TESIS DE LA UNIVERSIDAD DE ZARAGOZA
2023
130

Jessica Comín Polo

Caracterización molecular con técnicas de secuenciación masiva para el estudio de transmisión y evolución de *mycobacterium tuberculosis complex* en Aragón

Director/es
Samper Blasco, Sofía Luisa

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

CARACTERIZACIÓN MOLECULAR CON TÉCNICAS DE SECUENCIACIÓN MASIVA PARA EL ESTUDIO DE TRANSMISIÓN Y EVOLUCIÓN DE *MYCOBACTERIUM TUBERCULOSIS COMPLEX* EN ARAGÓN

Autor

Jessica Comín Polo

Director/es

Samper Blasco, Sofía Luisa

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Bioquímica y Biología Molecular

2023

CARACTERIZACIÓN MOLECULAR CON TÉCNICAS DE SECUENCIACIÓN MASIVA PARA EL ESTUDIO DE TRANSMISIÓN Y EVOLUCIÓN DE *Mycobacterium tuberculosis* complex EN ARAGÓN

Departamento de Bioquímica y Biología Molecular y Celular
Facultad de Ciencias

JESSICA COMÍN POLO

Tesis Doctoral 2023

Universidad Zaragoza
1542

Caracterización molecular con técnicas de secuenciación masiva para el estudio de transmisión y evolución de *Mycobacterium tuberculosis* complex en Aragón

Memoria presentada por:

Jessica Comín Polo

Para optar al título de Doctor por la Universidad de Zaragoza

Dirigida por:

Dra. Sofía Samper Blasco

IACS Instituto Aragonés de Ciencias de la Salud

GGM GRUPO DE GENÉTICA DE MICOBACTERIAS

1542

Departamento de Bioquímica y Biología Molecular y Celular

Universidad Zaragoza

«No se trata de buscar nuevas soluciones,

sino de plantear nuevas preguntas»

Albert Einstein

Dra. Sofía Samper Blasco, investigadora del Instituto Aragonés de Ciencias de la Salud en la Unidad de Investigación Traslacional del Hospital Universitario Miguel Servet,

Directora de la Tesis Doctoral de **JESSICA COMÍN POLO**, titulada:

Caracterización molecular con técnicas de secuenciación masiva para el estudio de transmisión y evolución de *Mycobacterium tuberculosis* complex en Aragón

Certifica que esta Tesis Doctoral ha sido realizada bajo su dirección, de acuerdo a los objetivos presentados en el Proyecto de Tesis aprobado por el Departamento de Bioquímica y Biología Molecular y Celular y que reúne los requisitos necesarios para optar al título de Doctor por la Universidad de Zaragoza, por lo que autoriza su presentación en la modalidad de compendio de publicaciones.

Zaragoza, 9 de enero de 2023

Sofía Samper Blasco

Esta tesis doctoral ha sido realizada en la Unidad de Investigación Traslacional del Hospital Universitario Miguel Servet, perteneciente al Instituto de Investigación Sanitaria de Aragón. Se ha elaborado adscrita al programa de Doctorado del Departamento de Bioquímica y Biología Molecular y Celular, siendo Jessica Comín Polo beneficiaria de un contrato predoctoral de la Diputación General de Aragón, convocadas en la ORDEN IIU/2023/2017 y cofinanciadas con el Programa Operativo FSE Aragón 2014-2020, "construyendo Europa desde Aragón".

Además, el trabajo se ha realizado en el contexto de dos proyectos de investigación:

- FIS15/0317 del Instituto de Salud Carlos III: Análisis de las diferencias de IS*6110* entre los miembros del complejo *Mycobacterium tuberculosis* y el papel de su localización en el origen de replicación.
- FIS18/0336 del Instituto de Salud Carlos III: Nuevas herramientas y estrategias para controlar la TB.

Agradecimientos

Quiero agradecer especialmente a Sofía toda su dedicación y apoyo, así como su cariño e implicación, ya que me han hecho el camino mucho más fácil y llevadero. ¡Eres la mejor!

