promedio
1	Scaffold_1_orf04709	Rv2512c	Transposase for insertion sequence element IS1081	6.84E+05
2	Scaffold_1_orf04711	Rv1067c	PE_PGRS19 PE-PGR	3.45E+05
3	Scaffold_1_orf00755	Rv3512	PE_PGRS56 PE-PGRS	1.91E+05
4	Scaffold_1_orf04705	Rv1067c	PE_PGRS19 PE-PGR	1.65E+05
5	Scaffold_1_orf01047	Rv3344c	PE_PGRS49 PE-PGR	1.63E+05
6	Scaffold_1_orf05988	Rv0279c	PE_PGRS4 PE-PGRS	1.10E+05
7	Scaffold_1_orf00760	Rv3512	PE_PGRS56 PE-PGRS	5.01E+04
8	Scaffold_1_orf05233	Rv0746	PE_PGRS9 PE-PGRS	4.35E+04
9	Scaffold_1_orf01045	Rv3345c	PE_PGRS50 PE-PGR	1.94E+04
10	Scaffold_1_orf05667	Rv0440	molecular chaperone GroEL	5.71E+03
11	Scaffold_1_orf04766	Rv3620c	peptidase M22	4.88E+03
12	Scaffold_1_orf04567	Rv2424c	Uncharacterised protein	3.11E+03
13	Scaffold_1_orf00204	Rv3873	PPE68 PPE FAMILY	2.02E+03
14	Scaffold_1_orf04443	Rv1221	RNA polymerase sigma factor SigE	1.97E+03
15	Scaffold_1_orf06032	Rv0251c	Hsp20/alpha crystallin family protein	1.63E+03
16	Scaffold_1_orf03409	Rv1876	bacterioferritin	1.61E+03
17	Scaffold_1_orf00594	Rv3615c	type VII secretion system ESX-1 filament-forming target EspC	1.57E+03
18	Scaffold_1_orf05889	Rv0287	type VII secretion protein EsxS	1.37E+03
19	Scaffold_1_orf02479	Rv2442c	50S ribosomal protein L21	1.30E+03
20	Scaffold_1_orf05797	Rv0350	fusion protein	1.04E+03
21	Scaffold_1_orf06109	Rv0199	Conserved membrane protein of uncharacterised function	9.22E+02
22	Scaffold_1_orf06090	Rv0211	Conserved membrane protein of uncharacterised function	8.43E+02
23	Scaffold_1_orf01539	Rv3035	PQQ enzyme repeat-containing protein	8.15E+02
24	Scaffold_1_orf00474	Rv3695	RDD family protein	8.09E+02
25	Scaffold_1_orf05484	Rv0600c	sensor histidine kinase	7.96E+02
26	Scaffold_1_orf00592	Rv3616c	type VII secretion system ESX-1 target EspA	7.63E+02
27	Scaffold_1_orf05059	Rv0867c	molybdenum cofactor biosynthesis protein E2	7.30E+02
28	Scaffold_1_orf02065	Rv2712c	Uncharacterised protein	6.91E+02
29	Scaffold_1_orf05622	Rv0467	isocitrate lyase	6.79E+02
30	Scaffold_1_orf05218	Rv0752c	acyl-CoA dehydrogenase FadE9	6.66E+02
31	Scaffold_1_orf02394	Rv2495c	branched-chain alpha-ketoacid dehydrogenase complex dihydrolipoyllysine-residue acyltransferase	6.38E+02
32	Scaffold_1_orf04588	Rv1131	Chain A, Crystal Structure Of Methylcitrate Synthase From Mycobacterium Tuberculosis	6.38E+02
33	Scaffold_1_orf03416	Rv1872c	L-lactate dehydrogenase	6.38E+02
34	Scaffold_1_orf05586	Rv0490	two-component sensor histidine kinase	6.25E+02
35	Scaffold_1_orf01140	Rv3290c	L-lysine 6-transaminase	5.79E+02
36	Scaffold_1_orf05886	Rv0290	type VII secretion system ESX-3 subunit EccD3	5.79E+02
37	Scaffold_1_orf00190	Rv3881c	type VII secretion system ESX-1 target EspB	5.65E+02
38	Scaffold_1_orf04156	Rv1614	prolipoprotein diacylglyceryl transferase lgt	5.40E+02
39	Scaffold_1_orf04589	Rv1130	MmgE/PrpD family protein	5.23E+02
40	Scaffold_1_orf03531	Rv1795	type VII secretion system ESX-5 subunit EccD5	5.17E+02
41	Scaffold_1_orf02930	Rv2157c	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	5.10E+02
42	Scaffold_1_orf05620	Rv0468	3-hydroxybutyryl-CoA dehydrogenase	5.00E+02
43	Scaffold_1_orf00912	Rv3417c	chaperonin GroEL	4.82E+02
44	Scaffold_1_orf04356	Rv1280c	periplasmic oligopeptide-binding lipoprotein oppA	4.77E+02
45	Scaffold_1_orf02453	Rv2459	integral membrane transport protein	4.72E+02

46	Scaffold_1_orf03145	Rv2028c	universal stress protein	4.65E+02
47	Scaffold_1_orf01529	Rv3043c	cytochrome ubiquinol oxidase subunit I	4.54E+02
48	Scaffold_1_orf04410	Rv1243c	PE_PGRS23 PE-PGRS	4.43E+02
49	Scaffold_1_orf01868	Rv2853	PE_PGRS48 PE-PGRS	4.23E+02
50	Scaffold_1_orf02998	Rv2115c	ATPase AAA	4.23E+02
51	Scaffold_1_orf05898	Rv0282	type VII secretion system ESX-3 AAA family ATPase EccA3	4.02E+02
52	Scaffold_1_orf04959	Rv0932c	phosphate-binding protein	3.69E+02
53	Scaffold_1_orf00497	Rv3679	anion transporter ATPase	3.69E+02
54	Scaffold_1_orf00711	Rv3534c	4-hydroxy-2-oxovalerate aldolase	3.69E+02
55	Scaffold_1_orf03356	Rv1908c	catalase/peroxidase HPI	3.68E+02
56	Scaffold_1_orf03142	Rv2030c	erythromycin esterase	3.66E+02
57	Scaffold_1_orf03880	Rv1398c	antitoxin	3.59E+02
58	Scaffold_1_orf00617	Rv3595c	PE_PGRS59 PE-PGR	2.96E+02
59	Scaffold_1_orf04504	Rv1183	RND transporter MmpL10	2.80E+02
60	Scaffold_1_orf01496	Rv3060c	transcriptional regulator, GntR family	2.79E+02
61	Scaffold_1_orf01296	Rv3193c	Conserved membrane protein of uncharacterised function	2.64E+02
62	Scaffold_1_orf03495	Rv1818c	PE_PGRS33 PE-PGR	2.61E+02
63	Scaffold_1_orf06290	Rv0086	hydrogenase HycQ	2.57E+02
64	Scaffold_1_orf05891	Rv0286	PPE4 PPE FAMILY P	2.53E+02
65	Scaffold_1_orf04074	Rv1664	polyketide synthase	2.49E+02
66	Scaffold_1_orf01675	Rv2950c	long-chain-fatty-acid--AMP ligase FadD29	2.10E+02
67	Scaffold_1_orf03550	Rv1784	FtsK/SpoIIIE family protein	1.97E+02
68	Scaffold_1_orf03344	Rv1917c	PPE34 PPE FAMILY	1.91E+02
69	Scaffold_1_orf01701	Rv2935	phthiocerol type I polyketide synthase PpsE	1.84E+02
70	Scaffold_1_orf04714	Rv1065	cysteine dioxygenase	1.77E+02
71	Scaffold_1_orf00910	Rv3418c	molecular chaperone GroES	1.74E+02
72	Scaffold_1_orf01540	Rv3034c	acetyltransferase	1.68E+02
73	Scaffold_1_orf02354	Rv2524c	DUF1729 domain-containing protein	1.65E+02
74	Scaffold_1_orf04516	Rv1174c	hemophore-related protein	1.58E+02
75	Scaffold_1_orf02401	Rv2490c	Uncharacterised protein	1.55E+02
76	Scaffold_1_orf01860	Rv3728	Truncated hydrogenase nickle incorporation protein	1.52E+02
77	Scaffold_1_orf03687	Rv1527c	polyketide synthase	1.20E+02
78	Scaffold_1_orf04183	Rv1594	quinolinate synthetase	1.03E+02
79	Scaffold_1_orf05167	Rv0784	putative deacetylase	7.68E+01
80	Scaffold_1_orf01026	Rv3350c	PPE56 PPE FAMILY	7.56E+01
81	Scaffold_1_orf06164	Rv0163	acyl-CoA thioesterase	7.20E+01
82	Scaffold_1_orf04791	Rv0818	DNA-binding response regulator	6.87E+01
83	Scaffold_1_orf04695	Rv1072	transmembrane protein	6.30E+01
84	Scaffold_1_orf03141	Rv2031c	alpha-crystallin	6.07E+01
85	Scaffold_1_orf03772	Rv1468c	PE_PGRS29 PE-PGR	4.74E+01
86	Scaffold_1_orf02898	Rv2178c	3-deoxy-7-phosphoheptulonate synthase class II	3.80E+01
87	Scaffold_1_orf02923	Rv2161c	F420-dependent oxidoreductase	3.79E+01
88	Scaffold_1_orf04654	Rv1945	HNH endonuclease	3.55E+01
89	Scaffold_1_orf05903	Rv0280	PPE3 PPE FAMILY P	3.05E+01
90	Scaffold_1_orf03737	Rv1493	methylmalonyl-CoA mutase	2.94E+01
91	Scaffold_1_orf04837	Rv1011	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase ispE	2.76E+01
92	Scaffold_1_orf01710	Rv2930	fatty-acid-CoA ligase fadD26	2.46E+01
93	Scaffold_1_orf04489	Rv1196	PPE18 PPE FAMILY	2.30E+01
94	Scaffold_1_orf02842	Rv2220	glutamine synthetase 1	2.29E+01
95	Scaffold_1_orf02590	Rv2379c	non-ribosomal peptide synthetase	2.11E+01
96	Scaffold_1_orf01280	Rv3202c	ATP-dependent DNA helicase	1.87E+01
97	Scaffold_1_orf05865	Rv0305c	PPE6 PPE FAMILY	1.82E+01
98	Scaffold_1_orf00317	Rv3800c	polyketide synthase	1.66E+01
99	Scaffold_1_orf05120	Rv0833	PE_PGRS13 PE-PGRS	1.46E+01
100	Scaffold_1_orf03111	Rv2048c	beta-ketoacyl synthase	8.31E+00
101	Scaffold_1_orf05790	Rv0355c	PPE8 PPE FAMILY	3.32E+00

Tabla Anexa 7. Términos GO significativamente enriquecidos por categoría asociados al 5% de los genes más expresados de ratón durante la infección con Mtb.

GO ID	Nombre GO	P-Value
A. Biological Process		
GO:0043903	regulation of symbiosis, encompassing mutualism through parasitism	3.23E-06
GO:0042981	regulation of apoptotic process	5.37E-06
GO:0002683	negative regulation of immune system process	1.05E-05
GO:0002478	antigen processing and presentation of exogenous peptide antigen	1.68E-04
GO:0043065	positive regulation of apoptotic process	2.71E-04
GO:0002831	regulation of response to biotic stimulus	3.67E-04
GO:0045597	positive regulation of cell differentiation	5.05E-04
GO:0070374	positive regulation of ERK1 and ERK2 cascade	5.33E-04
GO:0071356	cellular response to tumor necrosis factor	6.52E-04
GO:0019730	antimicrobial humoral response	8.82E-04
GO:0006915	apoptotic process	1.27E-03
GO:0071222	cellular response to lipopolysaccharide	2.15E-03
GO:0051851	modulation by host of symbiont process	2.57E-03
GO:1902105	regulation of leukocyte differentiation	3.73E-03
GO:0097237	cellular response to toxic substance	5.09E-03
GO:0030097	hemopoiesis	5.40E-03
GO:0001817	regulation of cytokine production	5.60E-03
GO:0045765	regulation of angiogenesis	6.22E-03
GO:0035458	cellular response to interferon-beta	6.48E-03
GO:0034341	response to interferon-gamma	7.20E-03
B. Cellular Component		
GO:0005615	extracellular space	1.11E-12
GO:0043209	myelin sheath	1.60E-08
GO:0042824	MHC class I peptide loading complex	2.65E-08
GO:0042612	MHC class I protein complex	1.80E-05
GO:0030670	phagocytic vesicle membrane	3.45E-05
GO:0062023	collagen-containing extracellular matrix	5.45E-05
GO:0045121	membrane raft	7.45E-05
GO:0005794	Golgi apparatus	2.92E-03
GO:0005839	proteasome core complex	5.17E-03
GO:0005764	lysosome	5.21E-03
GO:0022626	cytosolic ribosome	5.36E-03
GO:0031966	mitochondrial membrane	6.05E-03
GO:0009897	external side of plasma membrane	1.04E-02
GO:0036464	cytoplasmic ribonucleoprotein granule	1.26E-02
GO:0048471	perinuclear region of cytoplasm	1.41E-02
GO:0120025	plasma membrane bounded cell projection	2.02E-02
GO:0005634	nucleus	3.21E-02
C. Molecular Function		
GO:0044877	protein-containing complex binding	1.08E-09
GO:0042802	identical protein binding	4.48E-08
GO:0005198	structural molecule activity	2.29E-04
GO:0042605	peptide antigen binding	8.60E-04
GO:0030881	beta-2-microglobulin binding	8.66E-04
GO:0003924	GTPase activity	9.41E-04
GO:0008009	chemokine activity	1.33E-03
GO:0030234	enzyme regulator activity	1.69E-03
GO:0045236	CXCR chemokine receptor binding	2.98E-03
GO:0042608	T cell receptor binding	3.88E-03
GO:0004888	transmembrane signaling receptor activity	5.05E-03
GO:0016209	antioxidant activity	6.95E-03
GO:0004298	threonine-type endopeptidase activity	1.21E-02
GO:0050998	nitric-oxide synthase binding	1.47E-02
GO:0019843	rRNA binding	2.58E-02
GO:0042610	CD8 receptor binding	2.88E-02
GO:0005525	GTP binding	4.02E-02
GO:0046979	TAP2 binding	4.73E-02
GO:0046978	TAP1 binding	4.73E-02

Tabla Anexa 8. Vías KEGG más representadas en 5% de los genes más expresados de ratón durante la infección con Mtb.

KEGG ID	Vía	% de la vía representado por los genes expresados de ratón
03010	Ribosome	37.71
03050	Proteasome	27.66
00190	Oxidative phosphorylation	26.32
03060	Protein export	17.86
04612	Antigen processing and presentation	17.78
04714	Thermogenesis	14.71
04145	Phagosome	14.36
04142	Lysosome	12.21
04216	Ferroptosis	12.20
04657	IL-17 signaling pathway	12.09
04668	TNF signaling pathway	11.50
05152	Tuberculosis	10.44
04620	Toll-like receptor signaling pathway	10.10
04062	Chemokine signaling pathway	9.69
04380	Osteoclast differentiation	9.56
04670	Leukocyte transendothelial migration	9.48
04623	Cytosolic DNA-sensing pathway	9.38
00010	Glycolysis / Gluconeogenesis	9.09
04625	C-type lectin receptor signaling pathway	8.93
04210	Apoptosis	8.82

biology

MDPI

Article

Targeted RNA-Seq Reveals the *M. tuberculosis* Transcriptome from an In Vivo Infection Model

Fernanda Cornejo-Granados [1], Gamaliel López-Leal [1], Dulce A. Mata-Espinosa [2], Jorge Barrios-Payán [2], Brenda Marquina-Castillo [2], Edgar Equihua-Medina [1], Zyanya L. Zatarain-Barrón [2], Camilo Molina-Romero [2], Rogelio Hernández-Pando [2,*] and Adrian Ochoa-Leyva [1,*]

[1] Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autonóma de México, Cuernavaca 62210, Mexico; fernanda.cornejo@ibt.unam.mx (F.C.-G.); gamlopez@ccg.unam.mx (G.L.-L.); edgar.equihua@uaem.edu.mx (E.E.-M.)

[2] Sección de Patología Experimental, Departamento de Patología, Instituto Nacional de Ciencias Médicas y Nutrición "Salvador Zubirán", Vasco de Quiroga 15, Tlalpan, Sección XVI, Ciudad de México 14000, Mexico; dulmat@comunidad.unam.mx (D.A.M.-E.); jorge.barriosp@incmnsz.mx (J.B.-P.); brenda.marquina@comunidad.unam.mx (B.M.-C.); zyanyal@comunidad.unam.mx (Z.L.Z.-B.); camilomol@comunidad.unam.mx (C.M.-R.)

* Correspondence: rogelio.hernandezp@incmnsz.mx (R.H.-P.); adrian.ochoa@ibt.unam.mx (A.O.-L.)

check for updates

Citation: Cornejo-Granados, F.; López-Leal, G.; Mata-Espinosa, D.A.; Barrios-Payán, J.; Marquina-Castillo, B.; Equihua-Medina, E.; Zatarain-Barrón, Z.L.; Molina-Romero, C.; Hernández-Pando, R.; Ochoa-Leyva, A. Targeted RNA-Seq Reveals the *M. tuberculosis* Transcriptome from an In Vivo Infection Model. *Biology* **2021**, *10*, 848. https://doi.org/10.3390/biology10090848

Academic Editor: Wito Richter

Received: 29 June 2021
Accepted: 16 August 2021
Published: 31 August 2021

Simple Summary: High-throughput sequencing techniques such as RNA-seq allow a more detailed characterization of the gene expression profile during in vivo infections. However, using this strategy for intracellular pathogens such as *Mycobacterium tuberculosis* (Mtb) entails technical limitations. Some authors have resorted to flow cytometers to separate infected cells or significantly increase sequencing depth to obtain pathogens' gene expression. However, these options carry additional expenses in specialized equipment. We propose an experimental protocol based on differential cell lysis and a probe-based ribosomal depletion to determine the gene expression of Mtb and its host during in vivo infection. This method allowed us to increase the number of observed expressed genes from 13 using a traditional RNA-seq approach to 702. In addition, we observed the expression of genes essential for establishing the infection, codifying proteins such as PE-PGRS, lipoproteins lppN and LpqH, and three ncRNAs (small RNA MTS2823, transfer-messenger RNA RF00023, and ribozyme RF00010). We believe our method represents a valuable alternative to current RNA-seq approaches to study host–pathogen interactions and will help explore host–pathogen mechanisms in tuberculosis and other similar models of intracellular infections.

Abstract: The study of host-pathogen interactions using in vivo models with intracellular pathogens like Mycobacterium tuberculosis (Mtb) entails technical limitations, such as: (i) Selecting an efficient differential lysis system to enrich the pathogen cells; (ii) obtaining sufficient high-quality RNA; and (iii) achieving an efficient rRNA depletion. Thus, some authors had used flow cytometers to separate infected cells or significantly increase the sequencing depth of host–pathogen RNA libraries to observe the pathogens' gene expression. However, these options carry additional expenses in specialized equipment typically not available for all laboratories. Here, we propose an experimental protocol involving differential cell lysis and a probe-based ribosomal depletion to determine the gene expression of Mtb and its host during in vivo infection. This method increased the number of observed pathogen-expressed genes from 13 using the traditional RNA-seq approach to 702. After eliminating rRNA reads, we observed that 61.59% of Mtb sequences represented 702 genes, while 38.41% represented intergenic regions. Some of the most expressed genes codified for IS1081 (Rv2512c) transposase and eight PE-PGRS members, such as PGRS49 and PGRS50. As expected, a critical percent of the expressed genes codified for secreted proteins essential for infection, such as PE68, lppN, and LpqH. Moreover, three Mtb ncRNAs were highly expressed (small RNA MTS2823, transfer-messenger RNA RF00023, and ribozyme RF00010). Many of the host-expressed genes were related to the inflammation process and the expression of surfactant proteins such as the Sftpa and Sftpc, known to bind Mtb to alveolar macrophages and mi638, a microRNA with no previous associations with pulmonary diseases. The main objective of this study is to present the method, and a general catalog of the Mtb expressed genes at one point of the in vivo infection. We believe our method represents a different approach to the existing ones to study host–pathogen interactions in tuberculosis and other similar intracellular infections, without the necessity of specialized equipment.

Keywords: transcriptome; tuberculosis; host-pathogen; RNA-seq; in vivo infection

1. Introduction

RNA-seq approaches have helped to define the finely regulated host–pathogen interactions. However, its application for in vivo models of intracellular bacteria such as *M. tuberculosis* (Mtb) entails several limitations: (i) Selecting an efficient cell lysis system [1]; (ii) obtaining sufficient high-quality RNA [1,2]; and (iii) achieving an efficient rRNA depletion [1,2]. Hence, some studies have used infected cell cultures [3,4], labeled bacteria to separate infected from non-infected cells [1,5], increased the number of infecting bacilli to obtain enough RNA [6], or selectively amplified bacterial RNA/cDNA [7]. However, these strategies simplify the myriad factors involved in an in vivo model or modify the gene expression profile by altering the typical course of infection.

Currently, two studies describe the Mtb transcriptome during in vivo murine infection. The first [6] used DNA microarrays to compare the in vivo gene expression in BALB/c vs. BALB/c $^{SCID/SCID}$ vs. bacterial cultures at different time points of infection. The second [5] compared the bacteria transcriptome of Mtb Erdman marked with fluorescent reporter mCherry between alveolar and interstitial macrophages isolated directly from infected mouse lungs at day 14 post-infection using RNA-seq.

Undoubtedly, the use of RNA-seq offers a more sensitive and quantitative approach to determine the expression of genes and intergenic regions. However, the use of specialized equipment to separate infected from non-infected cells or significantly increasing the sequencing depth to enrich the information on the pathogens' gene expression entails additional expenses not available for all research laboratories worldwide. Overall, the purpose of this study is to present the efficiency and reproducibility of our experimental method to increase the number of observed pathogen expressed genes during tuberculosis in vivo infection. This method could be used in the future to explore the pathogen gene expression profile at different time points of the infection and use different Mtb-infecting strains.

Here, we used a well-characterized murine model for pulmonary tuberculosis [8,9] to obtain the Mtb and mouse transcriptome from lungs 21-days post-infection using an experimental approach involving differential cell lysis and a probe-based ribosomal depletion.

This murine model is based on the intratracheal instillation of live Mtb bacilli into male BALB/c mice and shows the evolution of the disease in two clear phases, an acute phase that spans from day 1–28 post-infection, characterized by inflammatory infiltrate and formation of granulomas, and an advance phase, from day 28 onwards, characterized by pneumonia, focal necrosis, and fibrosis [9]. Furthermore, we used male mice because of the essential differences in their immune response compared to female mice [10]. Female mice show a hyper-inflammatory response and better protection against tuberculosis, while male mice show anti-inflammatory responses probably favored by testosterone, and are more susceptible to infection.

2. Materials and Methods

2.1. Bacteria Cultures

The *M. tuberculosis* (Mtb) H37Rv strain was cultured in Middlebrook 7H9 broth (Millipore, Burlington, MA, USA, Cat. M0178) enriched with ADC Growth Supplement (Millipore, Burlington, MA, USA, Cat. M0553) at 37 °C. The optical density (OD) was monitored weekly, and the purity was assessed with a Zihel-Neelsen stain. As soon as the culture reached the mid-logarithmic phase (OD = 0.6), bacilli were harvested and aliquoted, adjusting to 2.5×10^5 colony-forming units (CFU) in 100 μL of phosphate-buffered saline

(PBS) and kept at −80 °C until use. Before inoculation, the frozen stock was thawed, diluted, and sonicated to disperse clumps.

2.2. Mice Infection

The experimental model of pulmonary tuberculosis used in this study has been previously described [8,9,11]. Briefly, pathogen-free male BALB/c mice of 6–8 weeks of age were anesthetized with sevoflurane (Sevorane®) (Abbvie, IL, EUA) and inoculated intratracheally using a feeding needle with 2.5×10^5 CFU of the H37Rv Mtb strain.

Twenty-one days post-infection, mice were euthanized, and both lungs were extracted aseptically, snap-frozen in liquid nitrogen, and kept at −80 °C until use. All the infection and lung extraction procedures were performed in a Class 3 biological safety cabinet.

2.3. M. tuberculosis Transcriptome

To analyze the gene expression for the pathogen and the host, we used both lungs of each mouse to obtain the bacteria and mouse RNA. We performed three strategies to enrich the mycobacterial RNA in the infected lungs and obtain the Mtb transcriptome. Two strategies were implemented during the RNA extraction, and the third, after the sequencing library construction.

2.4. I. M. tuberculosis Transcriptome

RNA extraction. The infected left side lungs were assigned randomly to each strategy, one lung to Strategy 1, and three lungs to Strategies 2 and 3.

(i) RNA extracted without differential lysis and centrifugation (Strategy 1): We pulverized one infected left lung using sterile mortars and pestles frozen with liquid nitrogen and placed the homogenate in a microfuge tube. Then, total RNA was extracted directly using the Quick RNA miniprep Kit (Zymo Research, Irvine, CA, USA; Cat.R1055) following the manufacturer's recommendations, including the digestion with DNAseI to remove the contaminating DNA. Finally, we quantified and assessed the quality of the resulting RNA by Qubit Fluorometer (Invitrogen, Waltham, MA, USA, Cat. Q32851) and Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA, Cat. 5067-4626), respectively.

(ii) RNA extracted with differential lysis and centrifugation (Strategy 2): We pulverized six infected left lungs independently using sterile mortars and pestles frozen with liquid nitrogen and placed each homogenate in microfuge tubes. Then, we added RLT buffer (Qiagen, Hilden, Germany, Cat.79216) with β mercaptoethanol as the initial lysis buffer to each tube and centrifuged at 14,000 rpm/4 °C for 5 min. We discarded the supernatant and kept the cream-colored pellet on ice. This procedure allowed the enrichment of bacterial cells in the pellet due to the mild-lysis produced by the RLT, which breaks most mouse cells keeping the bacterial cells intact due to its thicker wall membrane. Immediately after centrifugation, we continued the RNA extraction from the pellet using the Quick RNA miniprep Kit (Zymo Research, Irvine, CA, USA; Cat.R1055) following the manufacturer's recommendations, including the digestion with DNAseI to remove the contaminating DNA. Finally, we quantified and assessed the quality of the resulting RNA by Qubit Fluorometer and Agilent 2100 Bioanalyzer, respectively.

Construction of sequencing libraries. The resulting RNA extracted from each lung with Strategies 1 and 2 was treated independently with the Ribo-Zero rRNA Epidemiology Removal Kit (Illumina, San Diego, CA, USA; Cat.MRZE706) following the manufacturers' recommendations. Next, 700 µg of depleted RNA was used as input for the NEBNext Ultra RNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, MA, USA; Cat.E7530S), adjusting the Size Select conditions for insert sizes between 400 and 600 bp and the enrichment PCR to 15 cycles. The quantity and quality of all libraries were assessed by Qubit Fluorometer and Agilent 2100 bioanalyzer, respectively. Finally, the library from Strategy 1 and three libraries from strategy 2 were sequenced directly in a NextSeq 500 Mid Output cell in a 150-cycle paired-end format at the National Institute of Genomic Medicine

(INMEGEN) in Mexico City, Mexico. The remaining three libraries from Strategy 2 were further treated with the in-house ribosomal subtractive hybridization.

In House Ribosomal Substractive hybridization (Strategy 3).Three independent libraries constructed with RNA from Strategy 2 were hybridized with in-house ribosomal probes to reduce Mtb rRNA sequences. Briefly, we amplified the rRNA from the Mtb genome and sheared the amplicons on a Covaris instrument to an average size of 100–300 bp. The resulting fragments were processed with the NEBNext Fast DNA Library Prep Set for Ion Torrent (New England BioLabs, Ipswich, MA, USA; Cat.E6270S), further amplified with biotinylated primers and purified with AMPure XP beads (Beckman-Coulter, Pasadena, CA, USA; Cat.A63880).

Next, each library was hybridized with the in-house ribosomal probes for 72 hrs using a temperature ramp from 95 °C to 65 °C. After hybridization, we used magnetic streptavidin-coated beads (Dyabeads MyOne Streptavidin C1; Invitrogen, Waltham, CA, USA; Cat.65001) and washing buffers at a range temperature of 65–99 °C to gradually pull down the biotinylated probes and separate the captured rRNA. Finally, the non-captured fraction of the libraries was quantified by Qubit Fluorometer, and the quality was assessed by Agilent 2100 bioanalyzer. The non-captured fraction of the libraries were sequenced in a NextSeq 500 Mid Output cell in a 150-cycle paired-end format at the INMEGEN in Mexico City, Mexico.

2.5. Mouse Transcriptome

For the host transcriptome, we used the right lung of the same three mice used in Strategy 3 for the Mtb transcriptome.

RNA extraction. We pulverized each of the three lungs using a sterile mortar and pestle frozen with liquid nitrogen. Following the manufacturer's recommendations, the resulting homogenate was processed with the Quick RNA miniprep kit (Zymo Research, Irvine, CA, USA; Cat.R1055), including digestion with DNAseI to remove the contaminating DNA. After the extraction, we quantified and assessed the RNA quality by Qubit Fluorometer and Agilent 2100 bioanalyzer, respectively.