Agradecer la acogida en el grupo de Genética de Micobacterias y en especial al grupo EPIMOLA, por todo el trabajo de base que ha servido para la elaboración de esta tesis. Agradecer a Carmen, del Servicio General de Apoyo a la Investigación (SAI-Unizar), toda su paciencia, trabajo y los conocimientos que me ha transmitido. También mi agradecimiento a Dani, dispuesto siempre a rebuscar en las bases de datos la información epidemiológica solicitada. A la unidad de Micobacterias del Servicio de Microbiología del Hospital Miguel Servet, especialmente a Jesús, por enseñarme los intríngulis del día a día del servicio, ¡muchas gracias!

Mi agradecimiento a todo el personal de los Servicios de Genómica del CIBA, donde se ha llevado a cabo la mayor parte de la secuenciación de esta tesis doctoral. También quiero agradecer a la Unidad de Biocomputación, especialmente a Alberto, por hacerme la vida más fácil con la estadística y la informática.

Finalmente, agradecer a mi familia y amigos todo su apoyo y el haberme animado a seguir. ¡Sois los mejores!

Presentación por compendio de publicaciones

La presente tesis doctoral es un compendio de las siguientes publicaciones:

Publicación 1:

Comín J, Monforte ML, Samper S; Aragonese Working Group on Molecular Epidemiology of Tuberculosis (EPIMOLA), Otal I. **Analysis of *Mycobacterium africanum* in the last 17 years in Aragon identifies a specific location of IS*6110* in Lineage 6**. Sci Rep. 2021 May 14;11(1):10359. doi: 10.1038/s41598-021-89511-x. PMID: 33990628; PMCID: PMC8121931.

Factor de impacto de la revista *Scientific Reports* (2021): **4,997**

Q2 (Área MULTIDISCIPLINARY SCIENCES)

Publicación 2:

Comín J, Otal I, Samper S. **In-depth Analysis of IS*6110* Genomic Variability in the *Mycobacterium tuberculosis* Complex**. Front Microbiol. 2022 Feb 24;13:767912. doi: 10.3389/fmicb.2022.767912. PMID: 35283840; PMCID: PMC8912993.

Factor de impacto de la revista *Frontiers in Microbiology* (2021): **6,064**

Q1 (Área MICROBIOLOGY)

Publicación 3:

Comín J, Cebollada A; Aragonese Working Group on Molecular Epidemiology of Tuberculosis (EPIMOLA), Samper S. **Estimation of the mutation rate of *Mycobacterium tuberculosis* in cases with recurrent tuberculosis using whole genome sequencing**. Sci Rep. 2022 Oct 6;12(1):16728. doi: 10.1038/s41598-022-21144-0. PMID: 36202945; PMCID: PMC9537313.

Factor de impacto de la revista *Scientific Reports* (2021): **4,997**

Q2 (Área MULTIDISCIPLINARY SCIENCES)

Publicación 4:

Comin J, Chaure A, Cebollada A, Ibarz D, Viñuelas J, Vitoria MA, Iglesias MJ, Samper S. **Investigation of a rapidly spreading tuberculosis outbreak using whole-genome sequencing**. Infect Genet Evol. 2020 Jul;81:104184. doi: 10.1016/j.meegid.2020.104184. Epub 2020 Jan 10. PMID: 31931260.

Factor de impacto de la revista *Infection, Genetics and Evolution* (2020): **3,342**

Q3 (Área INFECTIOUS DISEASES)

Publicación 5:

Comín J, Cebollada A, Ibarz D, Viñuelas J, Vitoria MA, Iglesias MJ, Samper S. **A whole-genome sequencing study of an X-family tuberculosis outbreak focus on transmission chain along 25 years**. Tuberculosis (Edinb). 2021 Jan;126:102022. doi: 10.1016/j.tube.2020.102022. Epub 2020 Nov 28. PMID: 33341027.

Factor de impacto de la revista *Tuberculosis* (2021): **2,973**

Q3 (Área MICROBIOLOGY)

Publicación 6:

Comín J, Madacki J, Rabanaque I, Zúñiga-Antón M, Ibarz D, Cebollada A, Viñuelas J, Torres L, Sahagún J, Klopp C, Gonzalo-Asensio J, Brosch R, Iglesias MJ, Samper S. **The MtZ Strain: Molecular Characteristics and Outbreak Investigation of the Most Successful *Mycobacterium tuberculosis* Strain in Aragon Using Whole-Genome Sequencing**. Front Cell Infect Microbiol. 2022 May 24;12:887134. doi: 10.3389/fcimb.2022.887134. PMID: 35685752; PMCID: PMC9173592.

Factor de impacto de la revista *Frontiers in Cellular and Infection Microbiology* (2021): **6,073**
Q1 (Área MICROBIOLOGY)

Publicación 7:

Comín J, Cebollada A, Ibarz D, Viñuelas J, Sahagún J, Torres L, Iglesias MJ, Samper S. **Analysis of the twenty-six largest outbreaks of tuberculosis in Aragon using whole-genome sequencing for surveillance purposes**. Sci Rep. 2022 Nov 5;12(1):18766. doi: 10.1038/s41598-022-23343-1. PMID: 36335223; PMCID: PMC9637126.

Factor de impacto de la revista *Scientific Reports* (2021): **4,997**
Q2 (Área MULTIDISCIPLINARY SCIENCES)

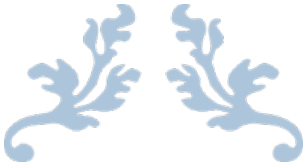
El factor de impacto y el quartil han sido extraídos del Journal Citation Report®

Trabajo 8:

AmpliSeq technology for rapid lineage and drug-resistance identification in clinical samples of *Mycobacterium tuberculosis*. Jessica Comín, Jesús Viñuelas, Carmen Lafoz, Alberto Cebollada, Daniel Ibarz, María-José Iglesias and Sofía Samper.

Enviado para su publicación.

Índice



INTRODUCCIÓN

1. Introducción

1.1. Historia de la tuberculosis

La tuberculosis (TB) ha tenido numerosos nombres a lo largo de la historia: tisis, consunción, *la plaga blanca*, *la ladrona de juventud*, etc. Sea como fuere, *Mycobacterium tuberculosis*, la bacteria causante de esta enfermedad, ha matado a más personas que ningún otro patógeno microbiano (1) (Figura 1). Se ha hipotetizado que el género *Mycobacterium* surgió hace más de 150 millones de años (2), que un progenitor temprano de *M. tuberculosis* habría infectado homínidos en África hace tres millones de años (3) y que el ancestro común de las cepas modernas de *M. tuberculosis* habría aparecido por primera vez hace entre 20000 y 15000 años (4,5). En momias egipcias de hace más de cuatro mil años de antigüedad se han encontrado deformidades esqueléticas típicas de la TB, como las causantes por la enfermedad de Pott (6), y los primeros registros escritos que describen la enfermedad datan de hace 3300 y 2300 años en la India y en China, respectivamente (7).

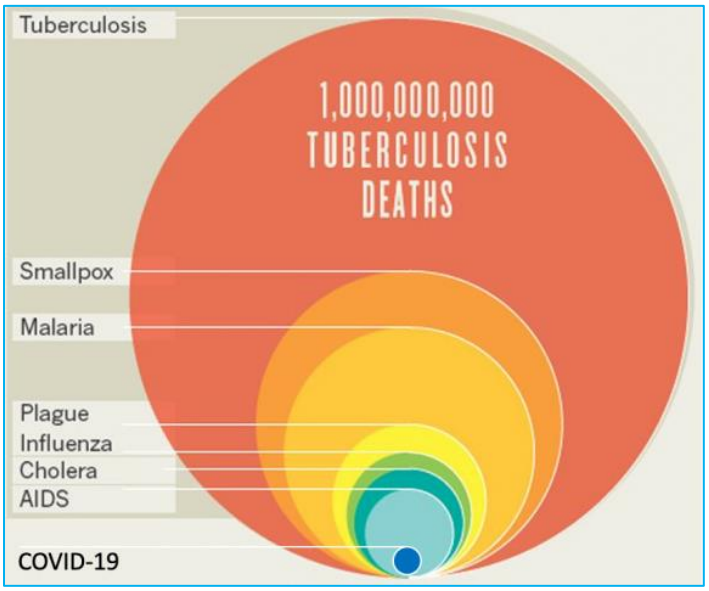


Figura 1. Muertes causadas por *M. tuberculosis* frente a otros microorganismos en los dos últimos siglos. *El círculo de la COVID-19 sería más grande a fecha del depósito de esta tesis (6.675.372 fallecidos a fecha de 23 de diciembre de 2022 según la Universidad John Hopkins).