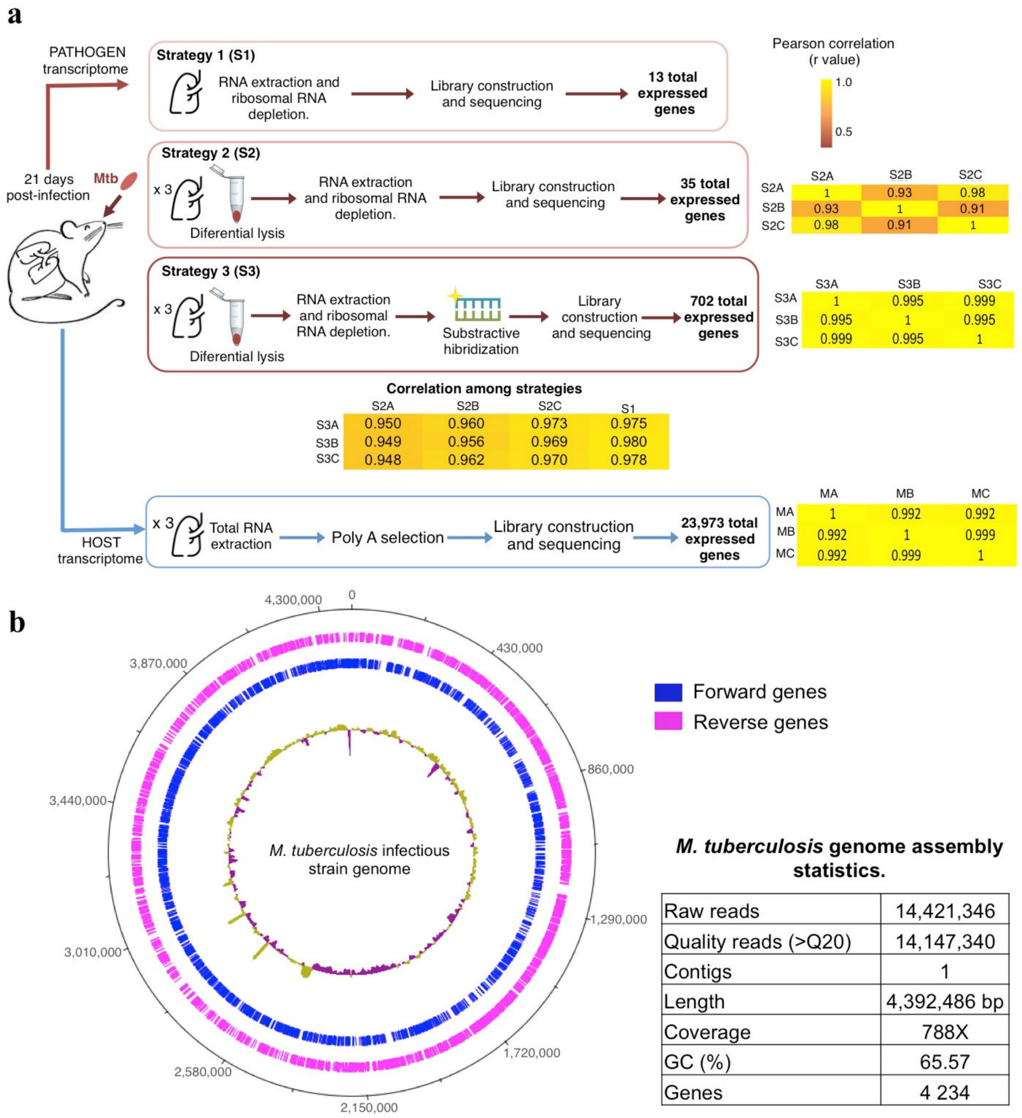
Construction of RNA sequencing libraries.Total RNA from each lung was treated with the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England BioLabs, Ipswich, MA, USA; Cat.E7490) following the manufacturer's recommendations. Then, 700 μg of the isolated mRNA was used as input for the NEBNext Ultra RNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, MA, USA; Cat.E7530S), adjusting the Size Select conditions for insert sizes between 400 and 600 bp and the enrichment PCR to 15 cycles. After the procedure, we quantified and assessed each library's quality by Qubit Fluorometer and Agilent 2100 bioanalyzer, respectively. The final libraries were sequenced in a NextSeq500 Mid Output cell in a 300-cycle paired-end format at the INMEGEN in Mexico City, Mexico.

2.6. Sequencing and Assembly of the M. tuberculosis Infectious Strain

We extracted the genomic DNA from a previously harvested aliquot of the infecting H37Rv Mtb strain using the Quick-DNA Fecal/Soil Microbe Miniprep Kit (Zymo Research, Irvine, CA, USA; Cat.D6010) following the manufacturer's recommendations. The quantity and quality of the resulting DNA were determined using agarose gel electrophoresis and Qubit fluorometer, respectively.

The sequencing library was constructed with the Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA, USA; Cat.FC-131-1024) following the manufacturer's recommendations and selecting an insert size of 400–600 bp. The final library was quantified with the Qubit fluorometer, and the size distribution was analyzed with Agilent 2100 bioanalyzer. This library was sequenced with the MiSeq Output cell in a 500 cycle paired-end format at the INMEGEN in Mexico City, Mexico.

The sequencing run produced 14,421,346 total paired reads, 98.1% of which remained after quality filters ($\geq$Q20) and adaptor removal with FASTX-toolkit (v0.0.13) (http://hannonlab.cshl.edu/fastx_toolkit/index.html, accessed on 28 June 2021), and Trimmomatic (v0.36) (http://www.usadellab.org/cms/index.php?page=trimmomatic released 2014, accessed on 28 June 2021). Then, we used SPADES (v.3.13.9) (https://github.com/ablab/spades, accessed on 28 June 2021) and MeDuSa (http://combo.dbe.unifi.it/medusa, accessed on 28 June 2021) to build the de novo assembly. This resulted in 1 contig with 4,392,486 bp and average sequence depth coverage of 788× (Figure 1b).


a
PATHOGEN transcriptome
21 days post-infection
Mtb
Strategy 1 (S1)
RNA extraction and ribosomal RNA depletion.
Library construction and sequencing
13 total expressed genes
Strategy 2 (S2)
x 3
Diferential lysis
RNA extraction and ribosomal RNA depletion.
Library construction and sequencing
35 total expressed genes
Strategy 3 (S3)
x 3
Diferential lysis
RNA extraction and ribosomal RNA depletion.
Substractive hibridization
Library construction and sequencing
702 total expressed genes
HOST transcriptome
x 3
Total RNA extraction
Poly A selection
Library construction and sequencing
23,973 total expressed genes
Pearson correlation (r value)
1.0
0.5
b
0
430,000
860,000
1,290,000
1,720,000
2,150,000
2,580,000
3,010,000
3,440,000
3,870,000
4,300,000
M. tuberculosis infectious strain genome
Forward genes
Reverse genes

	S2A	S2B	S2C
S2A	1	0.93	0.98
S2B	0.93	1	0.91
S2C	0.98	0.91	1

	S3A	S3B	S3C
S3A	1	0.995	0.999
S3B	0.995	1	0.995
S3C	0.999	0.995	1

Correlation among strategies

	S2A	S2B	S2C	S1
S3A	0.950	0.960	0.973	0.975
S3B	0.949	0.956	0.969	0.980
S3C	0.948	0.962	0.970	0.978

	MA	MB	MC
MA	1	0.992	0.992
MB	0.992	1	0.999
MC	0.992	0.999	1

***M. tuberculosis* genome assembly statistics.**

Raw reads	14,421,346
Quality reads (>Q20)	14,147,340
Contigs	1
Length	4,392,486 bp
Coverage	788X
GC (%)	65.57
Genes	4 234

Figure 1. (**a**) Experimental strategies to obtain the Mtb (S1, S2, and S3) and mouse (M) transcriptome profiles. Pearson correlations between three biological replicates (A, B, and C) and strategies (S) are shown in yellow heat maps. (**b**) Assembled genome of the *M. tuberculosis* infectious strain. Regions with depth coverage lower and above average (788×) are shown as purple and yellow inner circles, respectively. In addition, the statistic parameters of the assembly were reported in the table.

We used Glimmer (v3.02) [12] within Blast2Go (v5.2) [13] with the Prokaryotic Gene Finding feature to locate 4234 genes and determined a homolog for 99.24% of them analyzing against the non-redundant (nr) database with BLASTX, setting the E-value cut-off at 1.0×10^{-3}. The transcripts identified as members of the PPE-PGRS family by blast were further searched in the Mycobrowser database [14] to have a more specific description of the PPE-PGRS name. Additionally, all genes were associated with protein families through InterProScan, and functionally mapped to GO terms setting the following parameters:

E-value-hi-filer: 1.0 × 10^{-3}; annotation cut-off: 55 and GO weight: 5. Finally, the non-coding RNA regions present in the assembled Mtb genome were determined with Infernal (v1.1.3) [15] comparing against the Rfam 14.1 database. The resulting coordinates of the identified regions were added to the gff file with all the annotations.

2.7. Analysis of Host and Pathogen RNA-Seq Data

2.7.1. Analysis of *M. tuberculosis* Transcriptome

Pair-end reads were checked for quality > 20 Q with FASTX-toolkit (v0.0.13), (http://hannonlab.cshl.edu/fastx_toolkit/index.html, accessed on 28 June 2021) and adaptors were trimmed using Trimmomatic (v0.36) (http://www.usadellab.org/cms/index.php?page=trimmomatic, accessed on 28 June 2021). Next, as Avraham et al. 2016 [1] suggested, we built a composed database to analyze the mixed host–pathogen reads and minimize spurious read alignments. This database contained the mouse and Mtb rRNA sequences, mouse reference genome GRCm38.p6 (GenBank GCA_000001635.8), and the de novo assembly of the Mtb infectious genome. All read mappings were performed using SMALT (v0.7.6) (https://github.com/rcallahan/smalt, accessed on 28 June 2021), adjusting strict parameters to avoid cross-mapping reads that will affect transcript quantification. We only considered as positive reads that mapped with a minimum coverage of ≥80%.

The reads mapped to rRNA and mouse sequences were counted and separated using the Samtools suite (v1.3.1) (https://sourceforge.net/projects/samtools/files/samtools/, accessed on 28 June 2021). Finally, to determine the number of reads mapped to Mtb genes and non-coding regions, we intersected the aligned bam file with the corresponding gff file containing all the annotations using Bedtools (v2.26.0) (https://github.com/arq5x/bedtools2/releases, accessed on 28 June 2021). The gene count of each library was normalized by reads per kilobase per million mapped reads (RPKM), and Pearson rank correlations between replicates and between control and experimental libraries were calculated with the GraphPad Prism7 (GraphPad software, San Diego, CA, USA).

Further, we used Blast2Go (BioBam Bioinformatics, Valencia, Spain) to perform an InterPro and gene ontology (GO) enrichment analysis using Fisher's exact test, considering a significant *p*-value of <0.05 and the complete Mtb assembled genome as a reference. We used the KEGG database with a bi-directional best-hit method (BBH) and *Mycobacterium tuberculosis* as the reference gene set, then determined the percentage of each pathway represented by our set of expressed genes.

To analyze the presence of expressed genes corresponding to secreted proteins, we aligned the sequences of the 529 most expressed genes against the published secretome of *M. tuberculosis* [16], considering as positive an E-value < 0.001 and coverage > 70%. Additionally, we constructed 100 groups of 529 genes randomly selected from the 4234 total Mtb genes. Then, we compared each group with the Mtb secretome and determined the number of groups resulting in more than 85 proteins aligned with an E-value < 0.001 and >70% coverage.

Finally, to analyze if the distribution of the 12 genes from the Top 20 list in the 1138 kb region was a random event, we first divided our Mtb assembled genome into four sections, each accounting for ~25% of the total length. Next, we created a hundred groups of 20 random genes selected from the 702 observed expressed genes and observed their location. Lastly, for each group, we counted the number of genes located within the same region of the genome and considered a positive event if there were 12 or more genes located together.

2.7.2. Analysis of Mouse Transcriptome

Pair-end reads were checked for quality > 20 Q with FASTX-toolkit (v0.0.13) (http://hannonlab.cshl.edu/fastx_toolkit/index.html, accessed on 28 June 2021), and adaptors were trimmed using Trim_Galore (v0.4.2) (https://github.com/FelixKrueger/TrimGalore, accessed on 28 June 2021). Next, filtered reads were aligned against the *Mus musculus* NCBI reference genome GRCm38.p6 (GenBank GCA_000001635.8) using Bowtie2 (v2.3.5) (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml, accessed on 28 June 2021) with default parameters. Next, the raw count expression profiles were obtained using HTSeq (version 0.6.1) (https://htseq.readthedocs.io/en/master/history.html#version-0-6-1, accessed on 28 June 2021), then, gene count was normalized by RPKM.

We used the gene ontology (GO) database to analyze enriched GO terms in the Biological Process, Molecular Function, and Cellular Component categories using a Fisher's exact test, a significant *p*-Value of <0.05, and *Mus musculus* as the reference genome. Additionally, we used the KEGG database to analyze the most abundant KEGG pathways with a bi-directional best-hit method (BBH) and *Mus musculus* as the reference gene set.

3. Results and Discussion

3.1. The Differential Cell Lysis Protocol Concentrated the Number of Mycobacterial Cells and Increased the Extracted Bacterial RNA

As a first approach, we sequenced the total RNA extracted from one infected lung (Figure 1a) to explore the expression obtained following the typical RNA-seq strategy (Strategy 1, S1). Next, we analyzed the sequencing data mapping against a multifasta file containing the mouse reference genome, Mtb rRNA sequences, and the Mtb strain genome used for the infection (see materials and methods) (Figure 1b). The analysis showed that only 1.70% of sequences belonged to Mtb, representing 13 expressed genes (Table S1).

Thus, we standardized a differential cell lysis protocol to concentrate the number of mycobacterial cells before the RNA extraction (Strategy 2, S2) (Figure 1a). In this case, the Zihel-Neelsen staining showed a bacterial cell concentration (Figure 2), and the sequences from three biological replicates confirmed a ~4% increase in the proportion of Mtb sequences compared to S1 (Table S1).

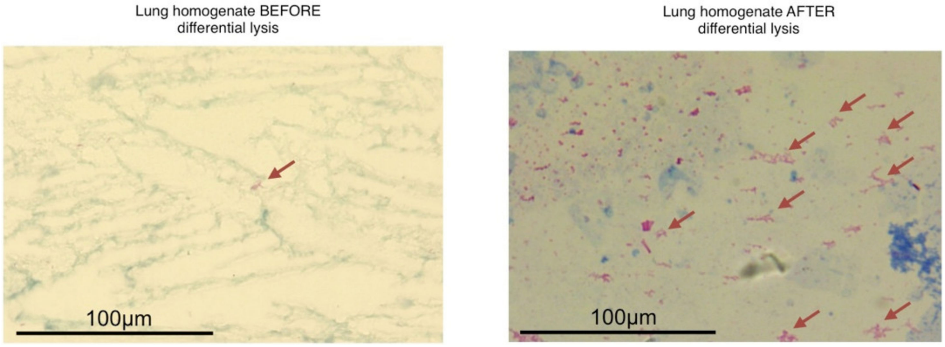
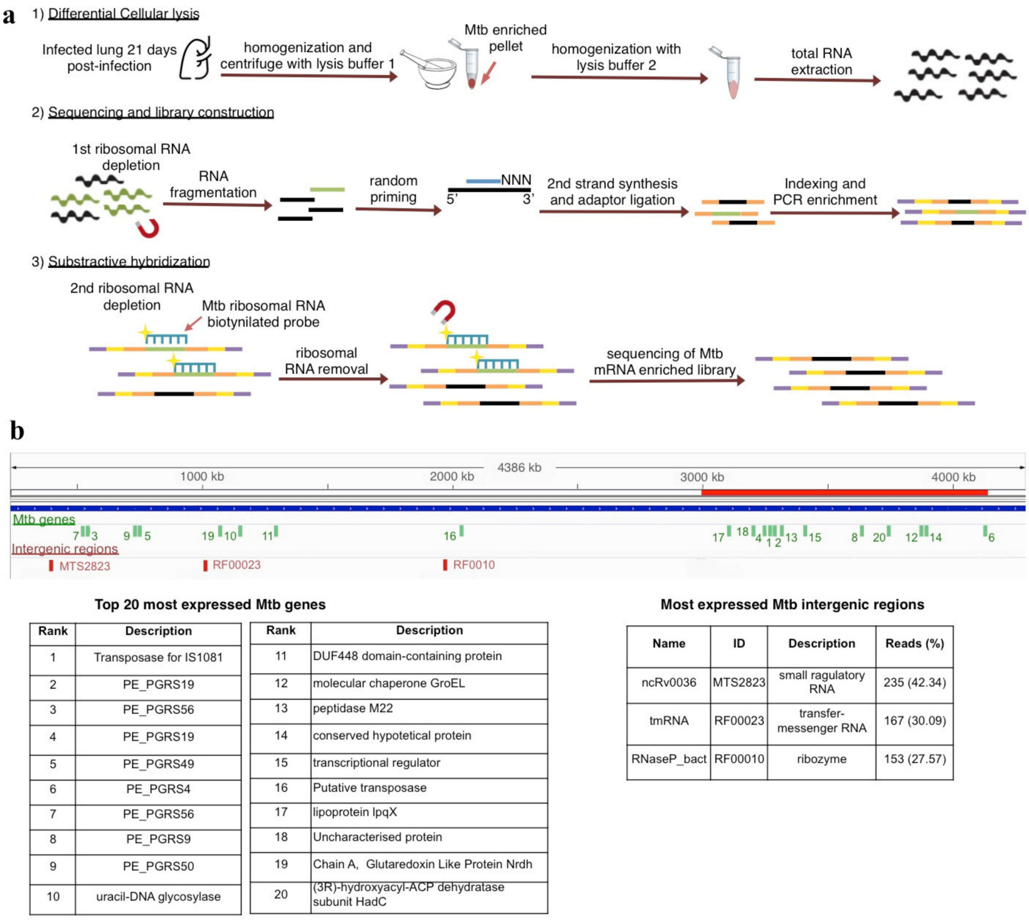


Figure 2. Zihel-Neelsen staining of the infected lung homogenate before and after differential cellular lysis. The red arrow shows the bacilli per field.

However, despite using a ribosomal depletion kit, ~97% of the Mtb sequences still belonged to ribosomal transcripts, and the remaining sequences represented the expression of 35 genes. Some studies have reported that commercial kits remove 70–85% of ribosomal RNA in pure mycobacterial cultures [17]. However, our samples derived from infected tissue, and the mix of host and mycobacteria cells might have reduced the depletion efficiency.

To reduce the amount of ribosomal RNA, we designed biotinylated probes to selectively subtract the ribosomal transcripts from the total RNA of three independent infected lungs (Strategy 3, S3) (Figures 1a and 3a) (see materials and methods). This method successfully decreased 24.56% of Mtb ribosomal transcripts, allowing the observation of 702 expressed genes (Table S2), a ~50 fold enrichment in the gene number compared to S1. Additionally, Pearson correlations of the gene expression abundance between strategies showed no bias due to the differential lysis and probe hybridization methods (r > 0.9) as well as good reproducibility among samples (Figure 1a).


a
1) Differential Cellular lysis
Infected lung 21 days post-infection
homogenization and centrifuge with lysis buffer 1
Mtb enriched pellet
homogenization with lysis buffer 2
total RNA extraction
2) Sequencing and library construction
1st ribosomal RNA depletion
RNA fragmentation
random priming
NNN
5'
3'
2nd strand synthesis and adaptor ligation
Indexing and PCR enrichment
3) Substractive hybridization
2nd ribosomal RNA depletion
Mtb ribosomal RNA biotynilated probe
ribosomal RNA removal
sequencing of Mtb mRNA enriched library
b
4386 kb
1000 kb
2000 kb
3000 kb
4000 kb
Mtb genes
7 3 9 5 19 10 11 16 17 18 4 1 2 13 15 8 20 12 14 6
Intergenic regions
MTS2823
RF00023
RF0010

Top 20 most expressed Mtb genes

Rank	Description
1	Transposase for IS1081
2	PE_PGRS19
3	PE_PGRS56
4	PE_PGRS19
5	PE_PGRS49
6	PE_PGRS4
7	PE_PGRS56
8	PE_PGRS9
9	PE_PGRS50
10	uracil-DNA glycosylase

Rank	Description
11	DUF448 domain-containing protein
12	molecular chaperone GroEL
13	peptidase M22
14	conserved hypotetical protein
15	transcriptional regulator
16	Putative transposase
17	lipoprotein lpqX
18	Uncharacterised protein
19	Chain A, Glutaredoxin Like Protein Nrdh
20	(3R)-hydroxyacyl-ACP dehydratase subunit HadC

Most expressed Mtb intergenic regions

Name	ID	Description	Reads (%)
ncRv0036	MTS2823	small ragulatory RNA	235 (42.34)
tmRNA	RF00023	transfer-messenger RNA	167 (30.09)
RNaseP_bact	RF00010	ribozyme	153 (27.57)

Figure 3. Final strategy to obtain Mtb gene expression profile and genome location of the twenty most expressed genes. (**a**) Detailed experimental procedure used to obtain the transcriptome profile in Strategy 3. (**b**) Genomic location of the Top 20 most expressed Mtb genes (green vertical lines) and the three most expressed intergenic regions (red vertical lines) across the Mtb-sequenced genome. The red horizontal line indicates the 1138 kb region of the Mtb genome (from 3,005,716 bp to 4,144,624 bp) that concentrated twelve of the top 20 most expressed genes. The tables indicate each expressed gene ranked from the most to the least expressed and the percentage of reads mapped to each intergenic region.

3.2. The MTb Gene Expression Profile Showed a High Abundance of Genes Codifiyng PE-PGRS Members and the Insertion Sequence IS1081

We used the data obtained with Strategy 3 (S3) to describe the gene expression profile of Mtb during the in vivo infection. Interestingly, the data showed that 61.59% of the non-ribosomal Mtb sequences represented 702 genes (Table S2), while 38.41% represented 30 intergenic regions. To avoid analyzing transitory transcripts, we only considered 529 genes with ≥2 reads in at least one sample and an average RPKM ≥ 1 for further analysis (Table S2).

First, we determined the InterPro and Gene Ontology (GO) annotations significantly ($p < 0.05$) enriched in the 529 most expressed genes as compared to the complete Mtb genome. The InterPro analysis showed a significant ($p < 0.05$) enrichment of eight protein domains related to fatty acid metabolism and two protein families representing type VII secretion system subunits (Table 1 and Table S3).

Table 1. InterPro domains and families significantly overrepresented in the 529 most expressed *M. tuberculosis* genes.

InterPro ID	Category	Description	*p*-Value
IPR014043	domain	Acyl transferase	6.267×10^{-7}
IPR009081	domain	Phosphopantetheine binding ACP domain	1.438×10^{-6}
IPR014030	domain	Beta-ketoacyl synthase, N-terminal	2.747×10^{-5}
IPR014031	domain	Beta-ketoacyl synthase, C-terminal	4.885×10^{-5}
IPR013968	domain	Polyketide synthase, ketoreductase domain	1.624×10^{-4}
IPR023836	family	EccCa-like, Actinobacteria	1.087×10^{-3}
IPR023837	family	EccCb-like, Actinobacteria	1.087×10^{-3}
IPR032821	domain	Ketoacyl-synthetase, C-terminal extension	1.190×10^{-3}
IPR006091	domain	Acyl-CoA oxidase/dehydrogenase, central domain	1.363×10^{-3}
IPR020807	domain	Polyketide synthase, dehydratase domain	1.757×10^{-3}
IPR009075	domain	Acyl-CoA dehydrogenase/oxidase C-terminal	1.995×10^{-3}
IPR002543	domain	FtsK domain	6.180×10^{-3}
IPR003029	domain	S1 domain	7.039×10^{-3}
IPR003495	domain	CobW/HypB/UreG, nucleotide-binding domain	7.039×10^{-3}
IPR023753	domain	FAD/NAD(P)	1.360×10^{-2}
IPR000788	domain	Ribonucleotide reductase large subunit, C-terminal	1.558×10^{-2}
IPR001030	domain	Aconitase/3-isopropylmalate dehydratase large subunit, alpha/beta/alpha domain	1.558×10^{-2}
IPR002300	domain	Aminoacyl-tRNA synthetase, class Ia	1.558×10^{-2}
IPR003714	domain	PhoH-like protein	1.558×10^{-2}
IPR004100	domain	ATPase, F1/V1/A1 complex, alpha/beta subunit, N-terminal domain	1.558×10^{-2}

As expected, the GO analysis showed a significant ($p < 0.05$) enrichment of terms related to lymphocyte and T-cell co-stimulation, T-cell receptors, nitrate reductase, zymogen binding, and scavenger response (Table 2 and Table S4). Accordingly, the KEGG pathway analyses of the 529 most expressed genes showed fatty acid metabolism, secretion, T-cell and lymphocyte co-stimulation, glyoxylate and dicarboxylate metabolism, fatty acid synthesis, secretion systems, and quorum sensing as some of the pathways covered ≥20% by the expressed genes (Table 3 and Table S5).

The InterPro, GO, and KEGG analysis showed an enrichment of terms related to energy production and macromolecules synthesis. This profile was expected and is consistent with the exponential growth phase of the bacteria during this infection stage [8,9]. Moreover, the abundant presence of pathways for glyoxylate metabolism indicated that the mycobacteria are adapting to a stressful environment [4] using lipids as a carbon source [18]. Additionally, the expression of genes related to T-cell stimulation and cell recruitment is consistent with the day post-infection we are analyzing (day 21), where the granuloma formation occurs [8,9].

The most over-expressed Mtb gene was the transposase for IS1081 (Rv2512c) (Table 4), an insertion sequence used as a molecular marker for *M. bovis* and Mtb but with no previous reports of expression during Mtb in vivo infections. In contrast, the insertion sequence IS6110, besides being considered the standard molecular marker for Mtb strains [19], has also been proposed as a promoter that can affect mycobacterial fitness given its location within the genome of some strains [19]. Some authors suggest it can affect the bacterial gene expression by causing a frameshift that affects the transcription or producing an RNA pseudoknot, which interferes with translation [20]. Additionally, increased transposase activity for IS6110 was observed in mice infected with Mtb H37Rv and in bacterial liquid cultures with nutrient deficiency [20]. These results suggest that the transposition of IS6110 may respond to stress conditions such as the one produced during an in vivo infection. However, the IS6110 transposase showed a low expression ranking under 300 of the 529 expressed genes in our data.

Table 2. Gene Ontology terms significantly enriched by category related to the 529 most expressed *M. tuberculosis* gene separated by category.

GO ID	GO Name	$p-$Value
A. Biological Process		
GO:0031295	T-cell co-stimulation	1.765×10^{-5}
GO:0031294	Lymphocyte co-stimulation	1.765×10^{-5}
GO:0009081	Branched-chain amino acid metabolic process	4.007×10^{-5}
GO:0009083	Branched-chain amino acid catabolic process	1.102×10^{-4}
GO:0006549	Isoleucine metabolic process	1.197×10^{-4}
GO:0000288	Nuclear-transcribed mRNA catabolic process, Deadenylation-dependent decay	2.663×10^{-4}
GO:0006573	Valine metabolic process	3.513×10^{-4}
GO:0060184	Cell cycle switching	4.878×10^{-4}
GO:0051728	Cell cycle switching, mitotic to meiotic cell cycle	4.878×10^{-4}
GO:0006574	Valine catabolic process	5.108×10^{-4}
B. Cellular Component		
GO:0005643	Nuclear pore	4.20×10^{-3}
GO:0000932	P-body	5.39×10^{-3}
GO:0009986	Cell surface	6.00×10^{-3}
GO:0098978	Glutamatergic synapse	6.18×10^{-3}
GO:1990527	Tec1p-Ste12p-Dig1p complex	7.04×10^{-3}
GO:0110165	Cellular anatomical entity	1.23×10^{-2}
GO:1990526	Ste12p-Dig1p-Dig2p complex	1.60×10^{-2}
GO:0030496	Midbody	1.81×10^{-2}
GO:0009325	Nitrate reductase complex	1.83×10^{-2}
GO:0016020	Membrane	2.56×10^{-2}
C. Molecular Function		
GO:0004085	Butyryl-CoA dehydrogenase activity	1.62×10^{-4}
GO:0005488	Binding	6.96×10^{-4}
GO:0052890	Oxidoreductase activity, acting on the CH-CH group of donors, with a flavin as acceptor	7.83×10^{-4}
GO:0043168	Anion binding	1.16×10^{-3}
GO:0003955	NAD(P)H dehydrogenase (quinone) activity	1.94×10^{-3}
GO:0017056	Structural constituent of nuclear pore	2.87×10^{-3}
GO:0004962	Endothelin receptor activity	2.94×10^{-3}
GO:0036094	Small molecule binding	3.15×10^{-3}
GO:1901363	Heterocyclic compound binding	4.09×10^{-3}
GO:0097159	Organic cyclic compound binding	4.10×10^{-3}

Table 3. KEGG pathways overrepresented by the 529 most expressed *M. tuberculosis* genes.

KEGG Id	Pathway	% of the Pathway Represented by the Expressed Genes
03020	RNA polymerase	75.00
00630	Glyoxylate and dicarboxylate metabolism	34.15
03018	RNA degradation	33.33
00562	Inositol phosphate metabolism	33.33
00910	Nitrogen metabolism	31.82
05152	Tuberculosis	30.77
00680	Methane metabolism	28.57
00290	Valine, leucine and isoleucine biosynthesis	28.57
00430	Taurine and hypotaurine metabolism	28.57
00020	Citrate cycle (TCA cycle)	28.13
00730	Thiamine metabolism	27.27
00983	Drug metabolism—other enzymes	27.27

Table 3. *Cont.*

KEGG Id	Pathway	% of the Pathway Represented by the Expressed Genes
00061	Fatty acid biosynthesis	26.67
02024	Quorum sensing	25.00
00010	Glycolysis/Gluconeogenesis	24.24
00270	Cysteine and methionine metabolism	24.24
00260	Glycine, serine, and threonine metabolism	24.00
03060	Protein export	23.53
01053	Biosynthesis of siderophore group nonribosomal peptides	22.22
03070	Bacterial secretion system	21.43

Table 4. Top 20 most expressed *M. tuberculosis* genes at day 21 post-infection.

Expression Ranking	Gen ID	Gene Description	Mean RPKM
1	Rv2512c	Transposase for insertion sequence element IS1081	6.84×10^5
2	Rv1067c	PE_PGRS19 PE-PGR	3.45×10^5
3	Rv3512	PE_PGRS56 PE-PGRS	1.91×10^5
4	Rv1067c	PE_PGRS19 PE-PGR	1.65×10^5
5	Rv3344c	PE_PGRS49 PE-PGR	1.63×10^5
6	Rv0279c	PE_PGRS4 PE-PGRS	1.10×10^5
7	Rv3512	PE_PGRS56 PE-PGRS	5.01×10^4
8	Rv0746	PE_PGRS9 PE-PGRS	4.35×10^4
9	Rv3345c	PE_PGRS50 PE-PGR	1.94×10^4
10	Rv0105c	uracil-DNA glycosylase	7.61×10^3
11	Rv2840c	DUF448 domain-containing protein	6.94×10^3
12	Rv0440	molecular chaperone GroEL	5.71×10^3
13	Rv3620c	peptidase M22	4.88×10^3
14	Rv0454	conserved hypotetical protein	4.58×10^3
15	Rv0967	transcriptional regulator	4.12×10^3
16	Rv2424c	Putative transposase	4.05×10^3
17	Rv1228	lipoprotein lpqX	3.66×10^3
18	Rv2424c	Uncharacterized protein	3.11×10^3
19	Rv3053c	Chain A, Glutaredoxin Like Protein Nrdh	3.09×10^3
20	Rv0637	(3R)-hydroxyacyl-ACP dehydratase subunit HadC	2.96×10^3

Notably, we observed that the genome of Mtb H37Rv has 33 copies for the IS6110 transposase while the IS1081 transposase has only 6, suggesting that the high expression of the transposase for IS1081 we observed was not due to a higher gene copy number. Thus, suggesting the importance of considering the role of IS1081 during the development of Mtb in different environments.