En la Grecia Antigua la TB era conocida como tisis. Hipócrates la describió así: «La tisis ataca principalmente entre los 18 y 35 años de edad. La consunción fue la enfermedad más importante que entonces prevaleció, y la única que ha demostrado ser fatal para muchas personas». Hipócrates fue el primer autor que atribuyó a la TB una causa infecciosa (8) y

Galeno, médico del emperador romano Marco Aurelio durante el siglo II, describió los síntomas de la TB (fiebre, sudoración, tos y esputo sanguinolento) y recomendaba para tratarla aire fresco, leche y viajes por mar (9). Tras la caída del Imperio Romano la TB se expandió por Europa entre los siglos VIII y XIX (10).

La descripción patológica y anatómica de la enfermedad fue referida por Francis Sylvius en su libro *Opera Medica* (1679), en el que describía los tubérculos, su progresión a abscesos, cavidades y empiema en los pulmones y otros órganos de pacientes afectados de consunción (11). En el siglo XVIII, a causa de la Revolución Industrial y sus condiciones de vida asociadas como malas instalaciones de trabajo, hogares superpoblados y mal ventilados, una sanidad muy primitiva y malnutrición, la TB se había convertido en epidémica en el oeste de Europa, con una mortalidad de 900 por cada 100000 habitantes (12). En 1839 Johann Lukas Schölein acuñó el término tuberculosis para sustituir los de consunción y plaga blanca, usados durante los siglos XVII y XVIII (7). La naturaleza infecciosa de la enfermedad fue demostrada por Jean-Antoine Villemin en 1865: observó que los soldados que se quedaban en los barracones durante más tiempo desarrollaban TB con más frecuencia que los que estaban en el campo. Para probar su teoría, consiguió infectar un conejo con «una pequeña cantidad de líquido purulento procedente de una cavidad tuberculosa» que había extraído de un individuo fallecido por TB. Finalmente, en 1882, Robert Koch consiguió identificar, aislar y cultivar el bacilo causante de la enfermedad, bautizado como bacilo de Koch en su honor (13). Los estudios de Koch sobre TB son un hito en la historia de la medicina, ya que inauguraron la era de la bacteriología en la higiene y supusieron la medicalización de sociedades enteras a finales del siglo XIX (14).

En la segunda mitad del siglo XIX existía la creencia generalizada de que la TB se podía curar con aire fresco, ejercicio y una buena nutrición. Este fue el principio de los sanatorios, el primero de los cuales fue fundado en 1854 en Alemania. Sin embargo, los pacientes graves apenas mejoraban en estos centros (14). Otra estrategia empleada fue el neumotórax terapéutico, llevado a cabo con éxito por primera vez en 1834 consiguiendo una curación completa del paciente (15).

En 1900, Albert Calmette y Camille Guérin comenzaron sus estudios para desarrollar una vacuna utilizando *M. bovis*, la bacteria causante de la TB bovina. Tras 230 pases seriados consiguieron una cepa de *M. bovis* atenuada, Bacillus Calmett y Guerín "BCG". La vacuna se administró por primera vez en 1921 y hoy continúa siendo la única vacuna para la prevención de la TB en humanos (16). El descubrimiento de los antibióticos durante el siglo XX supuso un cambio de paradigma para la TB. El primer antituberculoso efectivo fue la estreptomicina (1943), seguido de la isoniazida (1951), pirazinamida (1952), cicloserina (1952), etionamida (1956), rifampicina (1957) y etambutol (1962). A pesar de un horizonte tan prometedor, algunos estudios que se llevaron a cabo en los años 50 ya anticiparon los problemas de resistencias que se están sufriendo en la actualidad (17).

La Organización Mundial de la Salud (OMS) se ha comprometido a erradicar la TB para el año 2050. Sin embargo, a día de hoy continúa siendo un gran problema de salud a nivel

global, por lo que para conseguir ese objetivo son necesarias estrategias combinadas basadas en la mejora del tratamiento, del diagnóstico y de la estrategia de prevención (18).