The following eight most expressed genes belonged to the PE-PGRS family (Table 4), considered important Mtb antigens that are in close contact with the host immune system [21]. Particularly, PGRS49 and PGRS50 have been proposed as strong vaccine candidates [22] due to their strong antigenicity. On the other hand, PE-PGRS9 and PE-PGRS53 had been previously related to chronic Mtb infections [21,23]. Interestingly, only one previous in vivo study [6] reported the expression of PE-PGRS members, but not as the primarily expressed genes. Thus, to the best of our knowledge, this is the first RNA-seq study to highlight the role of some PE-PGRS transcripts during the early stages of tuberculosis in vivo infection. In line with the high expression of PE-PGRS, we also observed an increased expression of secretion system subunits necessary to export them, especially ESX-1, ESX-3, and ESX-5 [24].

Interestingly, 12 of the 20 most expressed genes were located within a 1138 kb region, representing 25.95% of the genome (Figure 3b and Table S2). To test if this observation was coincidental, we formed 100 groups of 20 random genes selected from the complete set of 702 expressed genes observed and determined their location. Notably, neither group showed more than nine genes within the same region (p = 0.00).

In a previous study, Talaat [6] defined a 34.1 kb region concentrating 20 significantly expressed genes during in vivo infection, and interestingly, our 1138 kb region contained five of those 20 genes. These results suggest that the transcription of this section of the genome is essential for the host–pathogen interaction.

3.3. An Increased Proportion of Expressed Genes Belong to the MTb Secretome

We compared how many of the 529 expressed genes corresponded to the Mtb secretome previously reported [16]. According to the previous publication [16], we expected that only ~12% (63 genes) belonged to the secretome, but we observed that 16.07% (85 genes) codified for secreted proteins (Table S6). These suggested a slight overabundance of the genes for secreted proteins in the transcriptome. To test if this observation was coincidental, we formed 100 groups of 529 random genes selected from 4234 total Mtb genes and compared their presence in the secretome. As a result, 41 from the 100 groups had more than 85 genes aligned with the secretome (E-value < 0.001 and >70% coverage). These results suggest that the over-abundance of genes for secreted proteins had a low probability of being a random event.

Among the expressed transcripts for secreted proteins, we observed seven lipoproteins, such as lppN (Rv2270) and LpqH (Rv3763), which are considered important for the host–pathogen interaction (Table S6). Some studies suggest they help internalize the bacteria into the macrophages [25]. Particularly LpqH has shown some effects that favor the host, such as increasing IL-12, while other effects favor the pathogen, such as inhibiting INF γ, and decreasing the expression of MCH-II and antigen processing [26–28]. In the study by Pisu [5], they detect some lipoproteins responsible for the degradation of triglycerides and cholesterol overexpressed in infected alveolar macrophages, such as Lpl and LipA. However, they did not detect overexpression of LpqH or lppN, probably because these proteins respond to different stimuli in other cell types.

On the other hand, 28.24% (24 genes) of the 85 expressed transcripts for secreted proteins were members of the PE-PGRS family (Table S6). Four of them were some of the most expressed genes (PGRS-49, PGRS56, PGRS9 y PGRS50) (Table 4). We also observed PE68, a gene codified in the region of difference 1 (RD1), which is absent in non-virulent Mtb strains such as BCG [29]. PE68 is considered an efficient immunomodulator that stimulates the host immune cellular response, even promoting tissue damage. Interestingly, although PE68 is not considered essential for the survival of Mtb, some studies suggest that alterations in its sequence affect strain virulence [29]. This transcript ranked in the 25th place of the 529 expressed genes.

3.4. A High Expression of Three ncRNAs during Mtb In Vivo Infection

From the whole Mtb sequences, 38.41% mapped to 30 intergenic regions. Previous reports showed that the expression of intergenic regions in liquid cultures during exponential growth is around 28%, while during the stationary phase, their expression increased to 58% [30]. These results suggest that the expression of intergenic regions varies during the development of Mtb in different environments.

Among the intergenic regions, we identified three ncRNAs, namely small regulatory RNA MTS2823 (ncRv0036) (42.34% of sequences), the transfer-messenger RNA RF00023 (30.09%), and the ribozyme RF0010 (27.57%) (Figure 3b).

There is scarce information about the role of ncRNAs during the Mtb infection. Mainly, MTS2823 is considered necessary for Mtb growth under different environments, and its expression is especially accumulated in stationary phase cultures [30], mouse lungs with chronic infections [30], and even in dormant mycobacteria [31]. Furthermore, previous reports suggest that MTS2823 may mediate the down-regulation of genes expressed in the exponential growth phase [30]. Although the mechanisms remain to be clarified, more in vivo studies are needed to determine if MTS2823 down-regulates genes during different stages of infection.

Our results suggest that MTS2823 also plays an essential role during the early stages of in vivo infection. Probably, this expression increases in later stages, favoring the maintenance of the chronic infection. Thus, the exciting idea is to favor its expression earlier to control the spread of infection. Contrary, sRNAs induced in the stationary phase such as MRS0997 and MTS1338 were not detected in our dataset [30].

3.5. The Host Gene Expression Showed an Inflammatory Profile Consistent with the Early Stage of Infection

Finally, we sequenced the host RNA from the same infected mice, and after assessing good biological reproducibility (Pearson r = 0.9946) (Figure 1a), we selected 15,677 genes with ≥10 reads in two samples for further analysis (Table S7).

The GO analysis considering the top 5% most expressed genes (Table S8) showed inflammation, antigen processing and presentation, cellular response to lipopolysaccharides, and bacterial evasion activities as some of the most significantly ($p < 0.05$) enriched terms. Accordingly, the KEGG pathway analysis showed antigen processing and presentation, phagolysosome, IL-17, and TNF-α signaling, and Toll receptors activities as some of the pathways represented in ≥10% (Table S9).

In the top 20 most expressed genes, we observed transcripts for two surfactant-associated proteins (Sftpc and Sftpa), three histocompatibility genes, and microRNA 638. Notably, Sftpa is known to act as an opsonin to enhance the ingestion of Mtb and other pulmonary pathogens by alveolar macrophages [32–34]. However, interestingly, this internalization pathway has also been shown to decrease reactive nitrogen intermediates levels, suggesting binding of Sftpa may be one mechanism by which Mtb reduced the macrophage cytotoxic response [34,35].

The dominant immune cellular response in the early infection has been previously described for this murine model (8), consistent with previous studies [36]. In contrast, the expression of genes related to T-cell and lymphocyte co-stimulation and antigen processing is typically associated with chronic infections [37] but maybe involved with granuloma formation and maintenance in early stages [38].

Interestingly, we observed microRNA mi6381 as one of the twenty most expressed genes. This microRNA has no previous reports associating it to tuberculosis or any other pulmonary disease. On the other hand, several in vitro studies suggest that micro RNAs such as mir155, mir135b, or mir146a play essential roles in mycobacterial infections [39–43]. Mainly mir155 was found overexpressed in mononuclear cells of tuberculosis patients than of healthy individuals [44]. However, we did not observe this miRNA in our list of expressed genes, suggesting that different miRNAs may be active in different infected tissues.

3.6. Comparisson of Mtb Gene Expression Using Other RNA-Seq Methods during In Vivo Infections

As mentioned before, to the best of our knowledge, two studies aim to describe the Mtb gene expression during an in vivo infection. Both studies have essential methodological similarities and differences with the present study, which we highlight in Table 5. Thus, it is essential to note that discrepancies in the observed expressed genes among studies could arise from the methodological and biological differences such as the murine model, the mycobacterial strain, or the criteria established by each group to consider genes as expressed for the analysis. In this sense, the study by Pisu et al. can be considered the most related to our work; however, it carries critical methodological differences (Table 5), which make the gene expression results not directly comparable between methods. Despite that, we compared our results with Pisu [5], who used RNA-seq to establish an "in vivo signature" of 180 genes expressed in macrophages, and Talaat [6], who used microarrays to determined 159 genes significantly expressed in in vivo conditions. We observed that only 16 and 17 of our expressed genes were observed in Pisu's and Talaat's studies, respectively (Figure 4), while five genes were shared by Talaat vs. Pisu. Thus, this suggested that the observed Mtb gene expression among the three studies depends on the different experimental methods and biological variants, such as the murine model and the bacterial strain used.

Table 5. Main characteristics of previous studies describing *M. tuberculosis* in vivo gene expression.

	Talaat et al., 2004 [6]	**Pisu et al., 2020 [5]**	**Cornejo-Granados et al., 2021**
Mtb strain	H37Rv	Erdman ATCC 35801 mCherry	H37Rv
Day post-infection analyzed	7, 14, 21 and 28	14	21
Murine model used	BALB/c and BALB/c SCID/SCID	C57BL/6J WT	BALB/c
Route of infection	Intranasal	Intranasal	Intratracheal
Infected tissue analyzed	Complete infected lungs	Alveolar and interstitial macrophages isolated from infected lungs	Complete infected lungs
RNA extraction method	Total RNA extraction with Tri Reagent (Trizol) from groups of 50 mice	Isolated cells were treated with Trizol and centrifuged to pellet mycobacterial cells. ~80% of the Trizol (containing host RNA) was removed. Finally, total RNA was extracted with fresh Trizol, and a proportion of the host RNA was added back.	Used a differential centrifugation of individual lungs with a mild-lysis buffer to pellet mycobacterial cells. Plus, removing the supernatant containing host RNA followed by total RNA extraction from the mycobacterial cells with a commercial kit.
Ribosomal RNA elimination	-	Ribo-Zero Gold rRNA Removal Kit (Epidemiology)	Ribo-Zero Gold rRNA Removal Kit (Epidemiology) and In-house Mtb ribosomal probes
Methodology for gene expression analysis	DNA Microarrays with oligonucleotides representing Mtb coding sequences.	RNA-seq	RNA-seq
Key insights	-The expression profile of Mtb in SCID mice is most similar to the profile when grown in broth. -Around 49 genes were only expressed in vivo, and 20 of these are contiguous in a delimited area of the Mtb genome.	-In vivo signature of 180 genes upregulated in macrophages. -Transcriptional signatures varied with macrophage ontology. -Alveolar macrophages showed a distinct up regulation of genes for the acquisition and use of fatty acids. -Interstitial macrophages showed an up regulation of genes for iron sequestration.	-About 62.59% of non-ribosomal sequences represented 702 genes, while 38.41% represented intergenic regions. -The transposase for IS1081 was the most expressed gene. -Eight genes for PE-PGRS members are among the most expressed genes.-Three highly expressed ncRNAs (MTS2823, RF00023, and RF00010)

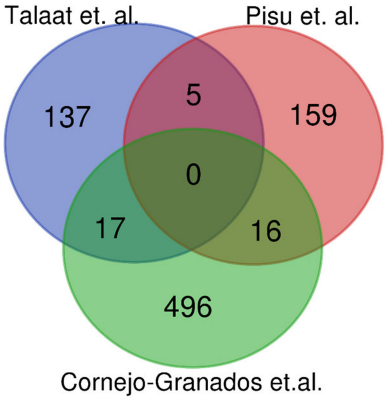


Figure 4. Venn diagram comparing the expressed genes between Talaat et al. [6], Pisu et al. [5], and Cornejo-Granados in vivo studies.

4. Conclusions

The experimental method using the differential cell lysis and a probe-based ribosomal depletion presented here allowed the observation of the most expressed genes and intergenic regions of Mtb directly from an in vivo infection, increasing the number of expressed genes observed, from 13 with the traditional approach to 702 using our experimental method.

We acknowledge that the number of genes analyzed could be broadened, increasing the sequencing depth. However, the main objective at this point was to demonstrate the technical efficacy and reproducibility of our experimental method. Additionally, it is important to mention that our analysis only assessed gene expression levels, from which we can only draw limited conclusions. We are conducting studies using this method and using different Mtb strains, allowing us to understand biological questions using the key value of RNA-seq in a comparative analysis.

The present study revealed the presence of highly expressed MTS2823, a non-coding RNA. Additionally, we observed mi638, a microRNA with no previous reports associating it to tuberculosis or any other pulmonary disease. The analysis we present is a general catalog of the gene expression at one point during the in vivo tuberculosis infection, and it does not intend to draw critical conclusions regarding the pathogen's physiology. In order to better understand the relevance of the expressed genes we observed, it is necessary to analyze and compare additional time points of the disease. In this regard, it is probable that our catalog of expressed genes only reflects the typical cross-talk during the evolution of the host–pathogen interaction. In this sense, our research group has ongoing experiments using this experimental approach to analyze and compare the gene expression during earlier and later stages of the infection and with Mtb strains from different genotypes.

Additionally, it is essential to note that this protocol is not the only experimental approach available for using RNA-seq in an in vivo infection model. As mentioned before, the study by Pisu et al. also uses RNA-seq to analyze the gene expression profile of infected macrophages isolated directly from a murine lung. Definitively, individual cell analysis seams an optimal approach for studying any infectious disease. However, this methodology entails additional expenses in specialized reagents and equipment not typically available for all laboratories. The advantage of our method is that it does not require specialized equipment to conduct a comparative analysis directly from an in vivo infection. Thus, we believe this approach is an efficient and reproducible method that is an alternative for the general scientific community to explore the in vivo gene expression of diverse Tb strains and clinical isolates with prevalence worldwide.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/biology10090848/s1. Table S1: Read count of the three strategies for the *M. tuberculosis* transcriptome. Table S2: *M. tuberculosis* genes expressed 21 days post-infection observed with Strategy 3. Table S3: InterPro domains and families significantly overrepresented in the 529 most expressed *M. tuberculosis* genes. Table S4: Gene Ontology terms significantly enriched by category related to the

529 most expressed *M. tuberculosis* gene. Table S5: KEGG pathways overrepresented by the 529 most expressed *M. tuberculosis* genes. Table S6: *M. tuberculosis* expressed genes belonging to the predicted secretome. Table S7: 15,677 most expressed mouse genes at day 21 post-infection with *M. tuberculosis*. Table S8: Gene Ontology terms significantly enriched by category related to the most expressed mouse genes. Table S9: KEGG pathways overrepresented by the most expressed mouse genes.

Author Contributions: Conceived and designed the experiments: A.O.-L., R.H.-P. and F.C.-G. Performed the experiments: F.C.-G., D.A.M.-E., J.B.-P., B.M.-C., Z.L.Z.-B. and C.M.-R. Analyzed the data: F.C.-G., G.L.-L. and E.E.-M. Contributed reagents/materials/analysis tools: A.O.-L. and R.H.-P. Wrote the paper: F.C.-G., G.L.-L., D.A.M.-E., J.B.-P., B.M.-C., Z.L.Z.-B., C.M.-R., E.E.-M., R.H.-P. and A.O.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the CONACyT grants SALUD-2014-C01-234188, Salud-2013-01-202859 and Ciencia de Frontera 2019 263986. This research was also funded by the DGAPA PAPIIT UNAM (IN215520). We also acknowledge the support of program Actividades de Intercambio Académico 2019 CIC-UNAM-CIAD.

Institutional Review Board Statement: All animal procedures were performed according to the national regulations on animal care and experimentation (NOM 062-ZOO-1999) after approval by the Animal Experimentation Committee at the National Institute of Medical Sciences and Nutrition México (CINVA 1329, approved on 23 February 2015). Our study was also carried out in accordance with the ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines.

Informed Consent Statement: Not applicable.

Data Availability Statement: The RNA-seq data and the *M. tuberculosis* genome used in this study have been deposited in NCBI under BioProject number PRJNA669742. All additional results are available in the Supplementary Materials.

Acknowledgments: We thank Juan Manuel Hurtado Ramírez for the informatics technical support and the computational server maintenance. F.C.-G., thanks to the Doctoral Biochemical Sciences Program, at IBT UNAM and CONACyT by doctoral fellowship: C.V.U.: 443238. G.L.-L. acknowledges the support of CONACyT as a postdoctoral fellow. We also thank the Unidad de Secuenciación Masiva from INMEGEN for sequencing technical support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Avraham, R.; Haseley, N.; Fan, A.; Bloom-Ackermann, Z.; Livny, J.; Hung, D.T. A highly multiplexed and sensitive RNA-seq protocol for simultaneous analysis of host and pathogen transcriptomes. *Nat. Protoc.* **2016**, *11*, 1477–1491. [CrossRef]
2. Westermann, A.J.; Gorski, S.A.; Vogel, J. Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* **2012**, *10*, 618–630. [CrossRef]
3. Schnappinger, D.; Ehrt, S.; Voskuil, M.I.; Liu, Y.; Mangan, J.A.; Monahan, I.M.; Dolganov, G.; Efron, B.; Butcher, P.D.; Nathan, C.; et al. Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages: Insights into the Phagosomal Environment. *J. Exp. Med.* **2003**, *198*, 693–704. [CrossRef]
4. Rohde, K.H.; Abramovitch, R.B.; Russell, D.G. Mycobacterium tuberculosis invasion of macrophages: Linking bacterial gene expression to environmental cues. *Cell Host Microbe* **2007**, *2*, 352–364. [CrossRef] [PubMed]
5. Pisu, D.; Huang, L.; Grenier, J.K.; Russell, D.G. Dual RNA-Seq of Mtb-Infected Macrophages In Vivo Reveals Ontologically Distinct Host-Pathogen Interactions. *Cell Rep.* **2020**, *30*, 335–350. [CrossRef] [PubMed]
6. Talaat, A.M.; Lyons, R.; Howard, S.T.; Johnston, S.A. The temporal expression profile of Mycobacterium tuberculosis infection in mice. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4602–4607. [CrossRef] [PubMed]
7. Waddell, S.J.; Laing, K.; Senner, C.; Butcher, P.D. Microarray analysis of defined Mycobacterium tuberculosis populations using RNA amplification strategies. *BMC Genom.* **2008**, *9*, 94. [CrossRef] [PubMed]
8. Hernandez-Pando, R.; Orozco, H.; Arriaga, K.; Sampieri, A.; Larriva-Sahd, J.; Madrid-Marina, V. Analysis of the local kinetics and localization of interleukin-1 alpha, tumour necrosis factor-alpha and transforming growth factor-beta, during the course of experimental pulmonary tuberculosis. *Immunology* **1997**, *90*, 607–617. [CrossRef] [PubMed]
9. Hernandez-Pando, R.; Orozcoe, H.; Sampieri, A.; Pavon, L.; Velasquillo, C.; Larriva-Sahd, J.; Alcocer, J.M.; Madrid, M.V. Correlation between the kinetics of Th1, Th2 cells and pathology in a murine model of experimental pulmonary tuberculosis. *Immunology* **1996**, *89*, 26–33. [PubMed]
10. Bini, E.I.; Mata Espinosa, D.; Marquina Castillo, B.; Barrios Payan, J.; Colucci, D.; Cruz, A.F.; Zatarain, Z.L.; Alfonseca, E.; Pardo, M.R.; Bottasso, O.; et al. The influence of sex steroid hormones in the immunopathology of experimental pulmonary tuberculosis. *PLoS ONE* **2014**, *9*, e93831. [CrossRef] [PubMed]

11. Lopez, B.; Aguilar, D.; Orozco, H.; Burger, M.; Espitia, C.; Ritacco, V.; Barrera, L.; Kremer, K.; Hernandez-Pando, R.; Huygen, K.; et al. A marked difference in pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes. *Clin. Exp. Immunol.* **2003**, *133*, 30–37. [CrossRef] [PubMed]
12. Delcher, A.L.; Harmon, D.; Kasif, S.; White, O.; Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **1999**, *27*, 4636–4641. [CrossRef] [PubMed]
13. Conesa, A.; Gotz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genom.* **2008**, *2008*, 619832. [CrossRef] [PubMed]
14. Kapopoulou, A.; Lew, J.M.; Cole, S.T. The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis* **2011**, *91*, 8–13. [CrossRef]
15. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29*, 2933–2935. [CrossRef]
16. Cornejo-Granados, F.; Zatarain-Barron, Z.L.; Cantu-Robles, V.A.; Mendoza-Vargas, A.; Molina-Romero, C.; Sanchez, F.; Del Pozo-Yauner, L.; Hernandez-Pando, R.; Ochoa-Leyva, A. Secretome Prediction of Two M. tuberculosis Clinical Isolates Reveals Their High Antigenic Density and Potential Drug Targets. *Front. Microbiol.* **2017**, *8*, 128. [CrossRef]
17. Wang, S.; Dong, X.; Zhu, Y.; Wang, C.; Sun, G.; Luo, T.; Tian, W.; Zheng, H.; Gao, Q. Revealing of Mycobacterium marinum transcriptome by RNA-seq. *PLoS ONE* **2013**, *8*, e75828. [CrossRef]
18. Russell, D.G.; VanderVen, B.C.; Lee, W.; Abramovitch, R.B.; Kim, M.J.; Homolka, S.; Niemann, S.; Rohde, K.H. Mycobacterium tuberculosis wears what it eats. *Cell Host Microbe* **2010**, *8*, 68–76. [CrossRef]
19. Roychowdhury, T.; Mandal, S.; Bhattacharya, A. Analysis of IS6110 insertion sites provide a glimpse into genome evolution of Mycobacterium tuberculosis. *Sci. Rep.* **2015**, *5*, 12567. [CrossRef]
20. Gonzalo-Asensio, J.; Perez, I.; Aguilo, N.; Uranga, S.; Pico, A.; Lampreave, C.; Cebollada, A.; Otal, I.; Samper, S.; Martin, C. New insights into the transposition mechanisms of IS6110 and its dynamic distribution between Mycobacterium tuberculosis Complex lineages. *PLoS Genet.* **2018**, *14*, e1007282. [CrossRef]
21. Delogu, G.; Sanguinetti, M.; Pusceddu, C.; Bua, A.; Brennan, M.J.; Zanetti, S.; Fadda, G. PE_PGRS proteins are differentially expressed by Mycobacterium tuberculosis in host tissues. *Microbes Infect.* **2006**, *8*, 2061–2067. [CrossRef]
22. Bettencourt, P.; Muller, J.; Nicastri, A.; Cantillon, D.; Madhavan, M.; Charles, P.D.; Fotso, C.B.; Wittenberg, R.; Bull, N.; Pinpathomrat, N.; et al. Identification of antigens presented by MHC for vaccines against tuberculosis. *NPJ Vaccines* **2020**, *5*, 2. [CrossRef]
23. Kruh, N.A.; Troudt, J.; Izzo, A.; Prenni, J.; Dobos, K.M. Portrait of a pathogen: The Mycobacterium tuberculosis proteome in vivo. *PLoS ONE* **2010**, *5*, e13938. [CrossRef] [PubMed]
24. Bitter, W.; Houben, E.N.; Luirink, J.; Appelmelk, B.J. Type VII secretion in mycobacteria: Classification in line with cell envelope structure. *Trends Microbiol.* **2009**, *17*, 337–338. [CrossRef] [PubMed]
25. Ocampo, M.; Curtidor, H.; Vanegas, M.; Patarroyo, M.A.; Patarroyo, M.E. Specific interaction between Mycobacterium tuberculosis lipoprotein-derived peptides and target cells inhibits mycobacterial entry in vitro. *Chem. Biol. Drug Des.* **2014**, *84*, 626–641. [CrossRef] [PubMed]
26. Stewart, G.R.; Wilkinson, K.A.; Newton, S.M.; Sullivan, S.M.; Neyrolles, O.; Wain, J.R.; Patel, J.; Pool, K.L.; Young, D.B.; Wilkinson, R.J. Effect of deletion or overexpression of the 19-kilodalton lipoprotein Rv3763 on the innate response to Mycobacterium tuberculosis. *Infect. Immun.* **2005**, *73*, 6831–6837. [CrossRef]
27. Noss, E.H.; Pai, R.K.; Sellati, T.J.; Radolf, J.D.; Belisle, J.; Golenbock, D.T.; Boom, W.H.; Harding, C.V. Toll-like receptor 2-dependent inhibition of macrophage class II MHC expression and antigen processing by 19-kDa lipoprotein of Mycobacterium tuberculosis. *J. Immunol.* **2001**, *167*, 910–918. [CrossRef]
28. Gehring, A.J.; Rojas, R.E.; Canaday, D.H.; Lakey, D.L.; Harding, C.V.; Boom, W.H. The Mycobacterium tuberculosis 19-kilodalton lipoprotein inhibits gamma interferon-regulated HLA-DR and Fc gamma R1 on human macrophages through Toll-like receptor 2. *Infect. Immun.* **2003**, *71*, 4487–4497. [CrossRef]
29. Jiang, Y.; Wei, J.; Liu, H.; Li, G.; Guo, Q.; Qiu, Y.; Zhao, L.; Li, M.; Zhao, X.; Dou, X.; et al. Polymorphisms in the PE35 and PPE68 antigens in Mycobacterium tuberculosis strains may affect strain virulence and reflect ongoing immune evasion. *Mol. Med. Rep.* **2016**, *13*, 947–954. [CrossRef]
30. Arnvig, K.B.; Comas, I.; Thomson, N.; Houghton, J.; Boshoff, H.I.; Croucher, N.; Rose, G.; Perkins, T.T.; Parkhill, J.; Dougan, G.; et al. Sequence-Based Analysis Uncovers an Abundance of Non-Coding RNA in the Total Transcriptome of Mycobacterium tuberculosis. *PLoS Pathog.* **2011**, *7*, e1002342. [CrossRef]
31. Ignatov, D.V.; Salina, E.G.; Fursov, M.V.; Skvortsov, T.A.; Azhikina, T.L.; Kaprelyants, A.S. Dormant non-culturable Mycobacterium tuberculosis retains stable low-abundant mRNA. *BMC Genom.* **2015**, *16*, 954. [CrossRef] [PubMed]
32. Pasula, R.; Downing, J.F.; Wright, J.R.; Kachel, D.L.; Davis, T.E., Jr.; Martin, W.J. Surfactant protein A (SP-A) mediates attachment of Mycobacterium tuberculosis to murine alveolar macrophages. *Am. J. Respir. Cell Mol. Biol.* **1997**, *17*, 209–217. [CrossRef] [PubMed]
33. Gaynor, C.D.; McCormack, F.; Voelker, D.R.; McGowan, S.; Schlesinger, L.S. Pulmonary surfactant protein A mediates enhanced phagocytosis of Mycobacterium tuberculosis by a direct interaction with human macrophages. *J. Immunol.* **1995**, *155*, 5343–5351.
34. Shepherd, V.L.; Lopez, J.P. The Role of Surfactant-Associated Protein A in Pulmonary Host Defense. *Immunol. Res.* **2001**, *23*, 111–120. [CrossRef]

35. Pasula, R.; Wright, J.R.; Kachel, D.L.; Martin, W.J. Surfactant protein A suppresses reactive nitrogen intermediates by alveolar macrophages in response to Mycobacterium tuberculosis. *J. Clin. Investig.* **1999**, *103*, 483–490. [CrossRef] [PubMed]
36. Shepelkova, G.; Pommerenke, C.; Alberts, R.; Geffers, R.; Evstifeev, V.; Apt, A.; Schughart, K.; Wilk, E. Analysis of the lung transcriptome in Mycobacterium tuberculosis-infected mice reveals major differences in immune response pathways between TB-susceptible and resistant hosts. *Tuberculosis* **2013**, *93*, 263–269. [CrossRef] [PubMed]
37. Shafiani, S.; Tucker-Heard, G.; Kariyone, A.; Takatsu, K.; Urdahl, K.B. Pathogen-specific regulatory T cells delay the arrival of effector T cells in the lung during early tuberculosis. *J. Exp. Med.* **2010**, *207*, 1409–1420. [CrossRef]
38. Ulrichs, T.; Kaufmann, S.H. New insights into the function of granulomas in human tuberculosis. *J. Pathol.* **2005**, *208*, 261–269. [CrossRef]
39. Malardo, T.; Gardinassi, L.G.; Moreira, B.P.; Padilha, É.; Lorenzi, J.C.C.; Soares, L.S.; Gembre, A.F.; Fontoura, I.C.; De Almeida, L.P.; Santos, I.K.F.d.M.; et al. MicroRNA expression signatures in lungs of mice infected with Mycobacterium tuberculosis. *Tuberculosis* **2016**, *101*, 151–159. [CrossRef]
40. Ghorpade, D.S.; Leyland, R.; Kurowska-Stolarska, M.; Patil, S.A.; Balaji, K.N. MicroRNA-155 Is Required for Mycobacterium bovis BCG-Mediated Apoptosis of Macrophages. *Mol. Cell. Biol.* **2012**, *32*, 2239–2253. [CrossRef]
41. Rajaram, M.; Ni, B.; Morris, J.D.; Brooks, M.N.; Carlson, T.K.; Bakthavachalu, B.; Schoenberg, D.; Torrelles, J.B.; Schlesinger, L.S. Mycobacterium tuberculosis lipomannan blocks TNF biosynthesis by regulating macrophage MAPK-activated protein kinase 2 (MK2) and microRNA miR-125b. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 17408–17413. [CrossRef]
42. Kumar, R.; Halder, P.; Sahu, S.K.; Kumar, M.; Kumari, M.; Jana, K.; Ghosh, Z.; Sharma, P.; Kundu, M.; Basu, J. Identification of a novel role of ESAT-6-dependent miR-155 induction during infection of macrophages withMycobacterium tuberculosis. *Cell. Microbiol.* **2012**, *14*, 1620–1631. [CrossRef]
43. Wang, J.; Yang, K.; Zhou, L.; Wu, M.; Wu, Y.; Zhu, M.; Lai, X.; Chen, T.; Feng, L.; Li, M.; et al. MicroRNA-155 Promotes Autophagy to Eliminate Intracellular Mycobacteria by Targeting Rheb. *PLoS Pathog.* **2013**, *9*, e1003697. [CrossRef]
44. Iwai, H.; Funatogawa, K.; Matsumura, K.; Kato-Miyazawa, M.; Kirikae, F.; Kiga, K.; Sasakawa, C.; Miyoshi-Akiyama, T.; Kirikae, T. MicroRNA-155 knockout mice are susceptible to Mycobacterium tuberculosis infection. *Tuberculosis* **2015**, *95*, 246–250. [CrossRef]

frontiers in Microbiology

ORIGINAL RESEARCH
published: 07 February 2017
doi: 10.3389/fmicb.2017.00128

Secretome Prediction of Two *M. tuberculosis* Clinical Isolates Reveals Their High Antigenic Density and Potential Drug Targets

Fernanda Cornejo-Granados*[1], *Zyanya L. Zatarain-Barrón*[2], *Vito A. Cantu-Robles*[1], *Alfredo Mendoza-Vargas*[3], *Camilo Molina-Romero*[4], *Filiberto Sánchez*[1], *Luis Del Pozo-Yauner*[5], *Rogelio Hernández-Pando*[2]* and *Adrián Ochoa-Leyva*[1]

[1] Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Mexico, [2] Experimental Pathology Laboratory, Department of Pathology, National Institute of Medical Science and Nutrition "Salvador Zubirán," Mexico City, Mexico, [3] Massive Sequencing Unit, National Institute of Genomic Medicine, Mexico City, Mexico, [4] Thoracic Oncology Unit, National Institute of Cancer, Mexico City, Mexico, [5] Laboratorio de Estructura de Proteínas, National Institute of Genomic Medicine, Mexico City, Mexico

OPEN ACCESS

Edited by:
Feng Gao,
Tianjin University, China

Reviewed by:
Prashant S. Phale,
IIT-Bombay, India
Marisol Ocampo,
Universidad del Rosario and Fundación Instituto de Inmunología de Colombia, Colombia

***Correspondence:**
Rogelio Hernández-Pando
rhdezpando@hotmail.com
Adrián Ochoa-Leyva
aochoa@ibt.unam.mx

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *01 November 2016*
Accepted: *18 January 2017*
Published: *07 February 2017*

Citation:
Cornejo-Granados F, Zatarain-Barrón ZL, Cantu-Robles VA, Mendoza-Vargas A, Molina-Romero C, Sánchez F, Del Pozo-Yauner L, Hernández-Pando R and Ochoa-Leyva A (2017) Secretome Prediction of Two M. tuberculosis Clinical Isolates Reveals Their High Antigenic Density and Potential Drug Targets. Front. Microbiol. 8:128. doi: 10.3389/fmicb.2017.00128

The Excreted/Secreted (ES) proteins play important roles during *Mycobacterium tuberculosis* invasion, virulence, and survival inside the host and they are a major source of immunogenic proteins. However, the molecular complexity of the bacillus cell wall has made difficult the experimental isolation of the total bacterial ES proteins. Here, we reported the genomes of two Beijing genotype *M. tuberculosis* clinical isolates obtained from patients from Vietnam (isolate 46) and South Africa (isolate 48). We developed a bioinformatics pipeline to predict their secretomes and observed that ∼12% of the genome-encoded proteins are ES, being PE, PE-PGRS, and PPE the most abundant protein domains. Additionally, the Gene Ontology, KEGG pathways and Enzyme Classes annotations supported the expected functions for the secretomes. The ∼70% of an experimental secretome compiled from literature was contained in our predicted secretomes, while only the 34–41% of the experimental secretome was contained in the two previously reported secretomes for H37Rv. These results suggest that our bioinformatics pipeline is better to predict a more complete set of ES proteins in *M. tuberculosis* genomes. The predicted ES proteins showed a significant higher antigenic density measured by Abundance of Antigenic Regions (AAR) value than the non-ES proteins and also compared to random constructed secretomes. Additionally, we predicted the secretomes for H37Rv, H37Ra, and two *M. bovis* BCG genomes. The antigenic density for BGG and for isolates 46 and 48 was higher than the observed for H37Rv and H37Ra secretomes. In addition, two sets of immunogenic proteins previously reported in patients with tuberculosis also showed a high antigenic density. Interestingly, mice infected with isolate 46 showed a significant lower survival rate than the ones infected with isolate 48 and both survival rates were lower than the one previously reported for the H37Rv in the same murine model. Finally, after a druggability analysis of the secretomes, we found potential drug targets such as cytochrome P450, thiol peroxidase, the Ag85C, and Ribonucleoside Reductase in the secreted proteins that could be used as drug targets for novel treatments against Tuberculosis.