1.2. Situación actual de la TB

1.2.1. En el mundo

La pandemia de COVID-19 ha hecho retroceder el progreso que se había conseguido en reducción de la carga de TB global y en la aportación de servicios esenciales, perdiéndose por el momento de vista los objetivos fijados por la OMS para 2020. Se esperaba conseguir un 20% de reducción en la tasa de incidencia, pero solo se logró alcanzar un 11%; y una reducción del 35% en el número de muertes por TB, llegando solo al 9,2% de reducción. La Región Europea de la OMS fue la que más se acercó al objetivo, con una reducción en el número de muertes del 26% (19).

El número de personas nuevamente diagnosticadas con TB cayó de 7,1 millones en 2019 a 5,8 millones en 2020, retrocediendo a niveles de diagnóstico de 2012, y muy alejado de los 10 millones de personas que se estima que desarrollaron TB en 2020. En 2021 se produjo un repunte de casos, con 6,4 millones de afectados. La mayoría de los casos de TB en 2020 se localizaron en las Regiones del Sureste Asiático (43%), África (25%) y el Pacífico Oeste (18%). Los 30 países con mayor carga de TB representan el 86% de todos los casos a nivel mundial, y de estos, ocho engloban dos tercios del total: India (26%), China (8,5%), Indonesia (8,4%), Filipinas (6%), Pakistán (5,8%), Nigeria (4,6%), Bangladesh (3,6%) y Sudáfrica (3,3%). El 56% de los casos globales se da en varones adultos, un 33% en mujeres y un 11% en niños. Además, un 8% de los casos son VIH positivo, alcanzando valores de hasta el 50% en la Región Africana de la OMS.

La reducción en el número de diagnósticos y acceso al tratamiento a causa de la pandemia de COVID-19 se ha traducido en un aumento en el número de muertes por TB. Las estimaciones más optimistas sugieren que 1,5 millones de personas fallecieron de TB en 2020, y 1,6 millones han fallecido en 2021, frente al 1,4 millones de 2019, retrocediendo hasta niveles de 2017. El número de muertos por TB ha aumentado en la mayoría de los 30 países que representan la carga de TB más alta a nivel mundial.

La TB resistente a los antibióticos continúa siendo un problema. La OMS utiliza cinco categorías para clasificar la resistencia: TB resistente a isoniazida, TB resistente a rifampicina, TB multirresistente (MDR, resistente a isoniazida y rifampicina), TB pre-extensamente resistente (pre-XDR, resistente a rifampicina y alguna fluoroquinolona) y TB XDR (resistente a rifampicina, a alguna fluoroquinolona y a un aminoglicósido inyectable y/o a bedaquilina y/o linezolid). En 2020, el 71% de los aislados de las personas diagnosticadas con cultivo bacteriológico se testó para la resistencia a rifampicina, superior al 61% de 2019. Entre ellos se encontraron 437000 casos de TB resistente a rifampicina y MDR, y otros 25681 de pre-XDR y XDR TB, cifras más bajas que las encontradas en 2019 y que concuerdan con el menor número de personas diagnosticadas en 2020 a nivel global. Sin embargo, en 2021 se detectaron 450000 nuevos casos de TB

resistente a rifampicina y MDR, un aumento del 3,1% con respecto a 2020 y que concuerda con el aumento global de casos de TB.

1.2.2. España y Aragón

España está considerado un país de baja carga de TB, con una incidencia de 7,3 casos por 100000 habitantes y alrededor de 3000 casos notificados en 2020, valores más bajos que en 2019 (más de 4000 casos notificados y una incidencia de 9,4 casos por cada 100000 habitantes), lo que concuerda con la caída de casos diagnosticados a causa de la COVID-19. Sin embargo, en 2021 la tasa ha ascendido a 8,2 casos por cada 100000 habitantes y alrededor de 3500 casos notificados. La tasa de mortalidad se encuentra en torno a 0,49 por cada 100000 habitantes. La incidencia fue más alta entre los varones (59%) que en las mujeres (34%) y niños (7%). Los principales factores de riesgo asociados con TB en España son el tabaquismo, el VIH, la desnutrición, el consumo nocivo de alcohol y la diabetes. Cabe destacar que en torno al 73% de los casos desarrollaron una TB pulmonar. En cuanto a la TB resistente a los antibióticos, en 2020 se confirmaron 15 casos de TB resistente a rifampicina/MDR y cuatro casos de pre-XDR/XDR TB, similar a lo obtenido en 2019 (20,21). En 2021, se detectaron 34 casos de TB resistente a rifampicina/MDR y un caso de pre-XDR/XDR TB.