Keywords: secretome, *Mycobacterium tuberculosis*, antigenic density, drug targets, AAR, clinical isolates, bioinformatics, immunogenicity

INTRODUCTION

Worldwide, *Mycobacterium tuberculosis* (*M. tuberculosis*) remains a highly prevalent pathogen. According to the WHO, there were 10.4 million new cases and 1.4 million deaths in 2015 (WHO, 2016). Additionally, 3.3% of the new cases and 20% of the previously treated ones correspond to multidrug-resistant (MDR) infections (WHO, 2016). Although the use and development of rapid molecular diagnostic tests like Xpert MTB/Rif® and GeneXpert Omni® has expanded, the development of new drugs and vaccines is necessary. Moreover, with the wide genetic variation within *M. tuberculosis* strains and the impact that this variability has on the clinical outcome (López et al., 2003; Pérez-Martínez et al., 2008), there is a great need to understand the molecular mechanisms leading from strain genotype to the clinical phenotype. The strain H37Rv is the most studied *M. tuberculosis* strain, and it is an important model for laboratory studies. Another important *M. tuberculosis* family of strains is the "Beijing" genotype, a member of Lineage 2 (East-Asia), which has caused great concern because of their enhanced virulence, their highly transmissible phenotypes, and their increasing prevalence worldwide (López et al., 2003).

The complete set of Excreted/Secreted (ES) proteins, which is often referred as the cell secretome, is involved in critical biological processes, like mechanisms of adhesion, cell migration, and invasion, cell-to-cell communication, signal transduction and potential infective strategies in disease mechanisms (Tjalsma et al., 2004). As a facultative intracellular pathogen, *M. tuberculosis* relies on its ability to survive within the host through the secretion of virulent proteins with the capacity to modulate a variety of host cellular pathways (Smith, 2003; Målen et al., 2007; Chande et al., 2015; Vargas-Romero et al., 2016). ES proteins are an important source of immunogenic proteins due to their ability to be recognized by the host immune system. They are also considered T-cell antigens that promote protective immune responses against *M. tuberculosis* (Daugelat et al., 1992; Målen et al., 2007; Zheng et al., 2013). This has led to focusing most vaccine and drug development efforts to the identification of mycobacterial secreted proteins.

Several experimental attempts have been made to determine the secretome of *M. tuberculosis* strains, using "traditional" techniques such as 2-D gel electrophoresis or based on "omics" approaches like liquid chromatography coupled with different types of MS analysis (Målen et al., 2007). However, the molecular complexity of the pathogen cell envelope, composed by mycolic acids, peptidoglycan, acyl lipids, etc., complicate the experimental analysis of ES proteins (Zhou et al., 2015). To address this limitation, bioinformatics methods can be used for the systematized prediction of ES proteins from available sequenced genomes (Gomez et al., 2015). In this regard, two predicted *M. tuberculosis* secretomes were previously reported using bioinformatics approaches. In one study, the genome of H37Rv was screened to predict their encoded ES proteins using several secretion predictors, resulting in a secretome of 825 proteins (Vizcaíno et al., 2010). However, only one protein from each predictor was selected and experimentally confirmed as secreted (Vizcaíno et al., 2010). In a second study, the authors reported a database composed of 276 secreted proteins for the H37Rv genome using different bioinformatics algorithms (Roy et al., 2013) and they found that 46 from 57 experimentally confirmed secreted proteins were predicted in their secretome (Roy et al., 2013). However, neither of the two studies provided annotation analysis, biochemical pathway mapping, protein domain content or antigenic potential of their *M. tuberculosis* predicted secretomes. Also, the two reported secretomes could still contain transmembrane proteins because the algorithms used in their ES predictions do not analyze this type of proteins, plus, success in their ES prediction was only evaluated against few experimentally secreted proteins. In the present study, we sequenced and assembled two genomes of *M. tuberculosis* clinical isolates members of the Beijing genotype and the total encoded proteins were independently analyzed to predict the ES proteins for each genome. The predicted ES proteins were then annotated regarding sequence similarity to other known proteins, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, gene ontologies (GO) and protein domains. Additionally, the antigenic density of the predicted secretomes was evaluated using the AAR value (Gomez et al., 2015). The secretomes for H37Rv, H37Ra and two BCG genomes were also predicted. Finally, a druggability analysis was also made for the predicted secretomes. We believe that our work could contribute to a better comprehension of the host-pathogen interactions in the *M. tuberculosis* infection.

METHODS

Bacterial Strains

We selected two *Mycobacterium tuberculosis* clinical isolates members of the Beijing genotype that we referred as isolates 46 and 48 in this manuscript. Prof D. van Soolingen at the National Institute of Public Health and the Environment (RIVM; Bilthoven, the Netherlands) kindly provided the isolate 46 that corresponds to the RIVM number 2002:1612 collected in Vietnam. This isolate is part of a tuberculosis isolates collection from a wide range of geographical origins. The isolate 48 corresponds to the previously described clinical strain code

1 reported in Tuberculosis patients attending primary health care clinics in the Western Cape Province of South Africa (Aguilar et al., 2010). This isolate was collected from the urban epidemiological field site in Cape Town during the period January 1993–December 2004 and kindly provided by Prof R. Warren from the Stellenbosh University (van der Spuy et al., 2009).

Sequencing and Assembly of *M. tuberculosis* Clinical Isolates

The bacterial genomic DNA (gDNA) of the 46 and 48 isolates was extracted from liquid cultures in logarithmic growth phase using the Quick-gDNA^{TM} MiniPrep kit after performing Zihel-Neelsen stains to assess the purity of *M. tuberculosis* cells. DNA libraries were constructed using the NebNext DNA Library protocol from Illumina (Cat. E6040S). Libraries were pair-end sequenced using Illumina GAIIx technology in the Unidad de Secuenciación Masiva from INMEGEN with a length of 72 bp per read and a sequencing depth of approximately 8 million paired reads per genome. The raw sequences were filtered using FastX-Toolkit and *de novo* assembled using Velvet (Zerbino and Birney, 2008). The resulting assemblies of each isolate are showed in Table S1. The final selected assemblies were analyzed with RAST (Aziz et al., 2008) to obtain the ORFs. We also extracted the ORFs from the H37Rv genome (GenBank: AL123456.3) using RAST to compare with the ones obtained for the clinical isolates.

Prediction of ES Proteins in *M. tuberculosis* Genomes

All the coding gene sequences were analyzed independently for each genome by the different feature-based tools indicated in **Figure 1**. SignalP 4.1 (Bendtsen et al., 2004) was used to predict classically secreted proteins (Sec-dependent), setting the option for prokaryote organisms and the positional limit of 70 residues for truncation and the rest of the parameters were set as default. SecretomeP 2.0 (Bendtsen et al., 2005a) was used to predict the non-classical secreted proteins selecting the default options for Gram-positive bacteria and all the resulting proteins with an N-N score $\geq$ 0.5 were considered as positives. TatP 1.0 (Bendtsen et al., 2005b) was used to determine the proteins secreted via the Tat pathway applying the default parameters and the resulting proteins with a Tat motif were considered as positives. Additionally, we used LipoP 1.0 (Juncker et al., 2003) to predict lipoprotein motifs in the first 70 amino acids of each sequence, for this program, all settings were set to default, and all the resulting proteins with a "cytoplasmic" prediction were removed. Finally, all the proteins considered as positive from each of the predictors were merged together and the resulting list was scanned by TMHMM 2.0 (Krogh et al., 2001). This tool allows the identification, localization, and orientation of transmembrane helices and all the proteins predicted with 0 transmembrane motifs were assigned directly as part of the secretome. The rest of the proteins (with $\geq$1 transmembrane motifs) were further analyzed with Phobius (Käll et al., 2007) to identify possible α-helical conformations in the N-terminal region of the proteins that belongs to a signal sequence and that could be mistakenly classified as a transmembrane region. If any of the analyzed proteins was predicted to have a signal sequence it was added to the list of ES proteins. The secretome of the *M. tuberculosis* H37Ra GenBank CP000611 and two *M. bovis* BCG (BCG Danish GenBank NZ_CUWH01000001 and BCG Pasteur GenBank AM408590) strains were also predicted using the same bioinformatics pipeline. For comparison, the proteins that are neither ES and transmembrane was defined as "intracellular proteins." Hence, the non-ES proteins consist of the transmembrane and the intracellular proteins.

Annotation and Comparative Analysis of ES Proteins

For identifying homolog proteins, ES proteins were analyzed using BLASTP against the non-redundant (nr) database using the Blast2GO (Conesa and Götz, 2008) with an E-value cut-off set at 1.0E^{-3} (Table S2). Both ES and non-ES proteins were functionally mapped to GO terms and annotated by setting the following parameters: *E*-value-hi-filter: 1.0E^{-3}; Annotation cut-off: 55; GO weight: 5 and Hsp-Hit Coverage cut-off:0. The ES proteins were also associated with protein families through InterProScan (Zdobnov and Apweiler, 2001). Blast2GO was then used to identify the over or under represented GO terms in the ES proteins, by setting the term filter *p*-value to $\leq$ 0.05. Additionally, the KAAS (Moriya et al., 2007) was used for mapping ES proteins to KEEG pathways using the BBH (bi-directional best hit) method to assign the representative genes data set and the orthologs for prokaryotes (Table S3). We also classified the enzymes according to the six enzymes commission classes using Blast2GO (Figure S1).

Construction of the Experimental Secretome

To validate the accuracy of our bioinformatics pipeline, we compared an experimental validated secretome (Figure S2) that we compiled from a literature search against our predicted secretomes. To construct the experimental secretome, we made a search at the NCBI database, and we retrieve all articles that experimentally reported excreted or secreted proteins for *M. tuberculosis*. After that, we ended up with 338 proteins that have been experimentally reported as secreted in different studies (Table S4). Then, we perform a BLASTP of the 338 proteins (*E*-value 1.0E^{-3}) against our predicted secretomes to assess how many experimental ES proteins matched with the predicted secretomes. Only the secreted proteins reported as markers for serodiagnosis by Zhou et al. (2015), were also included in the experimental secretome. To do this, we analyzed the complete set of Zhou et al. (2015) with TMHMM and only the proteins without transmembrane regions were included in our experimental secretome.

Calculation of the Abundance of Antigenic Regions (AAR)

The AAR is a value used to normalize the number of antigenic regions by the sequence length (Gomez et al., 2015). This value was calculated as the ratio between the sequence length and

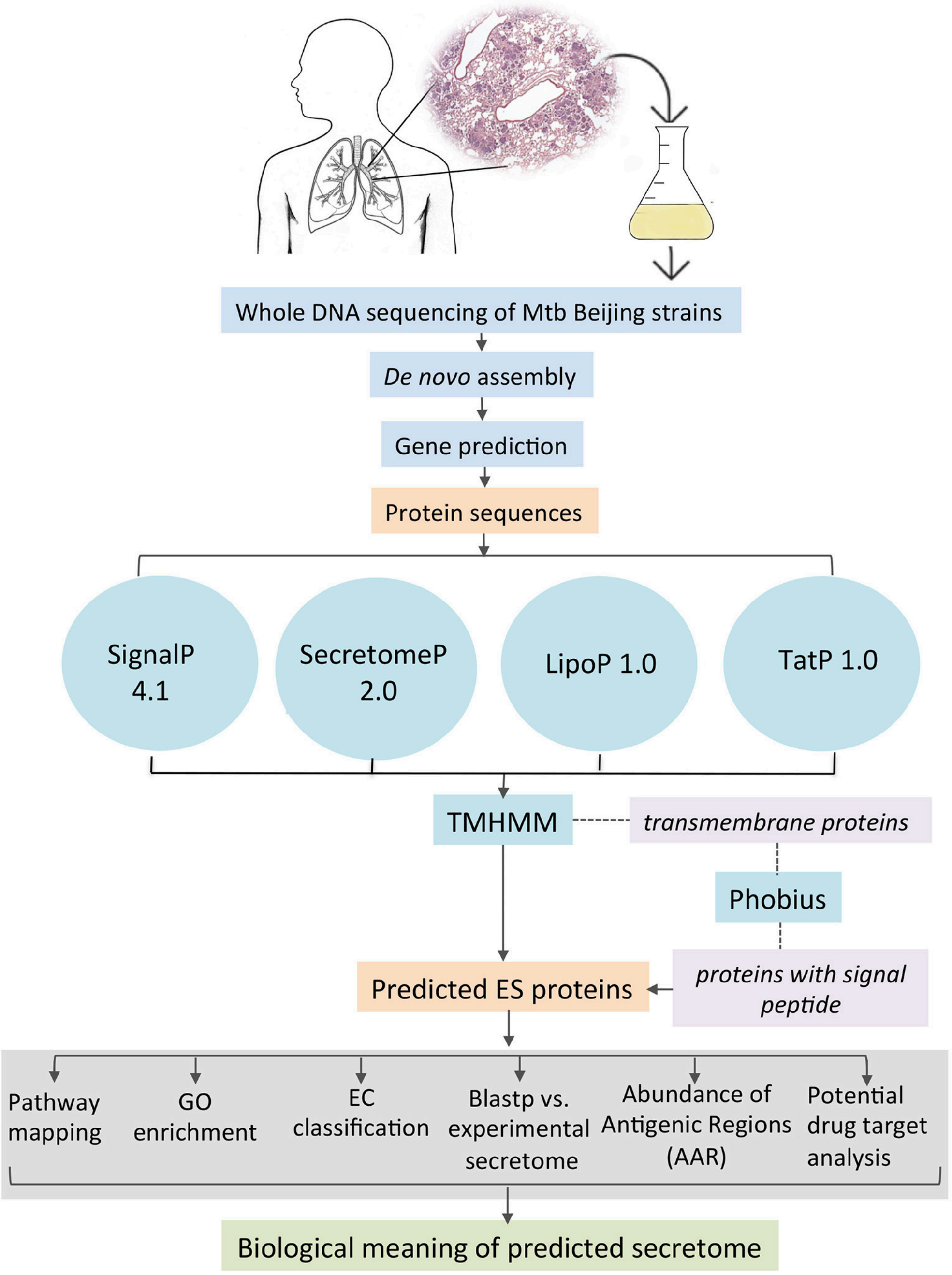

Whole DNA sequencing of Mtb Beijing strains
De novo assembly
Gene prediction
Protein sequences
SignalP 4.1
SecretomeP 2.0
LipoP 1.0
TatP 1.0
TMHMM
transmembrane proteins
Phobius
proteins with signal peptide
Predicted ES proteins
Pathway mapping
GO enrichment
EC classification
Blastp vs. experimental secretome
Abundance of Antigenic Regions (AAR)
Potential drug target analysis
Biological meaning of predicted secretome

FIGURE 1 | Bioinformatics pipeline to identify and annotate the secretomes of *M. tuberculosis* genomes.

the number of predicted antigenic regions for each protein and determines the number of amino acids that are needed to find one antigenic region within a protein sequence (Gomez et al., 2015). Hence, low AAR values mean that the protein has more antigenic regions (more antigenic density). We used the AAR value to evaluate the antigenic density of the different protein data sets. To this end, the number of antigenic regions for each protein sequences was obtained using BepiPred (Larsen et al., 2006) with the default settings (threshold 0.35) and normalized by sequence length (Gomez et al., 2015). The Mann-Whitney statistical test ($p < 0.001$) was used to establish if there is a significant difference between AAR values of protein data sets. To assess if the AAR observed in the predicted secretomes is significantly different to random constructed secretomes, we compared the AAR of the predicted secretomes to the AAR of 1000 protein datasets with 553, 519, and 548 randomly selected proteins from 46, 48, and H37Rv genomes, respectively. We then determined the AAR for each of the 1000 iterations and determined an empirical *p*-value by keeping track of the number of iterations equaled or exceeded the observed AAR for each corresponding secretome.

Potential Drug Target Analysis

We performed a BLASTP (E-value $1.0E^{-3}$) between the proteins of the 46, 48, and H37Rv secretomes to obtain the shared proteins in the three secretomes (core secretome). The resulting set of 449 shared proteins was further searched for sequence similarity against known drug targets available on the Drug Bank database (http://www.drugbank.ca/), setting the *E*-value to $1.0E^{-3}$ and the rest of the options to default. In Table S5 all the proteins that have similarity with a known drug target, as well as the drugs that can affect said target, are showed.

Survival and Drug Resistance Assays

The survival rate caused by the two clinical isolates was evaluated in 6- to 8-week-old male BALB/c mice as previously described (Hernandez-Pando et al., 1996). Briefly, two groups of 50 mice were each inoculated intratracheally with 2.5×10^5 bacilli of each of the two clinical isolates in 100 μL Phosphate-Buffered Saline (PBS) and survival rate was recorded since day 1–day 90 post-infection. The clinical isolates were also evaluated for drug resistance (rifampicin, ethambutol, streptomycin and isoniazid) with the BD BACTECH^{TM} MGIT^{TM} 960 Mycobacteria Culture System following the manufacturer's recommendations.

Nucleotide Sequence Accession Numbers

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accessions MSLU00000000 and MSLV00000000. The versions described in this paper are MSLU01000000 and MSLV01000000.

RESULTS

Genome Sequencing and Assembly

We sequenced the whole genome of the 46 and 48 *M. tuberculosis* clinical isolates, which were originated of patients from Vietnam and South Africa, respectively. Bacterial genomic DNA was extracted from liquid cultures and sequenced using Illumina GAIIx technology with a depth of approximately 8 million paired reads per genome. After genome assembly, we obtained 151 contigs for isolate 46 and 144 contigs for isolate 48 (Table S1) with approximately 71- and 72-fold genome coverage of a 4.3 Mb genome size, respectively. The Open Reading Frames (ORFs) were extracted for each genome, resulting in 4336 and 4310 proteins for isolate 46 and 48, respectively. Additionally, we also extracted the ORFs in the H37Rv genome to compare with the proteins of the clinical isolates.

Prediction of *M. tuberculosis* Secretomes

Proteins can be secreted through multiple secretory mechanisms. Thus, we utilized a combination of different bioinformatics tools based on Neuronal Network (NN), Hidden Markov Model (HMM), and Support Vector Machine (SVM) algorithms to predict the mycobacterial ES proteins encoded in our assembled genomes. To this end we utilized SignalP 4.1, SecretomeP 2.0, TatP 1.0, LipoP 1.0, followed by TMHMM 2.0 and Phobius (**Figure 1**). Such algorithms have a high performance in predicting signal peptides, protein subcellular localization, and transmembrane proteins. In addition, several of these algorithms have good performance to predict signal peptides in mycobacterial proteins (Restrepo-Montoya et al., 2009; Vizcaíno et al., 2010). SignalP and SecretomeP 2.0 were used to determine the classical and non-classical secreted proteins, respectively. Lipoproteins containing signal peptides were identified using LipoP 1.0 and proteins containing a twin-arginine signal peptide cleavage site were predicted using TatP 1.0. Then, the proteins predicted as secreted through the four algorithms were merged, yielding a set of 1956 different proteins for isolate 46, 1920 different proteins for isolate 48 and 1285 proteins for H37Rv. Next, each protein dataset was analyzed for the presence of transmembrane regions by TMHMM 2.0 algorithm. All the proteins containing transmembrane regions were removed, and the remaining proteins were considered the secretome for the genome 46, 48, and H37Rv. The proteins with transmembrane motifs that were removed from each isolate were re-analyzed with Phobius to identify proteins with a α-helical conformation in the amino section that could have been mistakenly classified as a transmembrane region (**Figure 1**). From Phobius analysis, only 1 protein for isolate 46 was predicted to have a signal peptide and therefore it was added to the secreted list of proteins. For isolate 48 and H37Rv no proteins were predicted to have a signal peptide by Phobius. Finally, the predicted secretome of isolates 46, 48, and H37Rv consisted of 553, 519, and 548 proteins, respectively, which represent ∼12% of the total proteins encoded in their *M. tuberculosis* genomes.

Annotation of *M. tuberculosis* Secretomes

Of the predicted secretomes, 448 proteins (81.01%) of isolate 46, 484 proteins (93.26%) of isolate 48, and 502 proteins (91.6%) of H37Rv strain showed significant similarity (BLASTP matches) with proteins deposited in the non-redundant (nr) database. Furthermore, according with the BLASTP match, most of the ES proteins for the three strains were identified as members of the PE, PPE, and PGRS families (Table S2) which are known as important in the bacterial virulence mechanisms (Brennan and Delogu, 2002). ES proteins were then annotated for Biological Process, Molecular Function, and Cellular Components with GO terms using Blast2GO. This resulted in 316 (57.14%) proteins for isolate 46, 283 (54.53%) proteins for isolate 48 and 309 (56.39%) proteins for H37Rv that were annotated with 304, 392, and 115 different GO terms, respectively. We analyzed whether any GO term showed a statistically significant over or under representation in the secretome as compared to the expected GO term distributions for the whole genome of each strain. For the three strains a significant over/under representation was only observed in the Cellular Component category. As expected, the most over represented GO terms in the secretomes were extracellular region, cell periphery and external encapsulating structure (**Figure 2**), which are the typical cellular components reported for secreted proteins (Gomez et al., 2015). While, GO

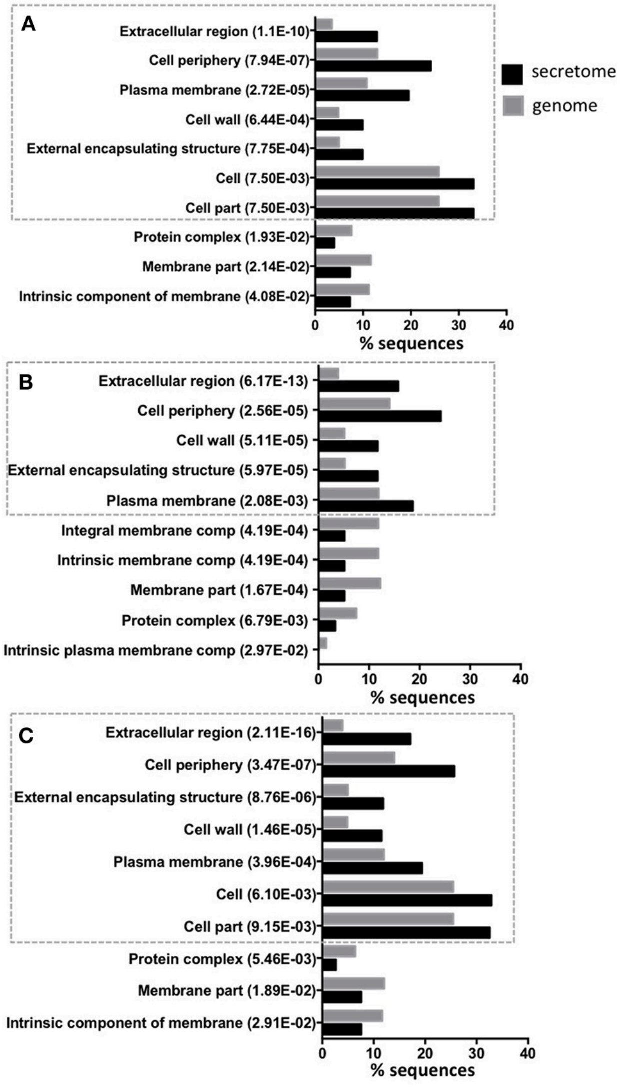


FIGURE 2 | Top 10 of significant over or under represented GO terms in predicted secretomes. The figure shows for each significant GO term in Cellular Component category the amount (percentage) of sequences annotated with this term in secretomes and genomes. The Y-axis shows significantly overrepresented (inside the ticked lines) or underrepresented GO terms in secretome and the X-axis gives the percentage of sequences of each term. The *p*-value in parentheses, shows the statistical assessment of over/under representation of GO term in secretomes when compared to genomes. Isolate 46 **(A)**, isolate 48 **(B)**, and reference strain H37Rv **(C)**.

terms that are non-related to cellular components of secreted proteins such as membrane part, protein complex or intrinsic membrane component were under represented in the secretome (**Figure 2**). We also used KAAS for mapping the ES proteins to KEGG pathways. After that, a total of 132 (23.9%), 117 (22.5%), and 121 (22.08%) ES proteins of isolates 46, 48 and H37Rv secretomes were mapped to 74, 75, and 76 KEGG pathways, respectively. The two most frequently mapped KEGG pathways for each secretome were: ABC transporters and pyrimidine metabolism. Additionally, we found two proteins involved in beta-Lactam resistance pathway (Table S3). The Tuberculosis KEGG pathway groups all human and *M. tuberculosis* proteins that have been involved in host-pathogen interactions and as expected five proteins were mapped to this pathway: lipoprotein LpqH, lipoprotein LprG, phosphate transport system substrate-binding protein pstS, acid phosphatase SapM, and the 6 kDa early secretory antigenic target ESAT-6 (Table S3). The full pathway annotations are available in Table S3.

We also classified the enzymes of the ES and non-ES proteins according to the six Enzyme Commission (EC) Classes. The results showed an overrepresentation of oxidoreductases, isomerases, and hydrolases in the ES proteins of isolate 46 as compared to the same enzyme types for the non-ES proteins (Figure S1). For isolate 48, there is an overrepresentation of the same kinds of proteins besides ligases (Figure S1B) while in strain H37Rv there is an overrepresentation of hydrolases, isomerases, oxidoreductases and ligases (Figure S1C). The annotation of protein domains contained in the secretomes was conducted using InterProScan and resulted in 274 protein domains for isolate 46, 234 for isolate 48 and 253 protein domains for H37Rv. The most represented protein domains are shown in **Table 1**. For the three secretomes, the most represented protein domains were PPE family C-terminal, and PE-PGRS family N-terminal. Interestingly, these protein domains are involved in the *M. tuberculosis* pathogenicity (Fishbein et al., 2015).

The Experimentally Reported ES Proteins Confirm the Accuracy of our Predicted Secretomes

To validate the accuracy of our bioinformatics pipeline to predict experimental secretomes, we compiled a protein dataset of 338 proteins (Table S4) experimentally reported as excreted/secreted in *M. tuberculosis* (see Methods) and determined how many proteins of this experimental secretome were also reported in our predicted secretomes. After that, we found that 227 (67.15%), 220 (65.09%), and 257 proteins (76.04%) of the experimental secretome were also contained in our predicted secretomes of isolates 46, 48, and H37Rv, respectively (Figure S2). These data indicates that ∼70% of the experimental secreted proteins were also included in our predicted secretomes, showing that our bioinformatics method is quite accurate. To asses, if the same number of experimental secreted proteins could be found in a list of randomly selected proteins, we constructed 1000 random secretomes consisting of groups of 553, 519, and 548 randomly selected proteins from the 46, 48, and H37Rv genomes, respectively and matched each random secretome against the experimental one. After that, we found that the maximum percentage of shared proteins obtained by a random secretome was of 40%, which is lower than the percentage obtained with our predicted secretomes (∼70%), indicating that our results are significantly different than random.

Additionally, we also matched the experimental secretome that we compiled from the literature search with the ones previously reported by Roy et al. (2013) and Vizcaíno et al. (2010). However, we found that only the 34.32 and 41.42% of the

TABLE 1 | Top 10 most represented protein domains in isolate 46, 48 and H37Rv secretomes.

InterPro code	InterPro description	Number of ES proteins (%)
ISOLATE 46		
IPR022171	PPE family C-terminal	18 (3.25)
IPR012338	Beta-lactamase/transpeptidase-like	18 (3.25)
IPR000084	PE-PGRS family N-terminal	17 (3.07)
IPR016040	NAD(P)-binding domain	15 (2.71)
IPR027417	P-loop containing nucleoside triphosphate hydrolase	14 (2.53)
IPR029058	Alpha/Beta hydrolase fold	12 (2.17)
IPR012336	Thioredoxin-like fold	11 (1.99)
IPR029063	S-adenosyl-L-methionine-dependent methyltransferase	11 (1.99)
IPR000253	Forkhead-associated (FHA) domain	8 (1.45)
IPR017853	Glycoside hydrolase superfamily	7 (1.27)
ISOLATE 48		
IPR000084	PE-PGRS family N-terminal	29 (5.59)
IPR012338	Beta-lactamase/transpeptidase-like	20 (3.85)
IPR022171	PPE family C-terminal	20 (3.85)
IPR029058	Alpha/Beta hydrolase fold	18 (3.47)
IPR016040	NAD(P)-binding domain	15 (2.89)
IPR012336	Thioredoxin-like fold	14 (2.7)
IPR027417	P-loop containing nucleoside triphosphate hydrolase	14 (2.7)
IPR029063	S-adenosyl-L-methionine-dependent methyltransferase	9 (1.73)
IPR000253	Forkhead-associated (FHA) domain	8 (1.54)
IPR001763	Rhodanese-like domain	6 (1.16)
REFERENCE STRAIN H37Rv		
IPR000084	PE-PGRS family, N-terminal	68 (12.41)
IPR022171	PPE family, C-terminal	22 (4.01)
IPR016040	NAD(P)-binding domain	14 (2.55)
IPR029058	Alpha/Beta hydrolase fold	9 (1.64)
IPR027417	P-loop containing nucleoside triphosphate hydrolase	8 (1.46)
IPR012338	Beta-lactamase/transpeptidase-like	8 (1.46)
IPR012336	Thioredoxin-like fold	5 (0.91)
IPR007312	Phosphoesterase	5 (0.91)
IPR026954	PknH-like extracellular domain	4 (0.73)
IPR005490	L,D-transpeptidase catalytic domain	4 (0.73)

experimental secretome was shared with the secretome reported by Roy et al. (2013) and Vizcaíno et al. (2010), respectively (Figure S2). These results suggest that the bioinformatics pipeline we used is better to predict a complete set of excreted/secreted proteins in *M. tuberculosis* genomes.

The AAR Value Reveals a High Antigenic Density in the Predicted and Experimental Secretomes

The AAR was used to calculate the antigenic density of the ES, intracellular, transmembrane and non-ES proteins of the isolates 46 and 48 and the H37Rv genomes. We found that the ES proteins have significantly more antigenic density (non-parametric Mann-Whitney test $p \leq 0.01$) than the rest of the proteins encoded in the *M. tuberculosis* genomes (**Figure 3** and **Table 2**). Interestingly, the secretomes of both clinical isolates had more antigenic density (AAR = ∼37.5) than the H37Rv secretome (AAR = 40.6) (**Table 2**). However, no significant difference between each secretome after performing the Mann-Whitney test was observed. This result suggests that no strain seems to have a more antigenic secretome than the other even there was a tendency to more antigenic secretomes for the clinical isolates respect to H37Rv.

To validate the biological significance of the high antigenic density (lower AAR values) observed in our secretomes, we also calculated the antigenic density for the experimental secretome obtained from the literature search (**Table 2**). Interestingly, the antigenic density for the experimental secretome was similar to the one obtained for the predicted secretomes (**Table 2**). It has been reported that some parasite secretomes have more antigenic density that non-secreted or transmembrane proteins (Gomez et al., 2015; Wang et al., 2015). So, we also analyzed if the high antigenic density is exclusive of the *M. tuberculosis* secretomes or if any set of similar number of proteins could obtain the same AAR value. To this end, we selected 1000 groups of 553, 519, and 548 randomly selected proteins from the 46, 48, and H37Rv genomes, respectively, and calculated the AAR for each group (see Methods). After that, we found that all the predicted secretomes had more significantly antigenic density ($p < 0.005$) than the randomly constructed ones. Hence, the high antigenic density obtained for the predicted secretomes is exclusive of that combination of secreted proteins. In addition, to test whether the antigenic density of our predicted secretomes was similar to other *M. tuberculosis* secretomes, we applied our bioinformatics pipeline to obtain the secretome of the H37Ra, BCG Danish, and BCG Pasteur genomes and calculated their AAR values (**Table 3**). We selected the genome of H37Ra because it is an attenuated strain closely related to the virulent H37Rv strain. We also selected two substrains of the *M. bovis* Bacille Calmette-Guérin (BCG) strain because it is the bacteria used in the Tuberculosis vaccine. Interestingly, the antigenic density for the BCG secretomes was very similar to the ones obtained for our clinical isolates (**Table 3**). However, the AAR values of the clinical isolates still show a lower tendency to have more antigenic density, followed by BCG and H37Ra and H37Rv strains (**Table 3**).

Taking the advantage of the fact that there is a lot of information about immunogenic proteins in Tuberculosis, we investigated if these immunogenic proteins have a high antigenic density. To this end, we selected two sets of proteins causing seropositive reactions in serum samples of Tuberculosis patients and determined their AAR values. The first set contains 57 secreted proteins resulted from a screening with 10 Tb patient and 3 healthy serum samples used as negative controls (Zhou et al., 2015). The second set contains 12 proteins characterized as serum biomarkers that can differentiate between both TB patients with active disease or recovered individuals; this second set was obtained from 189 patients (Deng et al., 2014). Interestingly, the AAR values were 38.5 for the former and 37.8 for the later protein

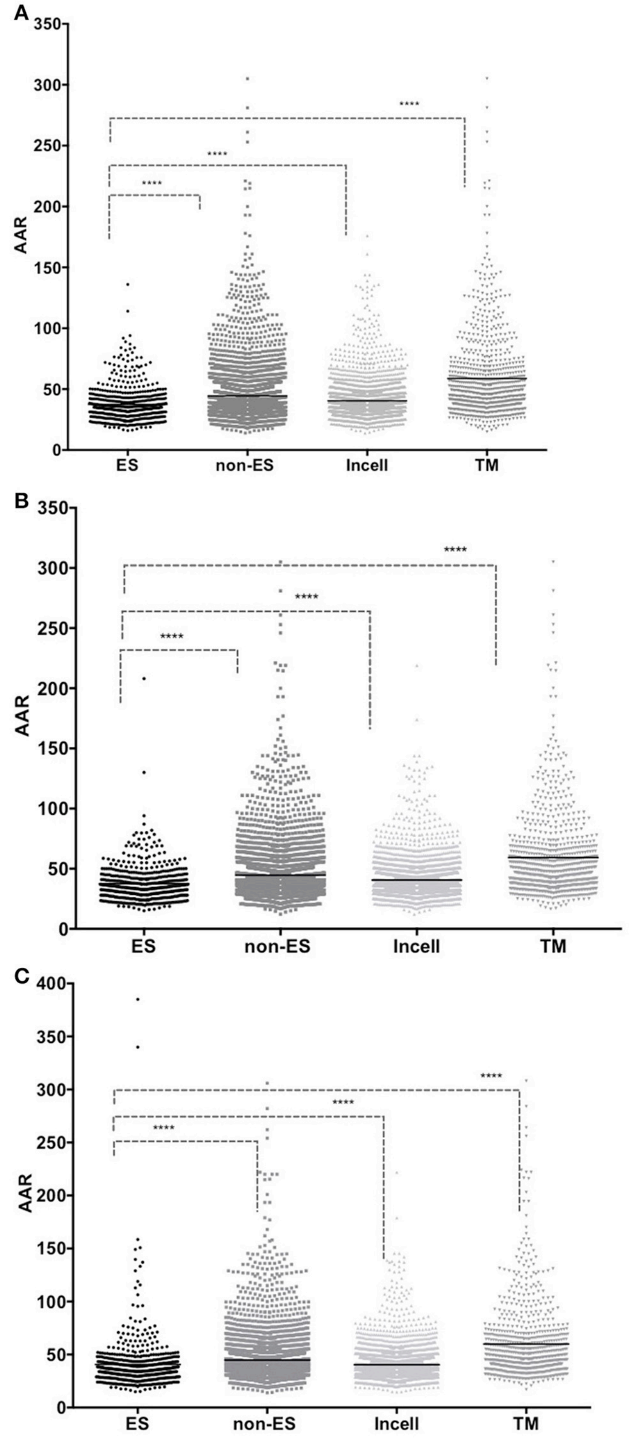


FIGURE 3 | Comparison between AAR values for Excreted/Secreted (ES), non Excreted/Secreted (non-ES) intracellular (Incell) and transmembrane (TM) proteins. The number of antigenic regions was calculated using the BepiPred method. The X-axis shows the protein group analyzed and the Y-axis shows the AAR obtained for each protein in that dataset. Low AAR values means that a protein has more antigenic density. The AAR comparisons between the proteins of isolate 46 **(A)**; isolate 48 **(B)** and H37Rv **(C)** are illustrated. A Mann-Whitney test was performed to compare the AAR within each group with a confidence level of 99% ($p \leq 0.01$), (***$p < 0.001$).

TABLE 2 | Abundance of Antigenic Regions (AAR) for *M. tuberculosis* predicted and experimental secretomes.

	ES Proteins		Non-ES proteins	
***M. tuberculosis* strain**	**Number of proteins in the dataset**	**AAR average**	**Number of proteins in the dataset**	**AAR average**
Beijing isolate 46	553	37.52	3702	44.54
Beijing isolate 48	519	37.55	3743	44.56
H37Rv reference strain	548	40.63	3788	44.74
Experimental secretome	338	38.99	-	-

TABLE 3 | Abundance of Antigenic Regions (AAR) for *M. tuberculosis* strains from different lineages.

	ES Proteins	
***M. tuberculosis* strain**	**Number of proteins in the dataset**	**AAR average**
Beijing isolate 46	553	37.52
Beijing isolate 48	519	37.55
H37Rv reference strain	548	40.63
H37Ra	554	40.52
BCG Danish	526	38.99
BCG Pasteur	564	38.89

dataset, suggesting that both sets of immunogenic proteins are composed of proteins with a high antigenic density.

Predicted Secretomes Suggest Novel Drug Targets

After a BLASTP (E-value 1.0E^{-3}) comparison between the 46, 48, and H37Rv secretomes we found that 449 proteins were shared between the three secretomes. This set of proteins was named the "*M. tuberculosis* core secretome" and we compared their sequence similarity against known drug targets available on the DrugBank database to determine if some ES proteins could be used as potential drug targets in the host-pathogen interactions (see Methods). Of the 449 ES proteins, only 26 showed homology with 91 known drug targets (Table S5). Notably, of all possible drug targets, only a few have a known inhibitor activity.

Survival and Drug Resistance Assays

We tested the survival rate of the 46 and 48 clinical isolates in a murine model (see Methods). Mice infected with isolate 46 showed the lowest survival rate, reaching 0% at day 48, while mice infected with isolate 48 survived until day 82 (**Figure 4**). After performing a log-rank (Mantel-Cox) test a significant difference was observed between the two survival rates with a p value ≤0.0001. Additionally, the two isolates also showed sensibility to rifampicin and ethambutol and resistance to streptomycin and isoniazid (see Methods).

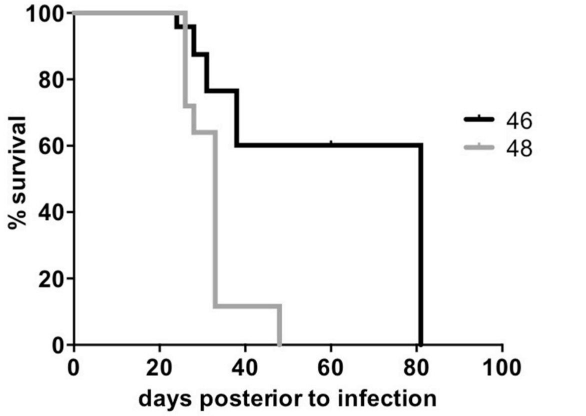


FIGURE 4 | Survival rate of Beijing clinical isolates. The two *M. tuberculosis* Beijing clinical isolates were used to infect BALB/c mice and record the survival rate. The X-axis shows the days post-infection and the Y-axis shows the percent of survival. A Log-rank (Mantel-cox) test was performed to compare the survival percentages within each group.

DISCUSSION

To the best of our knowledge, this study contains the most comprehensive *in silico* and experimental collection of *M. tuberculosis* secretomes and it is the first one that takes into account the secretome analysis of clinical isolates. Our results showed that our bioinformatics pipeline is quite accurate to predict a complete set of the secreted/excreted proteins in *M. tuberculosis* genomes. In this regard, our data indicates that ~70% of the experimental secretome was also predicted as secreted using our bioinformatics approach, while only the 34.32 and 41.42 % of the experimental secretome was found in the secretomes reported by Roy et al. (2013) and Vizcaíno et al. (2010), respectively. Furthermore, the maximum coincidence of the experimental secretome against 1000 randomly constructed secretomes was of 45.3%, indicating that our predicted secretomes are also significantly different to the random.

The predicted ES proteins for the three *M. tuberculosis* genomes represented around ~12% of the total genome proteins. Interestingly, this value is twice the percentage reported for the secretomes of parasite organisms like tapeworms (Gomez et al., 2015; Wang et al., 2015). We suggest that this difference between the percentages of secreted proteins among parasites could be associated with the fact that parasites such as *T. solium* are extracellular pathogens while *M. tuberculosis* is an intracellular parasite that requires different invasion mechanisms with their host. The annotation of the secretomes showed an enrichment of the antigenic protein families such as the PPE and PE-PGRS (**Table 1** and Table S2). It has been observed that these proteins may play a role in evasion of host immune responses, possibly via antigenic variation (Chaitra et al., 2008; Sampson, 2011; Akhter et al., 2012). As expected, several immunodominant antigens widely known for *M. tuberculosis* such as ESAT-6, Ag85C, CFP21, and CFP-10 (Silva et al., 2003; Wang et al., 2005) and several proteins of the Tuberculosis KEGG pathway (Table S3) were also present in our predicted secretomes. The overrepresentation of hydrolases and oxidoreductases in our predicted secretomes (Figure S1) is in agreement with the enrichment of this enzyme types reported in experimental *M. tuberculosis* secretomes (Målen et al., 2007). It is recognized that T-cells primarily mediate the immune response against an intracellular pathogen like *M. tuberculosis*. However, secreted and transmembrane proteins have also been identified to be targeted by B-cells in other intracellular bacteria like *Listeria* and *Chlamydia* (Grenningloh et al., 1997; Bannantine et al., 2000). In this regard, we chose BepiPred algorithm to analyze the antigenic density of the ES proteins, which predicts linear B-cell epitopes using Hidden Markov Models (Larsen et al., 2006). Interestingly, the predicted secretomes showed a significant higher antigenic density than the non-ES proteins (**Figure 3**). Additionally, a high antigenic density was also observed for the experimental secretome (**Table 2**). These results are in agreement with the high antigenic density reported for secretomes of 14 helminth species, including the human parasite *T. solium* (Gomez et al., 2015).

The antigenic density observed for the secretomes of the avirulent H37Ra and the virulent H37Rv strains was very similar, suggesting that the antigenic density is not associated with the virulence in these two strains. In fact, the avirulent phenotype of H37Ra is mainly associated to the loss of a secretion system (Zheng et al., 2008). The antigenic density between the two *M. b*ovis BCG strains was very similar (AAR= ~39) but it was higher than the antigenic density of the H37Rv (AAR= 40.6). In this case, the low virulence observed for the BCG strains is mainly attributed to the loss of RD1 locus, affecting the protein secretion pathway and the loss of cytolytic activity mediated by secreted ESAT-6, leading to reduced tissue invasiveness (Millington et al., 2011). However, the tendency to more antigenic density in the secretomes of BCG as compared to the H37Rv suggests why BCG strain has been the only one used as a vaccine so far. The higher antigenic density of all analyzed secretomes was the observed for the isolates 46 and 48 (AAR= ~37.5). This high antigenic density could be associated to the strong and sustained antigen stimulation for the granuloma formation and promoting the cell necrosis observed for Beijing strains (Flynn, 2004). Recently, it was suggested that *M. tuberculosis* uses mechanisms other than antigenic variation to evade T cells, indicating that antigenic variation is not a major mechanism of immune evasion in this pathogen (Coscolla et al., 2015). Interestingly, our data suggest that increasing the antigenic density of their secretomes could be one of the mechanisms associated with hypervirulent phenotypes of Beijing strains. In this regard, our survival assays performed in the tuberculosis murine model showed that isolates 46 and 48 have significantly lower survival rate than the one reported for the H37Rv (Hernandez-Pando et al., 1996). However, the antigenic density between two clinical isolates was very similar, AAR = 37.52 for isolate 46 and AAR = 37.55 for isolate 48, suggesting that the antigenic density is not the only mechanism responsible of the hypervirulent phenotype of these two Beijing strains.

The clinical isolates 46 and 48 showed a resistance to streptomycin and isoniazid, which are two of the first-line antibiotics used as treatment against Tuberculosis. It is necessary to explore not only the drugs acting at the intracellular level in *M. tuberculosis*, but also the drugs acting on proteins associated to the host-pathogen interactions. In this regard, it has been reported that several *M. tuberculosis* ES proteins interact with host cellular proteins to establish a successful infection modulating the host immune responses (Sreejit et al., 2014). The identification of this type of ES proteins may help us to design drugs against the host-pathogen interactions. For example, some virulence blockers are used to inhibit the secreted toxins of pathogens such as *Bacillus anthracis*, and *Clostridium tetani* (Moayeri et al., 2006; Clatworthy et al., 2007). Similarly, some studies have examined the potential of inhibiting extracellular molecules that participate in quorum sensing and help microorganisms such *Pseudomonas aeruginosa* in the biofilms formation (Duncan et al., 2012). Thus, the inhibition of ES proteins that are important for successful *M. tuberculosis* infection via disruption of host-pathogen interactions could help us to establish new opportunities for treatments against Tuberculosis.

In the core secretome we found homologous to known drug targets (Table S5), opening the possibility that known drugs could be used against ES proteins of *M. tuberculosis*. The list of drug targets includes the Ribonucleoside Reductase a homolog target used in cancer chemotherapy and several drugs including gallium nitrate and imexon were found to target this enzyme (Table S5). The gallium nitrate inhibits the activity of the Ribonucleoside Reductase and this drug has proven a high efficacy to treat Tuberculosis in murine models (Olakanmi et al., 2013). While imexon increases oxidative stress in target cells but also inhibits the Ribonucleoside Reductase (Roman et al., 2011). Nonetheless, there are not studies using this drug to Tuberculosis treatment. In basis on the above mentioned results, we suggest that several drugs (Table S5) usually used for cancer therapy, such as imexon and motexafin gadolinium could be explored as potential novel treatments against Tuberculosis through the modulation of the Ribonucleoside Reductase enzyme activity. However, it is important to mention that the human Ribonucleotide Reductase is also target of these cancer therapy drugs and secondary effects of a treatment using these drugs should be also evaluated. We also found several drugs that are in experimental phase (Table S5). Hence, their pharmacological action on the target protein is unknown. Some of them are: S-Oxy Cysteine, vitamin A, Pegvisomant, and Sofalcone which are target of the thiol peroxidase, short-chain dehydrogenase, cytochrome P450 (CYP139) and short chain dehydrogenase, respectively. Sofalcone suppress the production of NO and TNF in macrophages *in vitro* (Tanaka et al., 2009). Interestingly, these two citokines are responsible for the chronic inflammation and pneumonia in the late stages of *M. tuberculosis* infection. The CYP139 is immediately downstream of three polyketide synthase genes (pks17, 9, and 11) and upstream of genes encoding an ATP-binding ABC transporter, which is likely involved in the carriage (probably export) of macrolide molecules across the membrane (McLean and Munro, 2008). Hence, the CYP139 could be involved in the modification of these polyketide molecules prior to the transport process and it could be associated with the host-pathogen interactions.

The Ag85C is an essential protein for the cell wall synthesis in *M. tuberculosis* (Gobec et al., 2004). Recently, it was reported that Ebselen inhibits the growth of drug-resistant strains inhibiting the Ag85 complex (Favrot et al., 2013), demonstrating that it is an important target for the development of novel anti-Tuberculosis agents (Gobec et al., 2004). Thus, the two drugs reported in our analysis that also target the Ag85C (Table S5) could also be used to explore their effect on *M. tuberculosis*. Additionally, we also found a Class A beta-lactamase, which is a potential target for the Avibactam drug. Interestingly, it has been recently reported that clinical isolates of *M. tuberculosis* are susceptible to β-lactam/β–lactamase inhibitor combinations (Cohen et al., 2016). However, to date only the β-Lactamase inhibition by Avibactam has been proved in *M. abscessus* (Dubée et al., 2015). The Avibactam has no useful intrinsic antibacterial activity *per se*, it shows good results when it is combined with a β-lactam antibacterial as ceftazidime such as in the Avycaz which was recently used for treatment of intra-abdominal and urinary tract infections, including acute pyelonephritis, and it also has been used against carbapenemase-producing Enterobacteriaceae (Lucasti et al., 2013). These studies suggest that Avibactam in combination with other β-lactam antibiotics, such as Avycaz could be used against Tuberculosis.

The combination of SignalP, SecretomeP, TatP, and LipoP to predict the secreted proteins gives us the advantage to obtain proteins that were secreted through different secretion mechanisms and the use of TMHMM allows the elimination of secreted proteins containing transmembrane regions. We suggest that the combination of bioinformatics tools we designed allowed us a good match between experimental and predicted secretomes (Figure S2). For example, the secretome reported by Vizcaíno et al. (2010) also used the same 4 predictors that we used; however, they do not eliminate the transmembrane proteins in their secretome. While the secretome reported by Roy et al. (2013), only utilized the secreted proteins by SignalP followed by TATFIND1.4, PRED-LIPO, and the application of TMHMM analysis but only in a selected subgroup of proteins. Our bioinformatics pipeline could be useful to predict secretomes in other pathogen bacteria genomes. Undoubtedly, it would be ideal to support the secretome analysis with RNAseq data allowing us the identification of ES proteins that are differentially expressed during *M. tuberculosis* infection. For example, only the 41% of the *Taenia solium* secretome was reported as expressed (Gomez et al., 2015) while the 91% of the *E. multilocularis* secretome was differentially expressed between the parasite life-cycle stages (Wang et al., 2015). Our study contributes to increase the knowledge of the molecular mechanisms of host-pathogen interactions and we demonstrated how the ES proteins could be novel therapeutic targets against Tuberculosis using known drugs. Finally, a web server to calculate the AAR from protein datasets is under construction.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: FC, VC, AM, AO. Performed the experiments: FC, ZZ, VC, AM, CM, FS, AO. Analyzed the data: FC, VC, AO. Contributed reagents/materials/analysis tools: CM, LD, RH, AO. Wrote the paper: FC, ZZ, VC, AM, CM, FS, LD, RH, AO.

FUNDING

We acknowledge the support provided by CONACyT grants CB- 2013-223279 and SALUD-2014-C01-234188 and CONACyT contract Fon.Inst./58/2016.

ACKNOWLEDGMENTS

We thank Juan Manuel Hurtado Ramírez for informatics technical support and Unidad de Secuenciación Masiva from INMEGEN for sequencing technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.00128/full#supplementary-material

REFERENCES

Aguilar, D., Hanekom, M., Mata, D., Gey van Pittius, N. C., van Helden, P. D., Warren, R. M., et al. (2010). *Mycobacterium tuberculosis* strains with the Beijing genotype demonstrate variability in virulence associated with transmission. *Tuberculosis (Edinb).* 90, 319–325. doi: 10.1016/j.tube.2010.08.004

Akhter, Y., Ehebauer, M. T., Mukhopadhyay, S., and Hasnain, S. E. (2012). The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: perhaps more? *Biochimie* 94, 110–116. doi: 10.1016/j.biochi.2011.09.026

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bannantine, J. P., Griffiths, R. S., Viratyosin, W., Brown, W. J., and Rockey, D. D. (2000). A secondary structure motif predictive of protein localization to the chlamydial inclusion membrane. *Cell. Microbiol.* 2, 35–47. doi: 10.1046/j.1462-5822.2000.00029.x

Bendtsen, J. D., Kiemer, L., Fausbøll, A., and Brunak, S. (2005a). Non-classical protein secretion in bacteria. *BMC Microbiol.* 5:58. doi: 10.1186/1471-2180-5-58

Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: signalP 3.0. *J. Mol. Biol.* 340, 783–795. doi: 10.1016/j.jmb.2004.05.028

Bendtsen, J. D., Nielsen, H., Widdick, D., Palmer, T., and Brunak, S. (2005b). Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 6:167. doi: 10.1186/1471-2105-6-167

Brennan, M. J., and Delogu, G. (2002). The PE multigene family: a "molecular mantra" for *mycobacteria*. *Trends Microbiol.* 10, 246–249. doi: 10.1016/S0966-842X(02)02335-1

Chaitra, M. G., Shaila, M. S., and Nayak, R. (2008). Characterization of T-cell immunogenicity of two PE/PPE proteins of *Mycobacterium tuberculosis*. *J. Med. Microbiol.* 57, 1079–1086. doi: 10.1099/jmm.0.47565-0

Chande, A. G., Siddiqui, Z., Midha, M. K., Sirohi, V., Ravichandran, S., and Rao, K. V. (2015). Selective enrichment of mycobacterial proteins from infected host macrophages. *Sci. Rep.* 5:13430. doi: 10.1038/srep13430

Clatworthy, A. E., Pierson, E., and Hung, D. T. (2007). Targeting virulence: a new paradigm for antimicrobial therapy. *Nat. Chem. Biol.* 3, 541–548. doi: 10.1038/nchembio.2007.24

Cohen, K. A., El-Hay, T., Wyres, K. L., Weissbrod, O., Munsamy, V., Yanover, C., et al. (2016). Paradoxical hypersusceptibility of drug-resistant mycobacteriumtuberculosis to β-lactam antibiotics. *EBioMedicine* 9, 170–179. doi: 10.1016/j.ebiom.2016.05.041

Conesa, A., and Götz, S. (2008). Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* 2008:619832. doi: 10.1155/2008/619832

Coscolla, M., Copin, R., Sutherland, J., Gehre, F., de Jong, B., Owolabi, O., et al. (2015). *M. tuberculosis* T cell epitope analysis reveals paucity of antigenic variation and identifies rare variable TB antigens. *Cell Host Microbe* 18, 538–548. doi: 10.1016/j.chom.2015.10.008

Daugelat, S., Gulle, H., Schoel, B., and Kaufmann, S. H. (1992). Secreted antigens of *Mycobacterium tuberculosis*: characterization with T lymphocytes from patients and contacts after two-dimensional separation. *J. Infect. Dis.* 166, 186–190. doi: 10.1093/infdis/166.1.186

Deng, J., Bi, L., Zhou, L., Guo, S.-J., Fleming, J., Jiang, H.-W., et al. (2014). *Mycobacterium tuberculosis* proteome microarray for global studies of protein function and immunogenicity. *Cell Rep.* 9, 2317–2329. doi: 10.1016/j.celrep.2014.11.023

Dubée, V., Bernut, A., Cortes, M., Lesne, T., Dorchene, D., Lefebvre, A.-L., et al. (2015). β-lactamase inhibition by avibactam in *Mycobacterium abscessus*. *J. Antimicrob. Chemother.* 70, 1051–1058. doi: 10.1093/jac/dku510

Duncan, M. C., Linington, R. G., and Auerbuch, V. (2012). Chemical inhibitors of the type three secretion system: disarming bacterial pathogens. *Antimicrob. Agents Chemother.* 56, 5433–5441. doi: 10.1128/AAC.00975-12

Favrot, L., Grzegorzewicz, A. E., Lajiness, D. H., Marvin, R. K., Boucau, J., Isailovic, D., et al. (2013). Mechanism of inhibition of *Mycobacterium tuberculosis* antigen 85 by ebselen. *Nat. Commun.* 4, 2748. doi: 10.1038/ncomms3748

Fishbein, S., van Wyk, N., Warren, R. M., and Sampson, S. L. (2015). Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Mol. Microbiol.* 96, 901–916. doi: 10.1111/mmi.12981

Flynn, J. L. (2004). Mutual attraction: does it benefit the host or the bug? *Nat. Immunol.* 5, 778–779. doi: 10.1038/ni0804-778

Gobec, S., Plantan, I., Mravljak, J., Wilson, R. A., Besra, G. S., and Kikelj, D. (2004). Phosphonate inhibitors of antigen 85C, a crucial enzyme involved in the biosynthesis of the *Mycobacterium tuberculosis* cell wall. *Bioorg. Med. Chem. Lett.* 14, 3559–3562. doi: 10.1016/j.bmcl.2004.04.052

Gomez, S., Adalid-Peralta, L., Palafox-Fonseca, H., Cantu-Robles, V. A., Soberón, X., Sciutto, E., et al. (2015). Genome analysis of excretory/secretory proteins in *Taenia solium* reveals their Abundance of Antigenic Regions (AAR). *Sci. Rep.* 5:9683. doi: 10.1038/srep09683

Grenningloh, R., Darji, A., Wehland, J., Chakraborty, T., and Weiss, S. (1997). Listeriolysin and IrpA are major protein targets of the human humoral response against Listeria monocytogenes. *Infect. Immun.* 65, 3976–3980.

Hernandez-Pando, R., Orozco Estévez, H., Sampieri, A., and Pavón, L. (1996). Correlation between the kinetics of Thl/Th2 cells and pathology in a murine model of experimental pulmonary tuberculosis. *Immunology* 89, 26–33.