Aragón aplica un protocolo de vigilancia para TB que desde 2004 incluye el genotipado y la identificación de todas las cepas de TB aisladas en la Comunidad. En 2019 la incidencia de TB en Aragón fue de 9,9 casos por cada 100000 habitantes (134 casos notificados, un 60,5% de ellos eran varones). En 2020 la incidencia cayó hasta 5,9 casos por cada 100000 habitantes, con solo 80 casos notificados. Esta bajada es consistente con la disminución del número de diagnósticos a causa de la pandemia de COVID-19 (22,23). Sin embargo, en 2021 la incidencia subió a 7,8 casos por cada 100000 habitantes, con un total de 105 casos diagnosticados.

1.3. La enfermedad

1.3.1. Agente causal: *Mycobacterium tuberculosis*

El género *Mycobacterium* comprende más de 170 especies, la mayoría de ellas ambientales (24). Las micobacterias son aerobias, cilíndricas y no forman esporas. Debido a la composición especial de su pared bacteriana no captan bien los colorantes, pero sí resisten la decoloración por ácido o alcohol, de ahí que se las llame «acidorresistentes». Su envoltura celular es dinámica e inmunomoduladora, y varía a lo largo de la longitud de la célula y durante la infección. Esta variabilidad les permite manipular el sistema inmune humano, tolerar el tratamiento con antibióticos y adaptarse al ambiente variable del hospedador. La envoltura celular está compuesta por la membrana plasmática, la pared celular, los lípidos de superficie y la cápsula, además de estar poblada por numerosas proteínas. El núcleo de la pared contiene capas de lipomananos, arabinogalactano y ácidos micólicos, todo ello rodeado por una membrana externa cerosa, la «micomembrana», rica en ácidos micólicos. La estructura más externa es la cápsula, una matriz de glucanos y

proteínas secretadas que tiene un papel en las interacciones con el huésped y la virulencia (25) (Figura 2).