Juncker, A. S., Willenbrock, H., von Heijne, G., Brunak, S., Nielsen, H., and Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* 12, 1652–1662. doi: 10.1110/ps.0303703

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res.* 35, W429–W432. doi: 10.1093/nar/gkm256

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Larsen, J. E., Lund, O., and Nielsen, M. (2006). Improved method for predicting linear B-cell epitopes. *Immunome Res.* 2:2. doi: 10.1186/1745-7580-2-2

López, B., Aguilar, D., Orozco, H., Burger, M., Espitia, C., Ritacco, V., et al. (2003). A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clin. Exp. Immunol.* 133, 30–37. doi: 10.1046/j.1365-2249.2003.02171.x

Lucasti, C., Popescu, I., Ramesh, M. K., Lipka, J., and Sable, C. (2013). Comparative study of the efficacy and safety of ceftazidime/avibactam plus metronidazole versus meropenem in the treatment of complicated intra-abdominal infections in hospitalized adults: results of a randomized, double-blind, Phase II trial. *J. Antimicrob. Chemother.* 68, 1183–1192. doi: 10.1093/jac/dks523

Målen, H., Berven, F. S., Fladmark, K. E., and Wiker, H. G. (2007). Comprehensive analysis of exported proteins from *Mycobacterium tuberculosis* H37Rv. *Proteomics* 7, 1702–1718. doi: 10.1002/pmic.200600853

McLean, K. J., and Munro, A. W. (2008). Structural biology and biochemistry of cytochrome P450 systems in *Mycobacterium tuberculosis*. *Drug Metab. Rev.* 40, 427–446. doi: 10.1080/03602530802186389

Millington, K. A., Fortune, S. M., Low, J., Garces, A., Hingley-Wilson, S. M., Wickremasinghe, M., et al. (2011). Rv3615c is a highly immunodominant RD1 (Region of Difference 1)-dependent secreted antigen specific for *Mycobacterium tuberculosis* infection. *Proc. Natl. Acad. Sci. U.S.A.* 108, 5730–5735. doi: 10.1073/pnas.1015153108

Moayeri, M., Wiggins, J. F., Lindeman, R. E., and Leppla, S. H. (2006). Cisplatin inhibition of anthrax lethal toxin. *Antimicrob. Agents Chemother.* 50, 2658–2665. doi: 10.1128/AAC.01412-05

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185. doi: 10.1093/nar/gkm321

Olakanmi, O., Kesavalu, B., Pasula, R., Abdalla, M. Y., Schlesinger, L. S., and Britigan, B. E. (2013). Gallium nitrate is efficacious in murine models of tuberculosis and inhibits key bacterial Fe-dependent enzymes. *Antimicrob. Agents Chemother.* 57, 6074–6080. doi: 10.1128/AAC.01543-13

Pérez-Martínez, I., Ponce De León, A., Bobadilla, M., Villegas-Sepúlveda, N., Pérez-García, M., Sifuentes-Osornio, J., et al. (2008). A novel identification scheme for genus Mycobacterium, *M. tuberculosis* complex, and seven mycobacteria species of human clinical impact. *Eur. J. Clin. Microbiol. Infect. Dis.* 27, 451–459. doi: 10.1007/s10096-008-0459-9

Restrepo-Montoya, D., Vizcaíno, C., Niño, L. F., Ocampo, M., Patarroyo, M. E., and Patarroyo, M. A. (2009). Validating subcellular localization prediction tools with mycobacterial proteins. *BMC Bioinformatics* 10:134. doi: 10.1186/1471-2105-10-134

Roman, N. O., Samulitis, B. K., Wisner, L., Landowski, T. H., and Dorr, R. T. (2011). Imexon enhances gemcitabine cytotoxicity by inhibition of ribonucleotide reductase. *Cancer Chemother. Pharmacol.* 67, 183–192. doi: 10.1007/s00280-010-1306-0

Roy, A., Bhattacharya, S., Bothra, A. K., and Sen, A. (2013). A database for Mycobacterium secretome analysis: "MycoSec" to accelerate global health research. *OMICS* 17, 502–509. doi: 10.1089/omi.2013.0015

Sampson, S. L. (2011). Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clin. Dev. Immunol.* 2011, 497203–497211. doi: 10.1155/2011/497203

Silva, V. M., Kanaujia, G., Gennaro, M. L., and Menzies, D. (2003). Factors associated with humoral response to ESAT-6, 38 kDa and 14 kDa in patients with a spectrum of tuberculosis. *Int. J. Tuberc. Lung Dis.* 7, 478–484.

Smith, I. (2003). *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. *Clin. Microbiol. Rev.* 16, 463–496. doi: 10.1128/cmr.16.3.463-496.2003

Sreejit, G., Ahmed, A., Parveen, N., Jha, V., Valluri, V. L., Ghosh, S., et al. (2014). The ESAT-6 Protein of *Mycobacterium tuberculosis* interacts with β-2-Microglobulin (β2M) affecting antigen presentation function of macrophage. *PLoS Pathog.* 10:e1004446. doi: 10.1371/journal.ppat.1004446

Tanaka, H., Nakamura, S., Onda, K., Tazaki, T., and Hirano, T. (2009). Sofalcone, an anti-ulcer chalcone derivative, suppresses inflammatory crosstalk between macrophages and adipocytes and adipocyte differentiation: implication of heme-oxygenase-1 induction. *Biochem. Biophys. Res. Commun.* 381, 566–571. doi: 10.1016/j.bbrc.2009.02.086

Tjalsma, H., Antelmann, H., Jongbloed, J. D., Braun, P. G., Darmon, E., Dorenbos, R., et al. (2004). Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome. *Microbiol. Mol. Biol. Rev.* 68, 207–233. doi: 10.1128/MMBR.68.2.207-233.2004

van der Spuy, G. D., Kremer, K., Ndabambi, S. L., Beyers, N., Dunbar, R., Marais, B. J., et al. (2009). Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. *Tuberculosis (Edinb).* 89, 120–125. doi: 10.1016/j.tube.2008.09.003

Vargas-Romero, F., Guitierrez-Najera, N., Mendoza-Hernández, G., Ortega-Bernal, D., Hernández-Pando, R., and Castañón-Arreola, M. (2016). Secretome profile analysis of hypervirulent *Mycobacterium tuberculosis* CPT31 reveals increased production of EsxB and proteins involved in adaptation to intracellular lifestyle. *Pathog. Dis.* 74:ftv127. doi: 10.1093/femspd/ftv127

Vizcaíno, C., Restrepo-Montoya, D., Rodríguez, D., Niño, L. F., Ocampo, M., Vanegas, M., et al. (2010). Computational prediction and experimental assessment of secreted/surface proteins from *Mycobacterium tuberculosis* H37Rv. *PLoS Comput. Biol.* 6:e1000824. doi: 10.1371/journal.pcbi.1000824

Wang, B. L., Xu, Y., Li, Z. M., Xu, Y. M., Weng, X. H., and Wang, H. H. (2005). Antibody response to four secretory proteins from *Mycobacterium tuberculosis* and their complex antigen in TB patients. *Int. J. Tuberc. Lung Dis.* 9, 1327–1334.

Wang, S., Wei, W., and Cai, X. (2015). Genome-wide analysis of excretory/secretory proteins in Echinococcus multilocularis: insights into functional characteristics of the tapeworm secretome. *Parasit. Vectors* 8, 666. doi: 10.1186/s13071-015-1282-7

WHO (2016). *Global Tuberculosis Report.* Geneva.

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zheng, H., Lu, L., Wang, B., Pu, S., Zhang, X., Zhu, G., et al. (2008). Genetic basis of virulence attenuation revealed by comparative genomic analysis of *Mycobacterium tuberculosis* strain H37Ra versus H37Rv. *PLoS ONE* 3:e2375. doi: 10.1371/journal.pone.0002375

Zheng, J., Ren, X., Wei, C., Yang, J., Hu, Y., Liu, L., et al. (2013). Analysis of the secretome and identification of novel constituents from culture filtrate of bacillus Calmette-Guerin using high-resolution mass spectrometry. *Mol. Cell. Proteomics* 12, 2081–2095. doi: 10.1074/mcp.M113.027318

Zhou, F., Xu, X., Wu, S., Cui, X., Fan, L., and Pan, W. (2015). Protein array identification of protein markers for serodiagnosis of *Mycobacterium tuberculosis* infection. *Sci. Rep.* 5:15349. doi: 10.1038/srep15349

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

BMC Genomics

RESEARCH Open Access

Secretome characterization of clinical isolates from the *Mycobacterium abscessus* complex provides insight into antigenic differences

Fernanda Cornejo-Granados[1], Thomas A. Kohl[2,3], Flor Vásquez Sotomayor[4], Sönke Andres[4], Rogelio Hernández-Pando[5], Juan Manuel Hurtado-Ramirez[1], Christian Utpatel[2,3], Stefan Niemann[2,3], Florian P. Maurer[3,4,6*†] and Adrian Ochoa-Leyva[1*†]

Abstract

Background: *Mycobacterium abscessus* (MAB) is a widely disseminated pathogenic non-tuberculous mycobacterium (NTM). Like with the *M. tuberculosis* complex (MTBC), excreted / secreted (ES) proteins play an essential role for its virulence and survival inside the host. Here, we used a robust bioinformatics pipeline to predict the secretome of the *M. abscessus* ATCC 19977 reference strain and 15 clinical isolates belonging to all three MAB subspecies, *M. abscessus* subsp. *abscessus*, *M. abscessus* subsp. *bolletii*, and *M. abscessus* subsp. *massiliense*.

Results: We found that ~ 18% of the proteins encoded in the MAB genomes were predicted as secreted and that the three MAB subspecies shared > 85% of the predicted secretomes. MAB isolates with a rough (R) colony morphotype showed larger predicted secretomes than isolates with a smooth (S) morphotype. Additionally, proteins exclusive to the secretomes of MAB R variants had higher antigenic densities than those exclusive to S variants, independent of the subspecies. For all investigated isolates, ES proteins had a significantly higher antigenic density than non-ES proteins. We identified 337 MAB ES proteins with homologues in previously investigated *M. tuberculosis* secretomes. Among these, 222 have previous experimental support of secretion, and some proteins showed homology with protein drug targets reported in the DrugBank database. The predicted MAB secretomes showed a higher abundance of proteins related to quorum-sensing and Mce domains as compared to MTBC indicating the importance of these pathways for MAB pathogenicity and virulence. Comparison of the predicted secretome of *M. abscessus* ATCC 19977 with the list of essential genes revealed that 99 secreted proteins corresponded to essential proteins required for in vitro growth.

(Continued on next page)

* Correspondence: aochoa@ibt.unam.mx; fmaurer@fz-borstel.de
† Florian P. Maurer and Adrian Ochoa-Leyva contributed equally to this work.
[3] German Center for Infection Research (DZIF), Partner site Hamburg-Lübeck-Borstel, Borstel, Germany
[1] Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autonoma de México, Cuernavaca, Morelos, Mexico
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusions: This study represents the first systematic prediction and in silico characterization of the MAB secretome. Our study demonstrates that bioinformatics strategies can help to broadly explore mycobacterial secretomes including those of clinical isolates and to tailor subsequent, complex and time-consuming experimental approaches accordingly. This approach can support systematic investigation exploring candidate proteins for new vaccines and diagnostic markers to distinguish between colonization and infection. All predicted secretomes were deposited in the Secret-AAR web-server (http://microbiomics.ibt.unam.mx/tools/aar/index.php).

Keywords: Bioinformatics, Antigenicity, *M. abscessus* subspecies, In silico analysis, Vaccinology

Background

Non-tuberculous mycobacteria (NTM) are widely disseminated, mostly saprophytic and partly opportunistic bacteria. The prevalence of NTM in clinical specimens has increased globally, and in some industrialized countries, infections caused by NTM are becoming more common than tuberculosis (TB). Infections caused by *M. abscessus* (MAB) are particularly challenging to manage due to the extensive innate resistance of MAB against a wide spectrum of clinically available antimicrobials [1]. MAB causes mostly pulmonary and occasionally extrapulmonary infections that can affect all organs in the human body [2]. Current treatments for MAB induced pulmonary disease are long, associated with severe side effects and a cure rate below 50% [3–5]. MAB is comprised of three subspecies, *M. abscessus* subsp. *abscessus, M. abscessus* subsp. *bolletii* and *M. abscessus* subsp. *massiliense*, hereafter referred to as $MAB_{A,}$ MAB_B, and MAB_M, respectively [6]. MAB isolates can show smooth (S) and rough (R) colony morphotypes, a trait that relies on the presence (S) or absence (R) of surface-associated glycopeptidolipids (GPLs) and that correlates with the virulence of the strain [7–10]. Transitioning from high-GPL to low-GPL production is observed in sequential MAB isolates obtained from patients with chronic underlying pulmonary disease. In these patients, S-to-R conversion is thought to present a selective advantage as the aggregative properties of MAB R variants strongly affect intracellular survival. The selective advantage is also related to the loss of immunogenic GPLs. In addition, a propensity to grow as extracellular cords allows these low-GPL producing bacilli to escape innate immune defenses [10].

The complete set of proteins excreted / secreted (ES) by a bacterial cell is referred to as its secretome. The secretome is involved in critical biological processes such as cell adhesion, migration, cell-to-cell communication and signal transduction [11] ES proteins are considered an important source of molecules for serological diagnosis. Also, secreted proteins can be highly antigenic due to their immediate availability to the host immune system and are thus of interest in vaccinology [12, 13]. So far, there have been few efforts to experimentally determine the secretome of MAB, and in particular, the secretomes of clinical MAB isolates [14–17]. Nowadays, sequencing and bioinformatics strategies can be explored for the systematized prediction of ES proteins from bacterial genomes [18, 19]. Recently, a robust bioinformatics pipeline for predicting and analyzing the complete in silico secretome of two clinical *M. tuberculosis* (MTB) genomes was published showing higher overall agreement with an experimental secretome compiled from literature than two previously reported secretomes for *M. tuberculosis* H37Rv [19].

To gain further insights into MAB ES proteins and their association with virulence and pathogenicity we here sequenced and assembled the genomes of 15 clinical MAB isolates belonging to all three subspecies including S and R morphotypes. We then adapted the bioinformatics strategy previously established for MTB to predict and analyze the complete set of ES proteins encoded in these isolates and in the *M. abscessus* ATCC 19977 type strain, and compared it with our previous findings for MTB [19].

Results

Genome assembly, secretome prediction and annotation

We sequenced the genomes of 15 pulmonary and extrapulmonary (skin, tissue, lymph node, and blood) MAB isolates *obtained* from patients in Germany comprising all three MAB subspecies (Table 1 and Additional file 1: Table S1). For each genome, we obtained an average of 2,601,444 quality-filtered reads. After de novo assembly, we obtained from 38 to 78 contigs (mean = 58 contigs) with genome coverage of 217- to 368-fold (mean = 310-fold) and with an average of 5082 total proteins per genome (Additional file 3: Table S2). Also, we performed a Multilocus Sequence Typing (MLST) analysis at the Pasteur Institute site (https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_mycoabscessus_seqdef) to assess the genetic variability among the studied samples. This analysis assigns a Sequence Type (ST) to each strain by looking for sequence variations in seven housekeeping genes and providing information about philogenetic relationship [20]. We observed that eight out of 15 genomes had unique STs, three genomes were not defined and notably, two genomes belonged to ST 117

Table 1 Clinical isolates metadata and number of ES proteins

Strain	Accession number	Genome ID	Origin	Phenotype	Sequence Type (ST)	Total predicted proteins	ES proteins	% ES proteins
M. abscessus subsp. abscessus	GCA_ 015499845.1	4549-15	sputum	rough	1	5105	929	18
	GCA_ 015499865.1	11351-15	sputum	rough	63	5138	966	19
	GCA_ 015499835.1	8844-15	skin	smooth	246	4854	956	20
	GCA_ 015499805.1	3563-15	sputum	smooth	33	5239	968	18
	GCA_ 015499795.1	12389-15	sputum	smooth	47	5276	990	19
	GCA_ 015499765.1	2677-16	sputum	smooth	34	4900	919	19
	GCA_ 015499745.1	2572-17	tissue (breast implant)	NA	10-46-64-70-261	4847	874	18
M. abscessus subsp. massiliense	GCA_ 015499715.1	14479-15	sputum	rough	117	5120	962	19
	GCA_ 015499735.1	10896-16	sputum	rough	117	5109	950	19
	GCA_ 015499695.1	10003-15	sputum	smooth	98-245-271	4835	891	18
	GCA_ 015499655.1	16155-15	sputum	smooth	98-245-271	4884	898	18
M. abscessus subsp. bolletii	GCA_ 015499665.1	11702-16	sputum	rough	161	5079	931	18
	GCA_ 015499625.1	713-16	lymph node	rough	52	5456	1037	19
	GCA_ 015499615.1	7742-15	blood culture	smooth	333	4913	885	18
	GCA_ 015499585.1	13116-16	lymph node	smooth	52	5305	990	19
M. abscessus subsp. abscessus	CU458896.1	reference strain ATCC19977	–	–		4942	886	18
M. tuberculosis H37Rv	NC_000962.3	reference strain	–	–		4337	548	13

while other two belonged to ST 52, suggesting they could be highly related (Table 1).

We used a bioinformatics pipeline previously reported by our group [19] to predict the full secretome of all MAB clinical isolates and the widely used reference strain *M. abscessus* ATCC 19977 (GenBank CU458896.1) (Additional file 2: Fig. S1). We obtained an average of 939 ES proteins per genome, representing ~ 18% of the total proteome (Table 1). The predicted secretome for the MAB reference strain consisted of 886 proteins. All these proteins showed a BLASTP hit against the NR database but only 494 (55.8%) could be annotated with GO terms.

We analyzed the over-representation of GO terms in the secretome of *M. abscessus* ATCC 19977 as compared to the whole genome. The most significantly enriched GO-terms were: "lytic vacuole" (p = 9.37E-04), and "fungal-type vacuole" (p = 0.004) in Cellular Component (Fig. 1a), "serine-type carboxypeptidase" (p = 1.83E-04), and "serine-type D-Ala-D-Ala carboxypeptidase" (p = 1.83E-04) activities in Molecular Function (Fig. 1b) and, "response to inorganic substance" (p = 5.68E-04) and "cellular response to oxygen radical" (p = 0.001) in the Biological Process category (Fig. 1c). The KEGG pathway mapping of the ES proteins showed that 214 proteins (24.2%) could be assigned to 100 different KEGG pathways (Table 2), with the ABC transporter pathway being the most abundant (n = 13, 1.47%). Additionally, serine-type D-Ala-D-Ala carboxypeptidases (p = 1.83E-04) and peptidases (p = 8.40E-04) were the most significantly abundant enzymes according to the Enzyme Commission (EC) Classes (Additional file 6: Fig. S2), while the Mce/MiaD and PknH-like extracellular domains were the most enriched protein domains (Table 3). Of note, comparably few sequences were assigned to the PE/PPE category (n = 3). Notably, after comparing the predicted secretome of *M. abscessus*

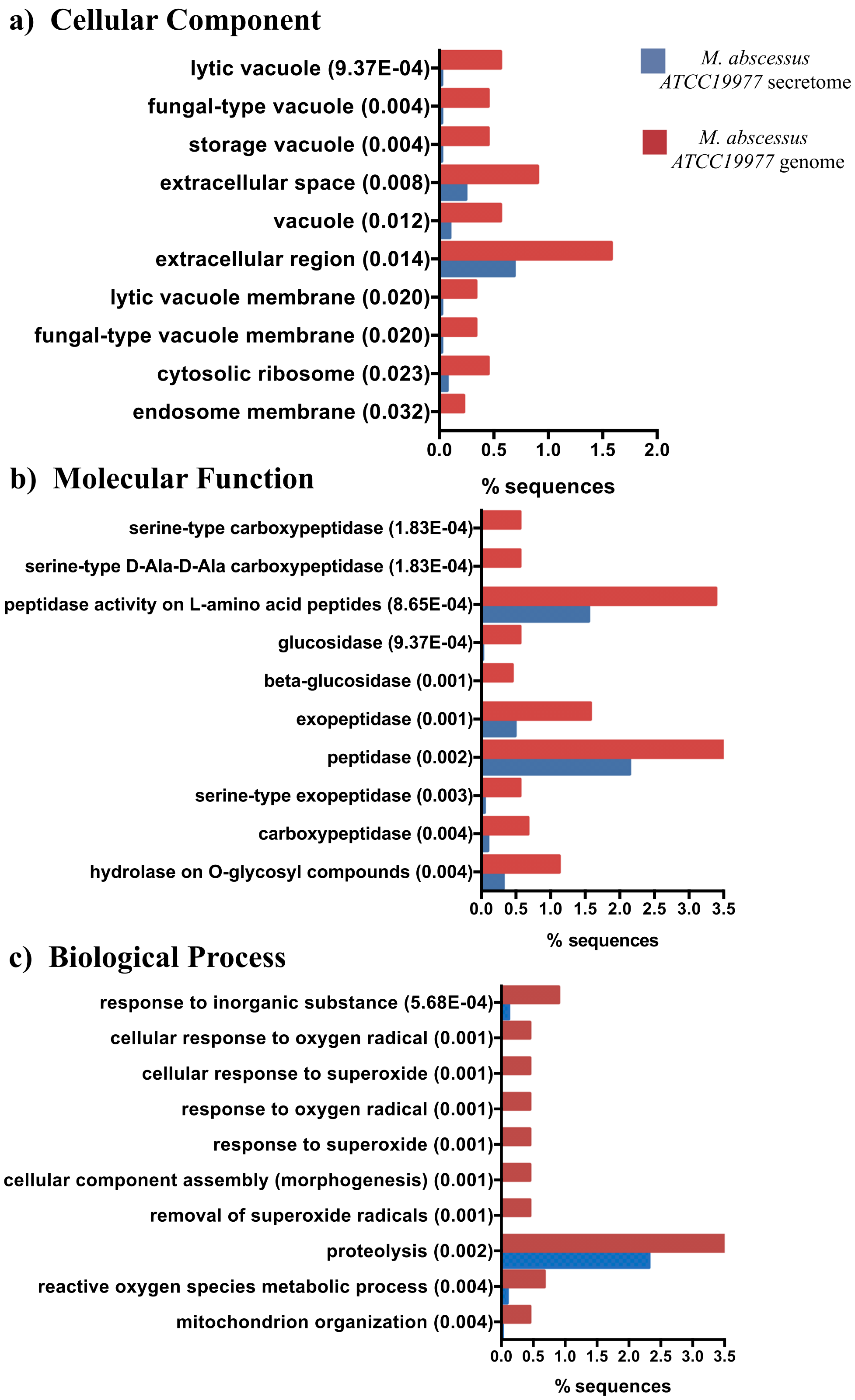


Fig. 1 GO enrichment analysis for the *M. abscessus* ATCC 19977 reference strain. Top 10 most enriched GO terms for the *M. abscessus* ATCC 19977 secretome (blue) and complete genome (red) in three categories: **a** Cellular Component, **b** Molecular Function and **c** Biological Process

Table 2 Top 10 KEGG pathways assigned for *M. abscessus* ATCC19977 ES proteins

Ranking	Pathway name	Number of represented ES proteins (%)
1	ABC transporters	13 (1.47)
2	Two-component system	9 (1.02)
3	Quorum sensing	6 (0.68)
4	Oxidative phosphorylation	4 (0.45)
5	Sulfur metabolism	4 (0.45)
6	Glycerolipid metabolism	4 (0.45)
7	Peptidoglycan biosynthesis	4 (0.45)
8	Protein export	4 (0.45)
9	Starch and sucrose metabolism	3 (0.34)
10	Glyoxylate and dicarboxylate metabolism	3 (0.34)

ATCC 19977 with a list of essential genes published by Laencina et al. [17], we found that 99 (11.17%) of the predicted ES proteins, corresponded to essential proteins required for in vitro growth.

Comparison of *M. abscessus* subspecies core secretomes

We analyzed the differences between the predicted secretomes of the three MAB subspecies. To this end, we defined the core secretome of each subspecies as the set of proteins shared between all secretomes of isolates belonging to MAB_A, MAB_B, and MAB_M, respectively. The resulting core secretomes contained 735 (MAB_A), 794 (MAB_B), and 813 (MAB_M) proteins (Fig. 2a).

Given that our study considered a limited number of de novo assembled genomes, we additionally compared the predicted core secretomes to 60 additional MAB genomes available in NCBI (Additional file 4: Table S3). We found that an average of 99.78, 99.12, and 98.59% of our core secretomes was also present in the additional MAB_A, MAB_B, and MAB_M genomes, respectively, further corroborating the validity of the predicted subspecies core secretomes for other MAB isolates.

We then determined the respective Abundance of Antigenic Regions (AAR) values to estimate antigenic densities for the protein sets in each core secretome. The average AAR values from most to least antigenic were: 40.24 for MAB_A, 40.75 for MAB_B, and 41.38 for MAB_M with no statistically significant difference between them.

Next, we identified the ES proteins shared between the MAB_A, MAB_B, and MAB_M core secretomes. We found that 704 proteins (86.5%) were shared among MAB_A, MAB_B, and MAB_M with an AAR value of 41.17 (Fig. 2b). The AAR values for the protein sets exclusively found in the MAB_A, MAB_B, or MAB_M secretome were 33.58, 41.22, and 43.13, respectively, with the MAB_A dataset showing a significantly lower AAR value indicating higher antigenicity than the others ($p < 0.1$; Fig. 2b).

Differences in core secretomes between R and S morphotypes

As MAB isolates with R and S morphotypes show differences in virulence and pathogenicity, we compared the predicted core secretomes of R and S isolates (Fig. 3). We observed that the core secretomes of R variants were larger (840, 924 and 845 proteins for MAB_A, MAB_M, and MAB_B) than those of the investigated S variants (764, 872 and 833 proteins, respectively) with no significant differences in antigenic densities as per mean AAR

Table 3 Top 10 most represented protein domains in *M. abscessus* ATCC19977 secretome

InterProcode	InterPro description	Number of ES proteins (%)
IPR003399	Mce/MlaD	19 (2.14)
IPR026954	PknH-like extracellular domain	15 (1.69)
IPR032407	Haemophore, haem-binding	10 (1.13)
IPR020846	Major facilitator superfamily domain	7 (0.79)
IPR013766	Thioredoxin domain	6 (0.68)
IPR000064	Endopeptidase, NLPC/P60 domain	6 (0.68)
IPR001638	Solute-binding protein family 3/N-terminal domain of MltF	6 (0.68)
IPR000675	Cutinase/acetylxylan esterase	6 (0.68)
IPR005490	L,D-transpeptidase catalytic domain	5 (0.56)
IPR000073	Alpha/beta hydrolase fold-1	5 (0.56)

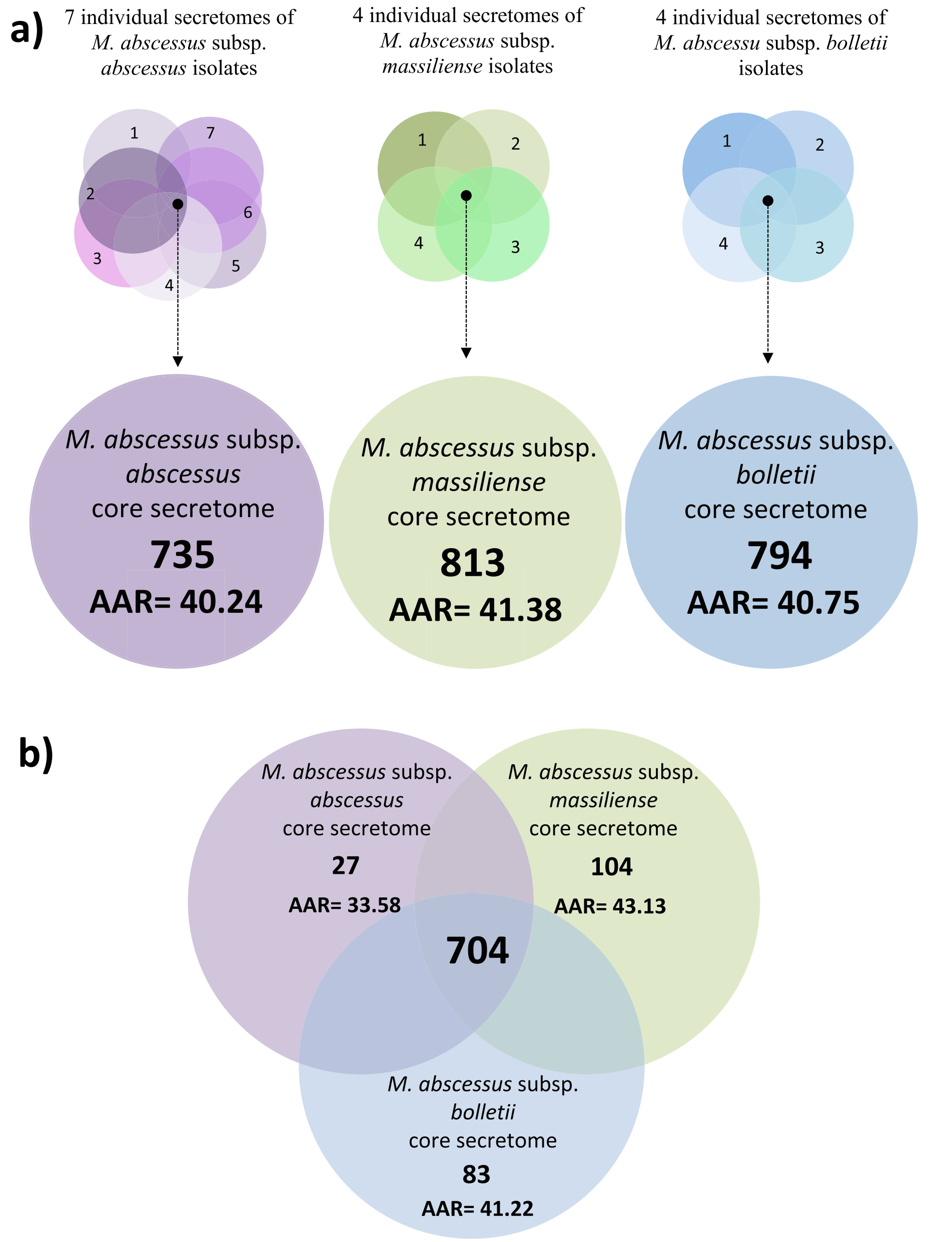

a)
7 individual secretomes of *M. abscessus* subsp. *abscessus* isolates
4 individual secretomes of *M. abscessus* subsp. *massiliense* isolates
4 individual secretomes of *M. abscessu* subsp. *bolletii* isolates
1
7
2
6
3
5
4
1
2
4
3
1
2
4
3
M. abscessus subsp. *abscessus* core secretome
735
AAR= 40.24
M. abscessus subsp. *massiliense* core secretome
813
AAR= 41.38
M. abscessus subsp. *bolletii* core secretome
794
AAR= 40.75
b)
M. abscessus subsp. *abscessus* core secretome
27
AAR= 33.58
M. abscessus subsp. *massiliense* core secretome
104
AAR= 43.13
704
M. abscessus subsp. *bolletii* core secretome
83
AAR= 41.22

Fig. 2 Venn diagram between the core secretomes of the three *M. abscessus subspecies.* **a** Number of total proteins contained in the core secretome of each subspecies. **b** Shared and unique proteins between the three subspecies as per BLASTP (E-value 1.0E-3)

value (Fig. 3). Intra-subspecies comparison of S and R secretomes revealed that 96.4, 90.7 and 95% of the identified ES proteins were found in both R and S morphotypes for MAB_A, MAB_M and MAB_B respectively. The number of unique proteins was larger in the core secretome of the R morphotypes (n = 93, 109, and 48 for MAB_A, MAB_M, and MAB_B) as compared to the S morphotypes (n = 9, 76, and 35, respectively; Fig. 3).

Interestingly, antigenic densities for the unique ES proteins of the R morphotypes were higher (AAR = 40.84, 36.71, and 35.59 for MAB_A, MAB_M, and MAB_B) than for the proteins exclusive to the S morphotypes irrespective of the subspecies (AAR = 45.43, 37.72, and 42.14; Fig. 3). To assess if the AAR values of these specific protein sets were different from same-sized protein sets randomly chosen from the respective core secretomes, we created 1000 random sets of 109, 93, 76, 48, 35 and 9 proteins and calculated the AAR value for each set. Then, we determined an empirical p-value based on the number of random protein sets that equaled or exceeded the AAR value for each protein dataset as was previously suggested by Cornejo-Granados et al. [19].

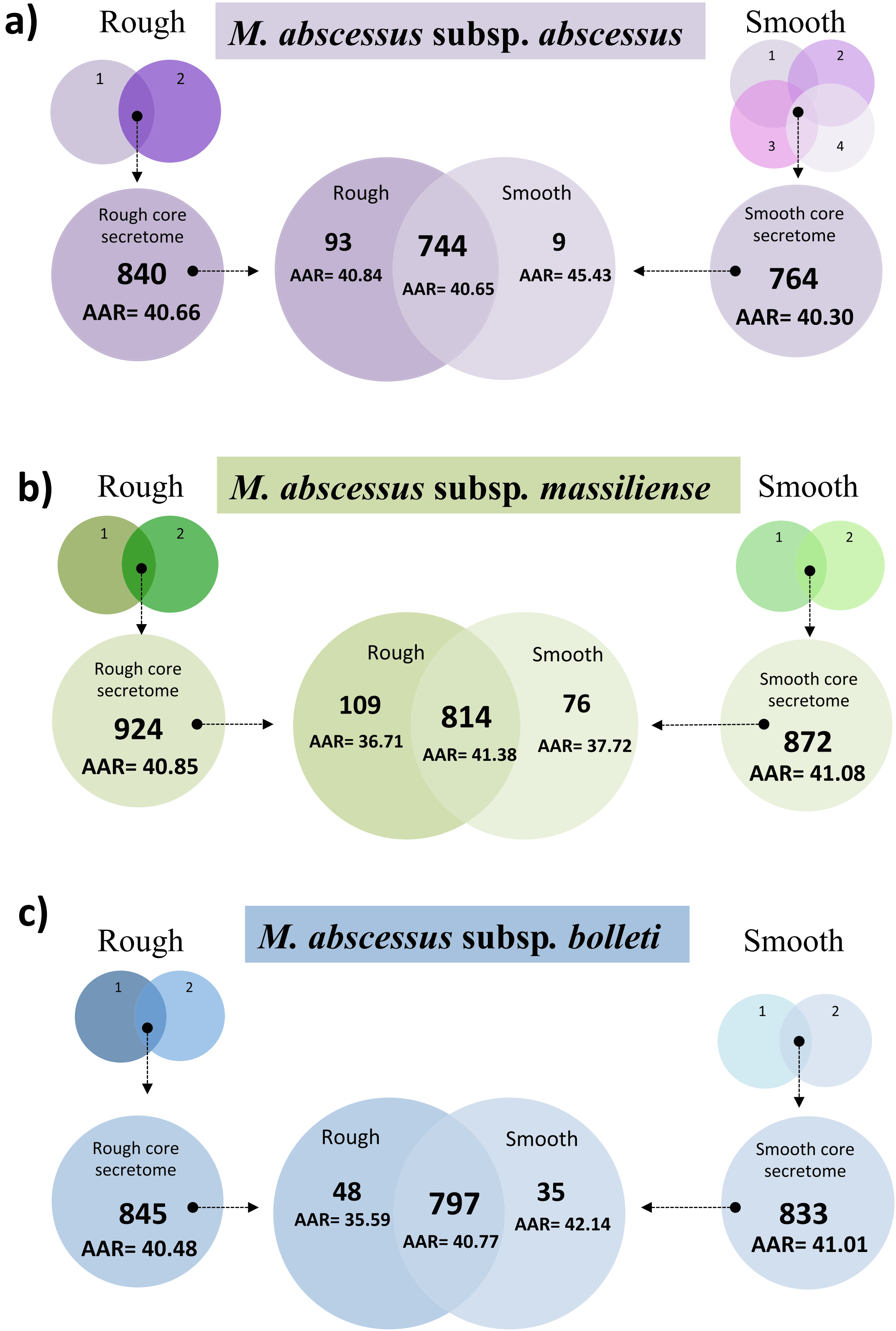


Fig. 3 Venn diagram between the core secretomes of the three *M. abscessus* subspecies by colony morphotype. We used BLASTP (E-value 1.0E-3) to assess the core secretomes for isolates with rough and smooth colony morphotypes **a** *M. abscessus* subsp. *abscessus*, **b** *M. abscessus* subsp. *massiliense* and **c** *M. abscessus* subsp. *bolletii*

We found that the ES proteins exclusive to the R morphotypes of MAB_M and MAB_B had significantly ($p < 0.05$) higher antigenic densities than randomly constructed protein sets (Additional file 5: Table S4).

Finally, we determined the MAB core secretomes by sample origin (pulmonary, extrapulmonary, blood). This resulted in 706 ES proteins shared among the ten pulmonary isolates, 758 proteins shared among the four extrapulmonary isolates, and 885 proteins for the single isolate grown from a blood sample. However, as per the GO, KEGG, and antigenicity analyses, we did not find any distinct characteristics specific to either sample source and, hence, type of infection.

Antigenicity of ES and non-ES proteins

It has previously been reported for different microorganisms including MTB that ES proteins tend to be more antigenic than non-ES proteins [18, 19, 21]. We thus tested if this was also true for the investigated MAB isolates. First, we found that the antigenic densities as indicated by mean AAR values were similar among all isolates irrespective of subspecies or morphotype within the same cell compartment, i.e. for ES, non-ES, intracellular (incell) and transmembrane (TM) proteins (Fig. 4). Second, we found that antigenic densities were significantly higher in ES proteins as compared to non-ES proteins in all isolates (AAR = 40.57 and 43.60, respectively; p-value < 0.0001) (Fig. 4). However, within the non-ES category, incell proteins showed even higher antigenic densities (AAR = 39.04) than the predicted ES proteins ($p < 0.0001$) while the lowest overall antigenic densities were observed for the TM category (AAR = 59.23; $p < 0.0001$).

Comparison of *M. abscessus* and *M. tuberculosis* secretomes

Lastly, we compared the predicted secretome of *M. abscessus* ATCC 19977 against the previously reported secretome of *M. tuberculosis* H37Rv [19]. We observed that the *M. abscessus* secretome was predicted to be almost equally antigenic (AAR = 39.63) than the *M. tuberculosis* secretome (AAR = 40.37) (Fig. 5). We found 337 MAB ES proteins (38.04%) with homology to proteins in the predicted MTB secretome (Fig. 5). Interestingly, 222 of these proteins had sequence homology with proteins experimentally reported as secreted in MTB (comparable experimental secretome data for MAB was not available to us) [19] (Additional file 7: Table S5). Furthermore, we determined the average AAR value of the 680 ES proteins shared among the 15 MAB isolates (AAR = 41.53). This value means that antigenic density was lower than for the predicted secretome of *M. tuberculosis* H37Rv (AAR = 40.37) and two clinical *M. tuberculosis* isolates belonging to the Beijing lineage (isolate C3 AAR = 37.52 and isolate C4 AAR = 37.55) (Table 4) [19]. Finally, we identified 13 ES proteins with homologues in both MAB and *M. tuberculosis*, which are listed as targets for various FDA approved drugs (Additional file 8: Table S6).

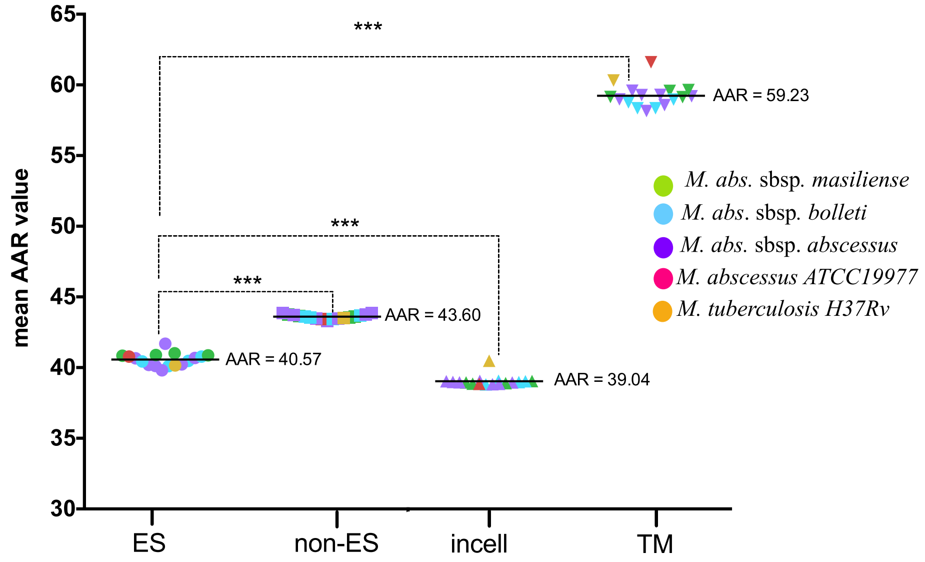


Fig. 4 Comparison between AAR values for Excreted/Secreted (ES), non Excreted/Secreted (non-ES), intracellular (incell) and transmembrane (TM) proteins. AAR values were calculated for each of the 15 genomes sequenced. The X-axis shows the cellular compartment and the Y-axis shows AAR values for the genomes of each subspecies: *M. abscessus* subsp. *abscessus* (green), *M. abscessus* subsp. *bolletii* (blue), *M. abscessus* subsp. *massiliense* (purple), *M. abscessus* ATCC19977 (red) and *M. tuberculosis* H37Rv (orange). Mann-Whitney tests were performed to compare the AAR of each group with a confidence level of 99% (***, $p < 0.001$)

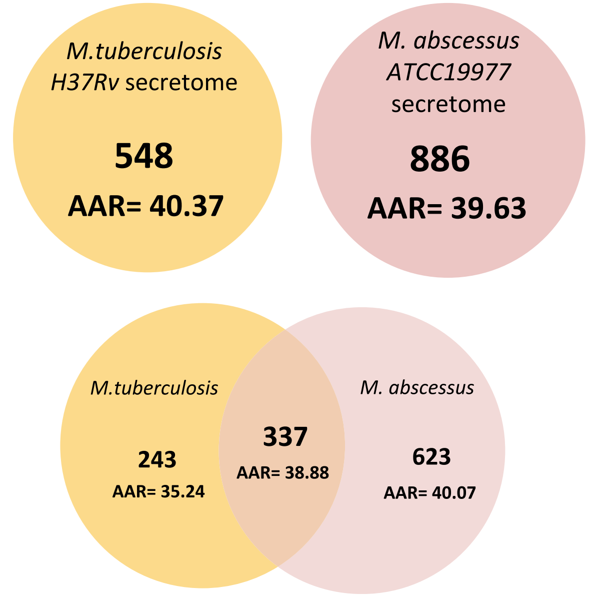


Fig. 5 Venn diagram between the predicted secretomes of *M. tuberculosis* H37Rv and *M. abscessus* ATCC 19977. We used BLASTP (E-value 1.0E-3) to compare the complete secretomes of both species

Discussion

This is the first study that proposes a method for prediction of MAB secretomes based on 15 clinical MAB isolates and the *M. abscessus* ATCC 19977 reference strain. Our results show that an average of 18% (939 proteins) of the total proteins encoded in the MAB core genome carry sequence patterns indicative of secretion. Notably, this percentage is 6% greater than the proportion previously reported for several MTB isolates (~ 12%) [19]. Nearly 200 species of mycobacteria have been identified with diverse lifestyles and a high degree of morphological, biochemical, and physiological diversity and a comparative genome analysis suggests that only a relatively small number of genes (1080) are shared between several *Mycobacterium* species [22, 23]. Moreover, loss of ancestral genes is a well described phenomenon in slowly growing mycobacteria such as MTB and, in particular, *M. leprae* [24]. In contrast, rapidly growing NTM such as MAB are considered to represent a more ancient evolutionary state, with larger genomes than those of MTB [23, 24]. Thus, it is not surprising that we found a larger number of ES proteins in MAB than MTB. Furthermore, the increased abundance of ES proteins in MAB as compared to MTB could be related to the ability of MAB to cause a different spectrum of disease and to adapt to different environmental settings requiring frequent interaction with a wide variety of host cells and organisms competing for the same ecological niche, likely involving cross species exchange of genetic information, for example by plasmid transfer [25–27]. A similar hypothesis has been suggested for fungal secretomes [28].

The GO and KEGG pathway annotations of the secretomes of *M. abscessus* ATCC 19977 and the MAB clinical isolates showed enrichment consistent with the characterization of previously reported mycobacterial secretomes [18, 19]. Interestingly and in line with the increased secretome size as compared to MTB, the KEGG pathway analysis showed a high abundance of the Quorum sensing pathway for the predicted MAB secretomes, which was not present in our previous MTB secretome pathway analysis [19]. The presence of a Quorum sensing pathway would be another similarity shared between MAB and non-mycobacterial pathogens commonly affecting patients with chronic lung disease such as *Pseudomonas aeruginosa* [29]. In addition, it could be related to the ability of MAB to form biofilms [30, 31], further contributing to the capacity of MAB to tolerate antibiotics and to persist over long periods in the environment [32–35].

The InterPro annotation showed that Mce domains were the most abundant (2.14%) domains in the MAB reference secretome, while PPE and PE-PGRS domains only corresponded to 0.3% of the ES protein sequences. This tendency is contrary to our observations for MTB [19], where the PPE and PE-PGRS domains accounted for ~ 12% of the secreted proteins and the Mce domains for only 0.5%. The lower quantity of predicted PE/PPE proteins in MAB was somewhat expected. *M. tuberculosis* has five ESX secretion systems, four of which encode PE/PPE proteins, while MAB has only two (ESX-3 and ESX-4) of which only the ESX-3 operon includes

Table 4 Abundance of Antigenic Regions (AAR) for *M. abscessus* and *M. tuberculosis* strains

Strain	Number of proteins in the dataset	Average AAR value
M. tuberculosis Beijing isolate C3[a]	553	37.52
M. tuberculosis Beijing isolate C4[a]	519	37.55
M. bovis BCG Pasteur	564	38.99
***M. abscessus* ATCC 19977**	**886**	**40.78**
M. tuberculosis H37Rv	548	40.37
M. abscessus clinical isolates	680	41.54

[a] Both Beijing isolates were previously reported in Cornejo-Granados et al. [19]

PE/PPE genes [36]. In contrast, Mce domains are known for participating in host cell entry by mycobacteria [37]. Thus, their higher abundance in MAB as compared to MTB highlights the importance of this pathway for MAB survival within the host. It needs to be mentioned though that Kumar et al. [37], also suggested that in low virulence bacteria, transport activities could be the primary function of Mce operons.

To compare the predicted secretomes according to colony morphotype, we first established the core secretome for the R and S variants per subspecies, thus eliminating individualities among the different isolates (Fig. 3). The high overall agreement between the core secretomes for both morphotypes of approximately 90% was expected, considering the fact that R variants can arise from the S morphotypes during persistent infection by loss of surface-exposed GPLs caused by mutations in the GPL synthesis pathway [26, 38–40]. However, both the higher number and the higher antigenic densities (lower AAR values) of the ES proteins exclusively found in R variants indicate that additional genetic changes may evolve during S-to-R conversion. Moreover, this observation raises the question whether some strains with additional genetic traits associated with virulence are able to undergo S-to-R conversion and cause disease due to R variants more easily than others. Genomic studies involving sequentially isolated S and R variants of the same strain obtained from individual patients over time will be required to better characterize the microevolution of MAB strains within the chronically infected host.

Similarly, the fact that MAB causes both chronic pulmonary disease (with R variants sometimes increasing over time) and extrapulmonary manifestations (mostly caused by S variants) led us to investigate whether differences exist in the predicted secretomes of isolates related to these clinical presentations. The absence of major differences in the GO, KEGG, and antigenicity analyses suggest that secretome variations do not influence MAB tissue tropism. Consequently, host characteristics such as severe immunosuppression may be the main driver for invasive MAB infections. Likewise, in the case of tissue infections, which often occur following surgical interventions, insufficient hygiene procedures and sterilization protocols for surgical equipment appear to be more relevant than pathobiological traits such as the secretome intrinsic to the causative MAB isolate [41].

Lastly, we observed that the predicted secretomes of all investigated clinical MAB isolates were less antigenic than the secretomes of *M. tuberculosis* H37Rv and two clinical *M. tuberculosis* isolates. Additionally, although there was no statistical difference, the isolates with a rough phenotype tended to be more antigenic than the isolates with a smooth phenotype. Previous evidence with *M. tuberculosis* [19] showed that clinical isolates from the Beijing phenotype showed increased virulence and less antigenic secretomes than the reference strain H37Rv. Thus, the diminished antigenicity of MAB could be viewed as a virulence trait in itself as it would support colonization of the host for extended time periods without immediate progression into clinical disease. However, further experimental tests on antigenicity are needed to demonstrate this observation.

This study represents the first systematic prediction and in silico characterization of the MAB secretome. We acknowledge that an important constraint in this study is the limited total number of genomes analyzed per subspecies and biological source. Thus, care must be taken to not over interpret the findings related to sample subcategories such as subspecies and morphotypes. Also, published experimental data on MAB secretomes are very limited and no systematic validation of the in silico findings reported herein could be performed against such datasets. Although more research will be needed to determine experimental secretomes in NTM, our study demonstrates that using bioinformatics strategies can help to broadly explore mycobacterial secretomes including those of clinical isolates and to tailor subsequent, complex and time-consuming experimental approaches accordingly. This approach can support a systematic investigation of mycobacterial secretomes exploring candidate proteins suitable for developing new vaccines and diagnostic markers to distinguish between colonization and infection.

Methods

Clinical isolates

We selected 15 MAB clinical isolates comprising members of all MAB subspecies (MAB_A, $n = 7$; MAB_B, $n = 4$; MAB_M, n = 4) and both S ($n = 8$) and R ($n = 6$) morphotypes (not determined, $n = 1$). The strains were isolated from different biological sources representing both pulmonary colonization / infection (sputum, $n = 10$) and extrapulmonary samples (skin, $n = 1$; soft tissue, $n = 1$; lymph nodes, $n = 2$; blood, $n = 1$) (Table 1 and Additional file 1: Table S1). For routine diagnostic purposes, species identification was performed using GenoType NTM-DR line probe assays (HAIN Lifescience, Nehren, Germany) and sequencing of the 16S and *rpoB* genes as described previously [42].

Whole genome sequencing and genome assemblies

Genomic DNA (gDNA) of the 15 MAB clinical isolates was extracted from solid cultures using a Centrimonium bromide chloroform DNA extraction protocol as previously described [43]. DNA libraries were constructed with the Nextera XT kit from Illumina and sequenced on the Illumina MiSeq benchtop platform with a v3 chemistry paired –end run and a read lenght of 2 × 300 bp. We processed the resulting reads with Trimmomatic [44], clipping the Illumina adapter sequences and trimming the reads with a sliding window of 20 bp looking for quality > 30 and discarding all reads shorter than 100 bp. Trimmed reads were used to construct de novo assemblies using SPADES [45] with default parameters and the --careful option enabled. Then, each assembly was analyzed with RAST [46] to obtain all the open reading frames (ORFs). Additionally, we predicted the ORFs from the deposited genome sequence of the *M. abscessus* ATCC 19977 type strain (GenBank CU458896.1) (Additional file 1: Table S1).

Secretome prediction

The complete set of predicted ORFs was independently analyzed for each genome using the bioinformatics pipeline previously reported by Cornejo-Granados et al. [19] and summarized in Additional file 2: Figure S1. Briefly, we used six different feature-based tools (SignalP, SecretomeP, LipoP, TatP, TMHMM and Phobius) [47–51] to identify ES proteins by the different secretion pathways and to remove the ones that had transmembrane domains (Additional file 2: Fig. S1). The proteins assigned as not-secreted (non-ES) were further classified into transmembrane proteins (TM) if they showed the presence of transmembrane domains with TMHMM 2.0 [50], and into intracellular proteins (incell) if they did not contain any transmembrane domains.

Annotation and comparative analysis of secreted proteins

To assign functional annotations to the proteins present in our genomes, we performed a BLASTP query of those proteins against the non-redundant (nr) complete database using Blast2GO [52] with an E-value cut-off set at 1.0E-3. Furthermore, all proteins were associated with protein families through InterProScan [53] and functionally mapped to Gene Ontology (GO) terms by setting the following parameters: E-value-hi-filter: 1.0E-3; Annotation cut-off: 55; GO weight: 5 and Hsp-Hit Coverage cut-off: 0. Blast2GO was then used to identify over- and under-represented GO and Enzyme Commission (EC) numbers in the ES proteins by setting the significance filter *p*-value to ≤0.05. Also, we used the KEGG Automatic Annotation Server (KAAS) database [54] to assign the pathway annotation to the secreted proteins using the BBH (bidirectional best hit) method and the reference gene data set assigned to *Mycobacterium*.

To determine differences between the predicted secretomes in relation to MAB subspecies and morphotype, we established core secretomes by performing a bidirectional best-hit BLASTP search (E-value 1.0E− 3) between the ES proteins of all genomes belonging to the respective subspecies and morphotypes. Then, we identified the shared and unique proteins for each comparison. Additionally, we determined the set of homologous ES proteins shared between the MAB reference strain ATCC 19977 and both *M. tuberculosis* H37Rv predicted and experimental secretomes [19]. The resulting proteins were further investigated for sequence similarities against known drug targets available on the Drug Bank database (http://www.drugbank.ca/), setting the E-value to 1.0E-3 and all other options to default. In Additional file 8: Table S6, we show all proteins that have similarity with an approved drug target, as well as the drugs that can affect said target.

Additionally, we analyzed the presence of the core secretomes in 20 *M. abscessus* genomes per subspecies downloaded from NCBI (Additional file 4: Table S3). To this end, each downloaded genome was analyzed with RAST to obtain all the open reading frames (ORFs). Next, we performed a BLASTP search (E-value 1.0E− 3) of each core secretome against each genome of the corresponding subspecies, and all hit proteins were considered homologs.

Calculation of the abundance of antigenic regions

The Abundance of Antigenic Regions (AAR) value is used to estimate the antigenic density of a protein by calculating the number of antigenic regions and normalizing it to the sequence length [18]. Of note, proteins with higher antigenic densities have lower AAR values. For this study, we calculated the AAR value for each protein in each data set using the Secret-AAR webserver (http://microbiomics.ibt.unam.mx/tools/aar/index.php) and reported the average unless stated otherwise [55]. Then, we used a Mann-Whitney statistical test to establish any significant differences between the AAR values of the different protein data sets.

Abbreviations

MAB: *Mycobacterium abscessus*; NTM: Non-tuberculous mycobacterium; MTBC: *M. tuberculosis complex*; MTB: *M. tuberculosis*; TB: Tuberculosis; ES: Excreted / secreted; R: Rough; S: Smooth; MAB_A: *M. abscessus* subsp. *abscessus*; MAB_B: *M. abscessus* subsp. *bolletii*; MAB_M: *M. abscessus* subsp. *massiliense*; non-ES: Not-secreted; TM: Transmembrane; incell: Intracellular; AAR: Abundance of Antigenic Regions; MLST: Multilocus Sequence Typing; ST: Sequence Type

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07670-7.

Additional file 1: Table S1. Complete metadata of the 15 clinical isolates of *M. abscessus* genomes sequenced.

Additional file 2: Figure S1. Bioinformatics pipeline to indentify and analyze the secreted proteins of *M. abscessus*.

Additional file 3: Table S2. Statistic data of the de novo assemblies for the sequenced isolates.

Additional file 4: Table S3. Comparison of the core secretome of each subspecies vs NCBI genomes.

Additional file 5: Table S4. AAR values for random constructed secretomes of the rough and smooth phenotypes.

Additional file 6: Figure S2. GO enrichment analysis of enzymes for *M. abscessus* ATCC19977. Percentage of sequences annotated with each GO term for the secreted proteins (blue) and the complete proteins in the genome (red).

Additional file 7: Table S5. List of 222 M. *abscessus* proteins with homologues in *M. tuberculosis* H37Rv and with previous experimental support for secretion according to Cornejo-Granados et al. [19].

Additional file 8: Table S6. Potential drug targets for 13 proteins shared between *M. tuberculosis* H37Rv and M. *abscessus* ATCC19977.

Acknowledgements
We would like to thank Julia Zallet and Vanessa Mohr for excellent technical assistance and Henrik Nielsen for his assistance with the SecrtomeP software.

Authors' contributions
Conceived and designed the experiments: FCG, AOL, FPM. Performed the experiments: FCG, JHR. Analyzed the data: FCG, TAK, FVS, SA, JHR, CU, SN, FPM, AOL. Contributed reagents/materials/analysis tools: AOL, RHP. Wrote and edited the manuscript: FCG, FVS, JHR, RHP, AOL, FPM. All authors read and approved the final version of the manuscript.

Funding
Parts of this work have been supported by the German Center for Infection Research and Grants by Joachim Herz Foundation, Hamburg, and Mukoviszidose Institut gGmbH, Bonn, the research and development arm of the German Cystic Fibrosis Association Mukoviszidose e.V. to F.P.M. We acknowledge the support provided by CONACyT grant CB- 2013-223279 and SALUD-2014-C01–234188 to A.O.L. This research also received support by the DGAPA PAPIIT UNAM (IN215520) to A.O.L. F.C.G. acknowledges the support of CONACyT as a Postgraduate fellow.

Availability of data and materials
The reference genomes analyzed for *M. abscessus* ATCC19977 and *M. tuberculosis* H37Rv were taken from NCBI, under GenBank IDs CU458896.1 and NC_000962.3, respectively. The Whole Genome Shotgun project has been deposited at NCBI, under BioProject PRJNA646278. It can be accessed with the link https://www.ncbi.nlm.nih.gov/bioproject/PRJNA646278.
All the predicted secretomes were deposited in the Secret-AAR web-server (http://microbiomics.ibt.unam.mx/tools/aar/index.php).
Additional data supporting the conclusions of this article are included within the article and its additional file(s).

Declarations

Ethics approval and consent to participate
Ethical review and approval was not required for the study as all work was performed on bacterial isolates archived at the strain repository of the National Reference Center for Mycobacteria in Borstel, Germany, in accordance with local legislation and institutional requirements. In particular, no data allowing identification of the affected patients was shared or released and no human DNA was sequenced or analyzed.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autonoma de México, Cuernavaca, Morelos, Mexico. [2]Molecular and Experimental Mycobacteriology, Research Center Borstel, Borstel, Germany. [3]German Center for Infection Research (DZIF), Partner site Hamburg-Lübeck-Borstel, Borstel, Germany. [4]National and WHO Supranational Reference Center for Mycobacteria, Research Center Borstel, Leibniz Lung Center, Borstel, Germany. [5]Experimental Pathology Section, National Institute of Medical Sciences and Nutrition Salvador Zubirán, Mexico City, Mexico. [6]Institute of Medical Microbiology, Virology and Hospital Hygiene, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.

Received: 3 March 2021 Accepted: 27 April 2021
Published online: 25 May 2021

References
1. Nessar R, Cambau E, Reyrat J-M, Murray A, Gicquel B. Mycobacterium abscessus: a new antibiotic nightmare. J Antimicrob Chemother. 2012;67(4):810–8. https://doi.org/10.1093/jac/dkr578.
2. Lee M-R, Sheng W-H, Hung C-C, Yu C-J, Lee L-N, Hsueh P-R. Mycobacterium abscessus complex infections in humans. Emerg Infect Dis. 2015;21(9):1638–46. https://doi.org/10.3201/2109.141634.
3. Sanguinetti M, Ardito F, Fiscarelli E, La Sorda M, D'Argenio P, Ricciotti G, et al. Fatal pulmonary infection due to multidrug-resistant Mycobacterium abscessus in a patient with cystic fibrosis. J Clin Microbiol. 2001;39(2):816–9. https://doi.org/10.1128/JCM.39.2.816-819.2001.
4. Jarand J, Levin A, Zhang L, Huitt G, Mitchell JD, Daley CL. Clinical and microbiologic outcomes in patients receiving treatment for Mycobacterium abscessus pulmonary disease. Clin Infect Dis. 2011;52(5):565–71. https://doi.org/10.1093/cid/ciq237.
5. Chen J, Zhao L, Mao Y, Ye M, Guo Q, Zhang Y, et al. Clinical efficacy and adverse effects of antibiotics used to treat Mycobacterium abscessus pulmonary disease. Front Microbiol. 2019;10:1977. https://doi.org/10.3389/fmicb.2019.01977.
6. Tortoli E, Kohl TA, Brown-Elliott BA, Trovato A, Leão SC, Garcia MJ, et al. Emended description of Mycobacterium abscessus, Mycobacterium abscessus subsp. abscessus and Mycobacteriumabscessus subsp. bolletii and designation of Mycobacteriumabscessus subsp. massiliense comb. nov. Int J Syst Evol Microbiol. 2016;66(11):4471–9. https://doi.org/10.1099/ijsem.0.001376.
7. Howard ST, Rhoades E, Recht J, Pang X, Alsup A, Kolter R, et al. Spontaneous reversion of Mycobacterium abscessus from a smooth to a rough morphotype is associated with reduced expression of glycopeptidolipid and reacquisition of an invasive phenotype. Microbiol Microbiol Soc. 2006;152:1581–90.
8. Abeles SR, Pride DT. Molecular bases and role of viruses in the human microbiome. J Mol Biol. 2014;426(23):3892–906. https://doi.org/10.1016/j.jmb.2014.07.002.
9. Ripoll F, Deshayes C, Pasek S, Laval F, Beretti J-L, Biet F, et al. Genomics of glycopeptidolipid biosynthesis in Mycobacterium abscessus and M. chelonae. BMC genomics. BioMed Central. 2007;8:114–9.
10. Gutiérrez AV, Viljoen A, Ghigo E, Herrmann J-L, Kremer L. Glycopeptidolipids, a Double-Edged Sword of the Mycobacterium abscessus Complex. Front Microbiol Front. 2018;9:1145.
11. Tjalsma H, Antelmann H, Jongbloed JDH, Braun PG, Darmon E, Dorenbos R, et al. Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome. Microbiol Mol Biol Rev. 2004;68(2):207–33. https://doi.org/10.1128/MMBR.68.2.207-233.2004.
12. Daugelat S, Guile H, Schoel B, Kaufmann SHE. Secreted antigens of Mycobacterium tuberculosis: characterization with T lymphocytes from patients and contacts after two-dimensional separation. J Infect Dis. 1992;166(1):186–90. https://doi.org/10.1093/infdis/166.1.186.
13. Zheng J, Ren X, Wei C, Yang J, Hu Y, Liu L, et al. Analysis of the secretome and identification of novel constituents from culture filtrate of bacillus Calmette-Guerin using high-resolution mass spectrometry. Mol Cell Proteomics. 2013;12(8):2081–95. https://doi.org/10.1074/mcp.M113.027318.
14. Gupta MK, Subramanian V, Yadav JS. Immunoproteomic identification of secretory and subcellular protein antigens and functional evaluation of the secretome fraction of Mycobacterium immunogenum, a newly recognized species of the Mycobacterium chelonae-Mycobacterium abscessus group. J Proteome Res. 2009;8(5):2319–30. https://doi.org/10.1021/pr8009462.
15. Shin A-R, Sohn H, Won CJ, Lee B, Kim WS, Kang HB, et al. Characterization and identification of distinct Mycobacterium massiliense extracellular proteins from those of Mycobacterium abscessus. J Microbiol. 2010;48:502–11 The Microbiological Society of Korea.
16. Yadav JS, Gupta M. Secretome differences between the taxonomically related but clinically differing mycobacterial species Mycobacterium abscessus and M. chelonae. JIOMICS. 2012;2:1–16.
17. Laencina L, Dubois V, Le Moigne V, Viljoen A, Majlessi L, Pritchard J, et al. Identification of genes required for Mycobacterium abscessus growth

in vivo with a prominent role of the ESX-4 locus. Proc Natl Acad Sci U.S.A. 2018;115:E1002–11 National Academy of Sciences.
18. Gomez S, Adalid-Peralta L, Palafox-Fonseca H, Cantu-Robles VA, Soberón X, Sciutto E, et al. Genome analysis of excretory/secretory proteins in Taenia solium reveals their abundance of antigenic regions (AAR). Sci Rep. 2015; 5(1):9683. https://doi.org/10.1038/srep09683.
19. Cornejo-Granados F, Zatarain-Barrón ZL, Cantu-Robles VA, Mendoza-Vargas A, Molina-Romero C, Sánchez F, et al. Secretome Prediction of Two M. tuberculosis clinical isolates reveals their high antigenic density and potential drug targets. Front Microbiol. 2017;8:128.
20. Macheras E, Konjek J, Roux AL, Thiberge JM, Bastian S, Leão SC, et al. Multilocus sequence typing scheme for the Mycobacterium abscessus complex. Res Microbiol. 2014;165(2):82–90. https://doi.org/10.1016/j.resmic.2013.12.003.
21. Wang S, Wei W, Cai X. Genome-wide analysis of excretory/secretory proteins in Echinococcus multilocularis: insights into functional characteristics of the tapeworm secretome. Parasit Vectors. 2015;8(1):666. https://doi.org/10.1186/s13071-015-1282-7.
22. Tortoli E, Fedrizzi T, Meehan CJ, Trovato A, Grottola A, Giacobazzi E, et al. The new phylogeny of the genus Mycobacterium: the old and the news. Infect Genet Evol. 2017;56:19–25. https://doi.org/10.1016/j.meegid.2017.10.013.
23. Malhotra S, Vedithi SC, Blundell TL. Decoding the similarities and differences among mycobacterial species. Yang R, editor. PLoS Negl Trop Dis. 2017;11: e0005883 Public Library of Science.
24. Bachmann NL, Salamzade R, Manson AL, Whittington R, Sintchenko V, Earl AM, et al. Key transitions in the evolution of rapid and slow growing Mycobacteria identified by comparative genomics. Front Microbiol Front. 2019;10:3019.
25. 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. Ahmed N, editor. PLoS ONE. 2009;4:e5660 Public Library of Science.
26. Ryan K, Byrd TF. Mycobacterium abscessus: shapeshifter of the mycobacterial world. Front Microbiol. 2018;9:2642. https://doi.org/10.3389/fmicb.2018.02642.
27. Waman VP, Vedithi SC, Thomas SE, Bannerman BP, Munir A, Skwark MJ, et al. Mycobacterial genomics and structural bioinformatics: opportunities and challenges in drug discovery. Emerg Microbes Infect. 2019;8(1):109–18. https://doi.org/10.1080/22221751.2018.1561158.
28. O'Toole N, Min XJ, Butler G, Storms R, Tsang A. Sequence-based analysis of fungal secretomes. Appl Mycol Biotechnol. 2013;6:277–96 Elsevier B.V.
29. Mukherjee S, Bassler BL. Bacterial quorum sensing in complex and dynamically changing environments. Nat Rev Microbiol Nature Publishing Group. 2019;17:371–82.
30. Orme IM, Ordway DJ. Host response to nontuberculous mycobacterial infections of current clinical importance. Andrews-Polymenis HL, editor. Infect Immun. 2014;82:3516–22 American Society for Microbiology Journals.
31. Clary G, Sasindran SJ, Nesbitt N, Mason L, Cole S, Azad A, et al. Mycobacterium abscessus smooth and rough morphotypes form antimicrobial-tolerant biofilm phenotypes but are killed by acetic acid. Antimicrob Agents Chemother. 2018;62:117 American Society for Microbiology Journals.
32. Kulka K, Hatfull G, Ojha AK. Growth of *Mycobacterium tuberculosis* biofilms. Washington D.C.: JoVE; 2012.
33. Maurer FP, Bruderer VL, Ritter C, Castelberg C, Bloemberg GV, Böttger EC. Lack of antimicrobial bactericidal activity in Mycobacterium abscessus. 2nd ed. Antimicrob Agents Chemothe. 2014;58:3828–36 American Society for Microbiology Journals.
34. Faria S, Joao I, Jordao L. General overview on Nontuberculous mycobacteria, biofilms, and human infection. J Pathog Hindawi. 2015;2015:809014–0.
35. Hunt-Serracin AC, Parks BJ, Boll J, Boutte CC. Mycobacterium abscessus cells have altered antibiotic tolerance and surface glycolipids in artificial cystic fibrosis sputum medium. Antimicrobial Agents Chemother. 2019;63:1370 American Society for Microbiology Journals.
36. Dumas E, Christina Boritsch E, Vandenbogaert M, de la Vega RC R, Thiberge J-M, Caro V, et al. Mycobacterial pan-genome analysis suggests important role of plasmids in the radiation of type VII secretion systems. Genome Biol Evol. 2016;8(2):387–402. https://doi.org/10.1093/gbe/evw001.
37. Kumar A, Chandolia A, Chaudhry U, Brahmachari V, Bose M. Comparison of mammalian cell entry operons of mycobacteria: in silico analysis and expression profiling. FEMS Immunol Med Microbiol. 2005;43(2):185–95. https://doi.org/10.1016/j.femsim.2004.08.013.
38. Catherinot E, Clarissou J, Etienne G, Ripoll F, Emile JF, Daffé M, et al. Hypervirulence of a rough variant of the Mycobacterium abscessus type strain. Infect Immun. 2007;75(2):1055–8. https://doi.org/10.1128/IAI.00835-06.
39. Roux A-L, Viljoen A, Bah A, Simeone R, Bernut A, Laencina L, et al. The distinct fate of smooth and rough Mycobacterium abscessus variants inside macrophages. Open Biol. 2016;6:160185 The Royal Society.
40. Bernut A, Herrmann J-L, Kissa K, Dubremetz J-F, Gaillard J-L, Lutfalla G, et al. Mycobacterium abscessus cording prevents phagocytosis and promotes abscess formation. Proc Natl Acad Sci U.S.A. 2014;111:E943–52 National Academy of Sciences.
41. Maurer F, Castelberg C, Braun von A, Wolfensberger A, Bloemberg G, Bottger E, et al. Postsurgical wound infections due to rapidly growing mycobacteria in Swiss medical tourists following cosmetic surgery in Latin America between 2012 and 2014. Euro Surveill. 2014;19:20905 European Centre for Disease Prevention and Control.
42. Adekambi T, Colson P, Drancourt M. rpoB-based identification of nonpigmented and late-pigmenting rapidly growing mycobacteria. J Clin Microbiol. 2003;41:5699–708 American Society for Microbiology Journals.
43. De Almeida IN, Da Silva CW, Rossetti ML, Costa ERD, De Miranda SS. Evaluation of six different DNA extraction methods for detection of Mycobacterium tuberculosis by means of PCR-IS6110: preliminary study. BMC Res Notes. 2013;6(1):561–6. https://doi.org/10.1186/1756-0500-6-561 BioMed Central.
44. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
45. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kilikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single cell sequencing. J Comput Biol. 2012;19(5):455–77. https://doi.org/10.1089/cmb.2012.0021.
46. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics. 2008;9(1):75. https://doi.org/10.1186/1471-2164-9-75.
47. Petersen TN, Brunak S, Heijne von G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6 Nature Publishing Group.
48. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. BMC Microbiol. 2005;5(1):58. https://doi.org/10.1186/1471-2180-5-58.
49. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S. Prediction of twin-arginine signal peptides. BMC Bioinformatics. 2005;6(1):167. https://doi.org/10.1186/1471-2105-6-167.
50. Sonnhammer EL, Heijne von G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998;6:175–82.
51. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res. 2007;35(Web Server):W429–32. https://doi.org/10.1093/nar/gkm256.
52. Conesa A, Götz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008;2008:619832.
53. Zdobnov EM, Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17(9):847–8. https://doi.org/10.1093/bioinformatics/17.9.847.
54. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35(Web Server):W182–5. https://doi.org/10.1093/nar/gkm321.
55. Cornejo-Granados F, Hurtado-Ramírez JM, Hernandez-Pando R, Ochoa-Leyva A. Secret-AAR: a web server to assess the antigenic density of proteins and homology search against bacterial and parasite secretome proteins. Genomics. 2019;111(6):1514–6. https://doi.org/10.1016/j.ygeno.2018.10.007.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Genomics 111 (2019) 1514–1516

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Short Communication

Secret-AAR: a web server to assess the antigenic density of proteins and homology search against bacterial and parasite secretome proteins

Fernanda Cornejo-Granados[a,1], Juan Manuel Hurtado-Ramírez[a,1], Rogelio Hernández-Pando[b], Adrián Ochoa-Leyva[a,*]

[a] Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Avenida Universidad 2001, Colonia Chamilpa, Cuernavaca, Morelos 62210, Mexico
[b] Experimental Pathology Section, National Institute of Medical Sciences and Nutrition "Salvador Zubirán", Mexico City 14000, Mexico

ARTICLE INFO

Keywords:
Antigenicity
Excreted-secreted proteins
AAR value
Immunoinformatics
Antigenic density

ABSTRACT

The secretome refers to all the Excreted/Secreted (ES) proteins of a cell, and these are involved in critical biological processes, such as cell-cell communication, and host immune responses. Recently, we introduced the Abundance of Antigenic Aegions (AAR) value to assess the protein antigenic density and to evaluate the antigenic potential of secretomes. Here, to facilitate the AAR calculation, we implemented it as a user-friendly webserver. We extended the webserver capabilities implementing a sequence-based tool for searching homologous proteins across secretomes, including experimental and predicted secretomes of *Mycobacterium tuberculosis* and *Taenia solium*. Additionally, twelve secretomes of helminths, five of Mycobacterium and two of Gram-negative bacteria are also available. Our webserver is a useful tool for researchers working on immunoinformatics and reverse vaccinology, aiming at discovering candidate proteins for new vaccines or diagnostic tests, and it can be used to prioritize the experimental analysis of proteins for druggability assays. The Secret-AAR web server is available at http://microbiomics.ibt.unam.mx/tools/aar/.

1. Introduction

The secretome is defined as the complete set of Excreted/Secreted (ES) proteins of a cell [1]. These proteins are involved in critical biological processes, such as mechanisms of adhesion, cell migration and invasion, cell-to-cell communication, signal transduction and in pathogenic bacteria the secretome plays a crucial role in parasitism modulating the host immune response and promoting the proliferation of infection [2–4]. Moreover, due to their direct exposure to the host immune system and their high antigenic density [5–8], the ES proteins are a vital source of immunogenic proteins, useful for vaccine development, druggability and diagnostic assays [9–11] [12]. The Abundance of Antigenic Regions (AAR) was the first metric developed to estimate and compare the antigenic density of a protein [5]. Our group previously published the predicted secretomes for the reference strains of the extracellular *T. solium* and the intracellular *M. tuberculosis* human pathogens [5,6], and applying the AAR we demonstrated that secretomes of both pathogens have more antigenic density than the non-secreted proteins of their genomes. The AAR has been rapidly accepted by other research groups to identify antigenic proteins for druggability assays against host-pathogen interactions [6], to evaluate the antigenicity of universal peptides for neglected tropical diseases [13] and to correlate the antigenic density with the immunogenic properties of proteins [14]. The AAR was also applied to demonstrate that the secretome is significantly more antigenic than the complete proteome of organisms such as *Taenia crassiceps* [8], *Echinococcus multilocularis* [7], and several *Aspergillus* species [15]. Currently, no web server or software allows obtaining the AAR directly from the fasta sequence in a user-friendly manner. Thus, to make the AAR more accessible to a broader community, we report the Secret-AAR web server for the automatic AAR calculation of any protein at a genome-scale level.

Three predicted secretomes have been reported for *M. tuberculosis* [6,16,17]. Additionally, our group published an experimental secretome compiled from a literature search [6]. To the best of our knowledge, the *M. tuberculosis* secretome published by our group is the most accurate prediction, demonstrating 70% of coincidence with the experimental secretome compiled from literature [6]. Also, our group established the bioinformatics strategy to predict the secretome of *T. solium* and validated it against transcriptomic data [5]. Now, to facilitate the access to this data, we also included in the Secret-AAR

[*] Corresponding author.
E-mail address: aochoa@ibt.unam.mx (A. Ochoa-Leyva).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ygeno.2018.10.007
Received 25 July 2018; Received in revised form 10 October 2018; Accepted 10 October 2018
Available online 11 October 2018

webserver the option to identify (via BLAST) if a protein belongs to the most comprehensive experimental and predicted collection of secretomes for *T. solium* and *M. tuberculosis* [5,6]. These two human pathogens which still represent a significant problem in worldwide public health and part of their success on infection relies on the production of ES proteins for the modulation of the host immune system. Additionally, we included other 19 secretome databases for searching homologous proteins across different bacterial and parasite species. Twelve secretomes are of helminths, five of Mycobacterium strains and two of Gram-negative bacteria.

2. Methods

2.1. Web interface

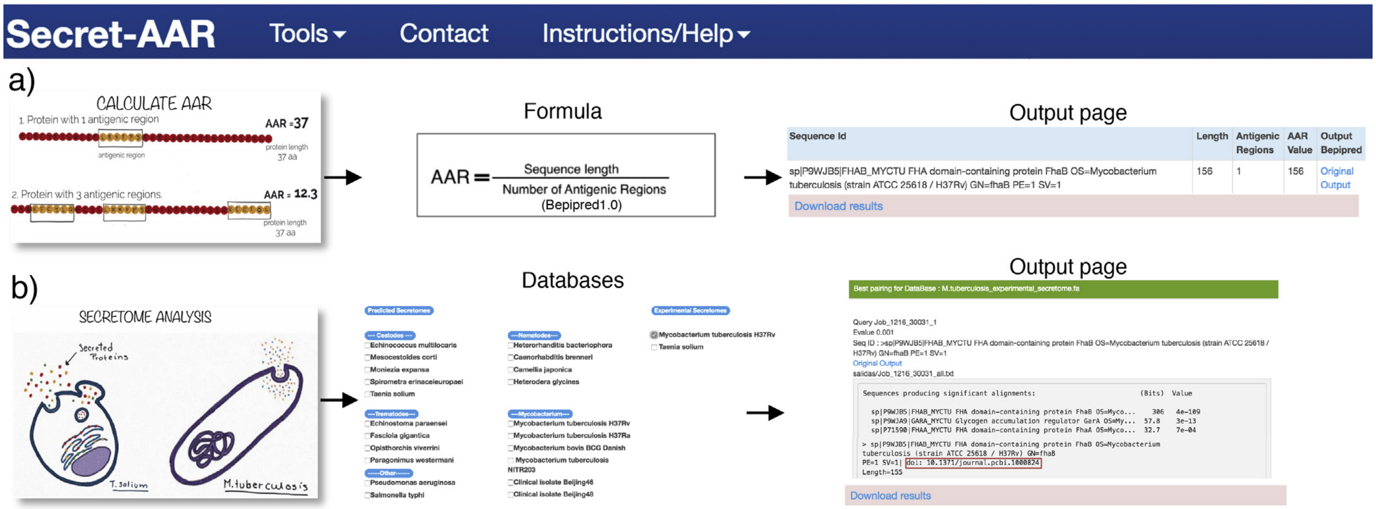
The web server has two input pages: one to calculate the AAR value (AAR value input page) and another to analyze if the protein of interest is part of the experimental and/or predicted secretome of some relevant human pathogens (Secretome analysis input page.) To use Secret-AAR (http://microbiomics.ibt.unam.mx/tools/aar/index.php) the user only needs to choose the desired tool from the home page or the drop-down menu on the header and submit the protein sequence or set of sequences in fasta format, either by pasting them into the text box or by uploading them as a text file. The web server supports up to 1000 protein queries per request with a maximum length of 5000 amino acids in a single sequence. Additionally, the user can find examples of sequences by clicking the 'Example Sequences' button and by clicking on the "Download Results" button the user can retrieve the results in a plain text format to save the file locally. Further, in the Instructions/Help option, the users can access a tutorial with a more detailed description of the web server. After closing the window of the web server all submitted information and results are automatically deleted.

2.2. The AAR value input page

The AAR value is the ratio between the sequence length to the number of predicted antigenic regions [5] (Fig. 1a). The number of antigenic regions for each protein is calculated using the Bepipred1.0 [18] algorithm, setting the threshold score to 0.35, and considering only the epitopes with at least six amino acids long, as was previously suggested [19]. Hence, low AAR values mean that a protein has more epitope density (more antigenic regions) and vice versa. To obtain the AAR value, Secret-AAR uses an in-house developed script to identify the number of antigenic regions with a minimum of six amino acids of the Bepipred output and also identify the sequence length of the query protein. After that, our script calculates the AAR value through normalization of the number of antigenic regions by sequence length of each protein. Importantly, the algorithm established in the webserver, does not calculate the AAR if the protein has no antigenic regions that cover the minimum window size of 6 amino acids. The output page contains a table showing the length, the number of antigenic regions, and the AAR value for each sequence (Fig. 1a). Also, each sequence has an "Original Output" link that the user can use to access the original Bepipred1.0 prediction results and visualize the location and specific length of the potential antigenic regions identified.

2.3. Secretome analysis input page

To help guiding the search for potential drug targets or diagnostic molecules against some important human pathogens, we included in the web server the option to analyze if the protein of interest has been reported as an experimental or predicted ES protein [5,6] (Fig. 1b). To this end, Secret-AAR takes the input sequence and BLAST it against the predicted and experimental secretomes obtained for several species (Fig. 1b). Additionally to the predicted and experimental secretomes previously reported for the reference strains of *M. tuberculosis* and *T. solium* [5,6], we included the predicted secretomes for other 19 organisms. Two of *M. tuberculosis* clinical isolates Beijing46 and Beijing48, also reported by our group [6] and twelve secretomes for cestodes, trematodes, and nematodes obtained from the Helminth Secretome database [20]. Also, in this work, we obtained the predicted secretomes for three *M. tuberculosis* strains H37Ra, BCG Danish, and NITR203 and for other two important human pathogens, *Pseudomonas aeruginosa* and *Salmonella typhi*, all of which were obtained using the bioinformatics pipeline previously published [6]. The user can select one or more databases from this list to analyze their proteins at the same time. The output page contains the blast results including the alignment, sequence identity and coverage between the protein of interest and the best hits for the selected database. Also, if the matching sequence has experimental support as ES protein, the DOI of the original article is included in the final section of the protein header (Fig. 1b).



$$AAR = \frac{\text{Sequence length}}{\text{Number of Antigenic Regions (Bepipred1.0)}}$$

Sequence Id	Length	Antigenic Regions	AAR Value	Output Bepipred
sp|P9WJB5|FHAB_MYCTU FHA domain-containing protein FhaB OS=Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv) GN=fhaB PE=1 SV=1	156	1	156	Original Output

Download results

Predicted Secretomes

--- Cestodes ---
☐ Echinococcus multilocaris
☐ Mesocestoides corti
☐ Moniezia expansa
☐ Spirometra erinaceieuropaei
☐ Taenia solium

---Trematodes---
☐ Echinostoma paraensei
☐ Fasciola gigantica
☐ Opisthorchis viverrini
☐ Paragonimus westermani

------Other------
☐ Pseudomonas aeruginosa
☐ Salmonella typhi

---Nematodes---
☐ Heterorhanditis bacteriophora
☐ Caenorhabditis brenneri
☐ Camellia japonica
☐ Heterodera glycines

---Mycobacterium---
☐ Mycobacterium tuberculosis H37Rv
☐ Mycobacterium tuberculosis H37Ra
☐ Mycobacterium bovis BCG Danish
☐ Mycobacterium tuberculosis NITR203
☐ Clinical isolate Beijing46
☐ Clinical isolate Beijing48

Experimental Secretomes

☑ Mycobacterium tuberculosis H37Rv
☐ Taenia solium

Best pairing for DataBase : M.tuberculosis_experimental_secretome.fa

Query Job_1216_30031_1
Evalue 0.001
Seq ID : >sp|P9WJB5|FHAB_MYCTU FHA domain-containing protein FhaB OS=Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv) GN=fhaB PE=1 SV=1
Original Output
salidas/Job_1216_30031_all.txt

```
Sequences producing significant alignments:                          (Bits)  Value

  sp|P9WJB5|FHAB_MYCTU FHA domain-containing protein FhaB OS=Myco...    306   4e-109
  sp|P9WJA9|GARA_MYCTU Glycogen accumulation regulator GarA OS=My...   57.8    3e-13
  sp|P71590|FHAA_MYCTU FHA domain-containing protein FhaA OS=Myco...   32.7    7e-04

> sp|P9WJB5|FHAB_MYCTU FHA domain-containing protein FhaB OS=Mycobacterium
tuberculosis (strain ATCC 25618 / H37Rv) GN=fhaB
PE=1 SV=1| doi: 10.1371/journal.pcbi.1000824
Length=155
```

Download results

Fig. 1. Basic workflow of the Secret-AAR web server.
a) The input sequence is used for the calculation of the AAR value, and the output includes protein length, number of antigenic regions, AAR value and the original output of Bepipred in which the antigenic region can be observed directly in the sequence. b) The input sequence is used for searching homologous proteins across 23 secretomes using BLASTP. The output includes the typical BLAST results, and if the matching sequence has experimental support as ES protein, the DOI of the original article is included in the final section of the protein header.

3. Discussion and conclusions

Secret-AAR is a user-friendly tool to obtain the AAR value of any protein at a genome-scale level in an automatic manner. This metric can be used to prioritize the experimental analysis of potential protein targets for vaccine development, druggability, and biomarker candidates. Additionally, our web server can be used for searching homologous proteins across secretomes, including the most comprehensive predicted and experimental collection of secretomes for *T. solium* and *M. tuberculosis*. This webserver is under continuous improvement, thus, additional secretome databases can become available in the future.

Secret-AAR is available at http://microbiomics.ibt.unam.mx/tools/aar/. The scripts to calculate the AAR and secretome databases are freely available as an open-source and can be downloaded from https://github.com/8aLab/Web-Server-Secret-AAR.

Author's contributions

FCG and JMHR developed the server and performed the analyses. RHP discussed the manuscript and AOL developed the idea. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgements

This work was supported by the Consejo Nacional de Ciencia y Tecnología (CONACyT) of Mexico [SALUD-2014-C01-234188, CB-2013-223279]. F.C.G. acknowledges the support of CONACyT as a doctoral fellow.

References

[1] S. Ranganathan, G. Garg, Secretome: clues into pathogen infection and clinical applications, Genome Med. 1 (2009) 113, https://doi.org/10.1186/gm113.

[2] H. Tjalsma, H. Antelmann, J.D.H. Jongbloed, P.G. Braun, E. Darmon, R. Dorenbos, et al., Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome, Microbiol. Mol. Biol. Rev. 68 (2004) 207–233, https://doi.org/10.1128/MMBR.68.2.207-233.2004.

[3] A.G. Chande, Z. Siddiqui, M.K. Midha, V. Sirohi, S. Ravichandran, K.V.S. Rao, Selective enrichment of mycobacterial proteins from infected host macrophages, Sci. Rep. 5 (2015) 13430, , https://doi.org/10.1038/srep13430.

[4] F. Vargas-Romero, N. Guitierrez-Najera, G. Mendoza-Hernández, D. Ortega-Bernal, R. Hernandez-Pando, M. Castañón-Arreola, Secretome profile analysis of hypervirulent Mycobacterium tuberculosis CPT31 reveals increased production of EsxB and proteins involved in adaptation to intracellular lifestyle, Pathogen. Dis. 74 (2016), https://doi.org/10.1093/femspd/ftv127.

[5] S. Gomez, L. Adalid-Peralta, H. Palafox-Fonseca, V.A. Cantu-Robles, X. Soberón, E. Sciutto, et al., Genome analysis of Excretory/Secretory proteins in Taenia solium reveals their Abundance of Antigenic Regions (AAR), Sci. Rep. 5 (2015) 9683, , https://doi.org/10.1038/srep09683.

[6] F. Cornejo-Granados, Z.L. Zatarain-Barrón, V.A. Cantu-Robles, A. Mendoza-Vargas, C. Molina-Romero, F. Sánchez, et al., Secretome prediction of two M. tuberculosis clinical isolates reveals their high antigenic density and potential drug targets, Front. Microbiol. 8 (2017) 128, , https://doi.org/10.3389/fmicb.2017.00128.

[7] S. Wang, W. Wei, X. Cai, Genome-wide analysis of excretory/secretory proteins in Echinococcus multilocularis: insights into functional characteristics of the tapeworm secretome, Parasit. Vectors 8 (2015) 666, , https://doi.org/10.1186/s13071-015-1282-7.

[8] G.M. García-Montoya, J.A. Mesa-Arango, J.P. Isaza-Agudelo, S.P. Agudelo-Lopez, F. Cabarcas, L.F. Barrera, et al., Transcriptome profiling of the cysticercus stage of the laboratory model Taenia crassiceps, strain ORF, Acta Trop. 154 (2016) 50–62, https://doi.org/10.1016/j.actatropica.2015.11.001.

[9] H. Målen, F.S. Berven, K.E. Fladmark, H.G. Wiker, Comprehensive analysis of exported proteins from Mycobacterium tuberculosis H37Rv, Proteomics 7 (2007) 1702–1718, https://doi.org/10.1002/pmic.200600853.

[10] J. Zheng, X. Ren, C. Wei, J. Yang, Y. Hu, L. Liu, et al., Analysis of the secretome and identification of novel constituents from culture filtrate of bacillus Calmette-Guerin using high-resolution mass spectrometry, Mol. Cell. Proteomics 12 (2013) 2081–2095, https://doi.org/10.1074/mcp.M113.027318.

[11] M. Niederweis, O. Danilchanka, J. Huff, C. Hoffmann, H. Engelhardt, Mycobacterial outer membranes: in search of proteins, Trends Microbiol. 18 (2010) 109–116, https://doi.org/10.1016/j.tim.2009.12.005.

[12] P. Tucci, G. González-Sapienza, M. Marin, Pathogen-derived biomarkers for active tuberculosis diagnosis, Front. Microbiol. 5 (2014) 549, , https://doi.org/10.3389/fmicb.2014.00549.

[13] S. Miles, M. Navatta, S. Dematteis, G. Mourglia-Ettlin, Identification of universal diagnostic peptide candidates for neglected tropical diseases caused by cestodes through the integration of multi-genome-wide analyses and immunoinformatic predictions, Infect. Genet. Evol. 54 (2017) 338–346, https://doi.org/10.1016/j.meegid.2017.07.020.

[14] R.J. Bobes, J. Navarrete-Perea, A. Ochoa-Leyva, V.H. Anaya, M. Hernández, J. Cervantes-Torres, et al., Experimental and theoretical approaches to investigate the immunogenicity of Taenia solium-derived KE7 antigen, Infect. Immun. 85 (2017), https://doi.org/10.1128/IAI.00395-17 e00395–17.

[15] R.P. Vivek-Ananth, K. Mohanraj, M. Vandanashree, A. Jhingran, J.P. Craig, A. Samal, Comparative systems analysis of the secretome of the opportunistic pathogen Aspergillus fumigatus and other Aspergillus species, Sci. Rep. 8 (2018) 296, , https://doi.org/10.1038/s41598-018-25016-4.

[16] A. Roy, S. Bhattacharya, A.K. Bothra, A. Sen, A database for Mycobacterium secretome analysis: "MycoSec" to accelerate global health research, OMICS 17 (2013) 502–509, https://doi.org/10.1089/omi.2013.0015.

[17] C. Vizcaíno, D. Restrepo-Montoya, D. Rodríguez, L.F. Niño, M. Ocampo, M. Vanegas, et al., Computational prediction and experimental assessment of secreted/surface proteins from Mycobacterium tuberculosis H37Rv, PLoS Comput. Biol. 6 (2010) e1000824, , https://doi.org/10.1371/journal.pcbi.1000824.

[18] J.E.P. Larsen, O. Lund, M. Nielsen, Improved method for predicting linear B-cell epitopes, Immun. Res. 2 (2006) 2, , https://doi.org/10.1186/1745-7580-2-2.

[19] L. Berglund, J. Andrade, J. Odeberg, M. Uhlén, The epitope space of the human proteome, Protein Sci. 17 (2008) 606–613, https://doi.org/10.1110/ps.073347208.

[20] G. Garg, S. Ranganathan, Helminth secretome database (HSD): a collection of helminth excretory/secretory proteins predicted from expressed sequence tags (ESTs), BMC Genomics 13 (Suppl. 7) (2012) S8, https://doi.org/10.1186/1471-2164-13-S7-S8.