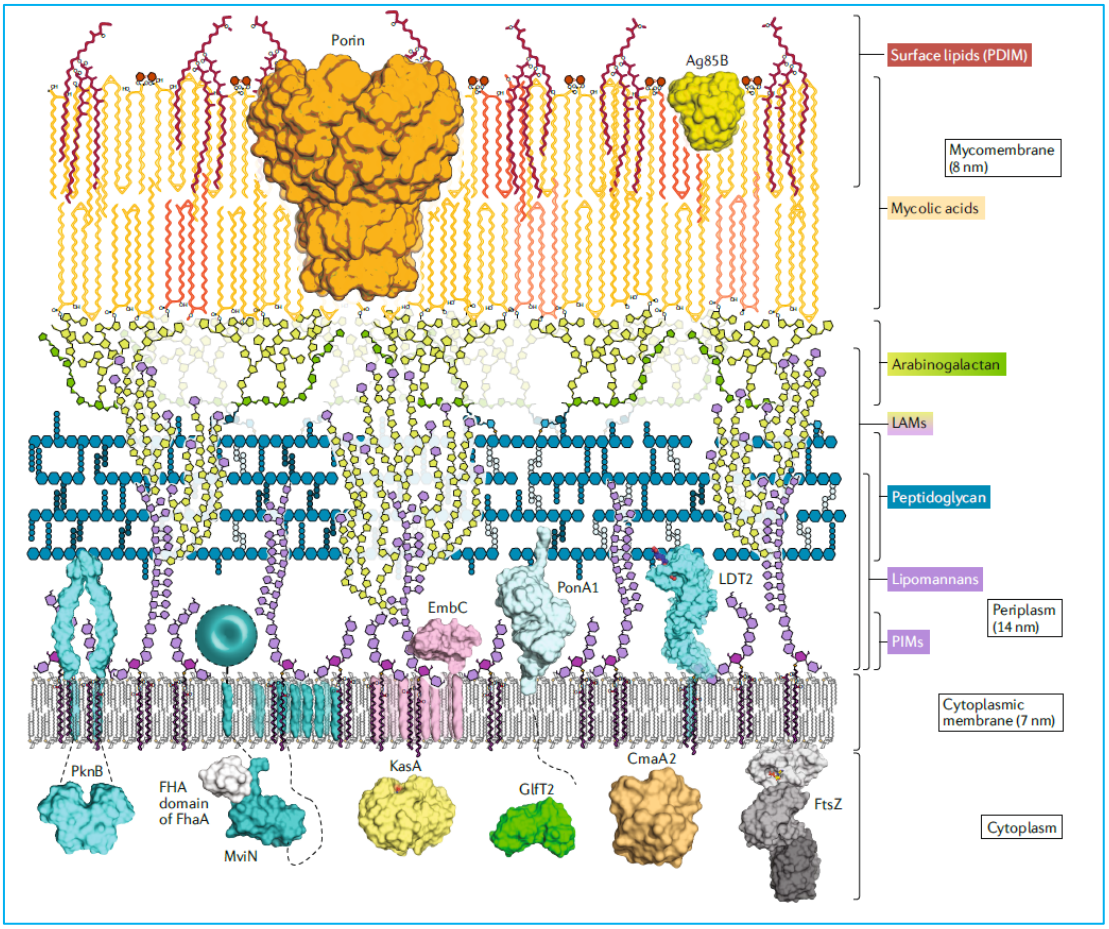


Figura 2. Estructura de la envoltura celular de las micobacterias, incluyendo las proteínas embebidas en ella que juegan un papel en su construcción y modulación. PIMs = fosfoinositol manósidos. LAMs = lipoarabinomananos (25).

Tradicionalmente, las micobacterias se dividen en dos grupos: de rápido crecimiento y de lento crecimiento, siendo a este último al que pertenecen los tres principales patógenos del género (*Mycobacterium tuberculosis* complex o MTBC, *M. leprae* y *M. ulcerans*) (26). Además, algunas bacterias de este género llamadas «no tuberculosas» pueden causar enfermedad, principalmente en personas inmunodeprimidas, como por ejemplo *M. abscessus*, *M. avium*, *M. marinum*, *M. xenopi*, *M. gordonae* y *M. kansasii.*

M. canettii es la bacteria del género *Mycobaterium* más cercana al ancestro común del MTBC, y aunque forma parte del complejo, se diferencia del resto en que puede sufrir transferencia horizontal de genes y recombinaciones internas de forma frecuente (3,27). Se han aislado hasta la fecha menos de 100 genomas de *M. canettii*, todos ellos procedentes del cuerno de África. La transmisión entre humanos no se ha demostrado, lo que sugiere

que se adquiere a través del ambiente, aunque no se ha encontrado ningún reservorio animal o ambiental (28).

1.3.2. *Mycobacterium tuberculosis* complex y sus linajes

El MTBC engloba varias especies bacterianas que comparten un 99,9% de identidad genética, pero se diferencian en el hospedador primario. El complejo se divide en dos grandes grupos: linajes adaptados al ser humano y linajes adaptados a los animales (29).

Los linajes adaptados al ser humano incluyen a *M. tuberculosis sensu stricto* y a *M. africanum*. Los humanos somos el único hospedador conocido donde la infección y transmisión de ambos patógenos ocurre de forma eficiente (30). Hace algunas décadas se postuló que el origen de la TB humana era *M. bovis*, el patógeno que produce la TB en el ganado bovino. Sin embargo, esta hipótesis se descartó completamente ya que los genomas de algunos linajes de *M. tuberculosis* son más ancestrales que el de *M. bovis* (5,31) y, además, este ha perdido varios genes aún presentes en *M. tuberculosis* (32).

La primera división de linajes del MTBC basada en diferencias de regiones genómicas viene dada por la deleción de la región TbD1: las cepas con esta región delecionada se consideran evolutivamente modernas mientras que aquellas que la conservan se consideran ancestrales (5,32). *M. africanum*, *M. bovis* y otras cepas animales no tienen la región TbD1 delecionada, pero comparten la deleción de otra región, la RD9, que no está delecionada en las cepas modernas ni en otras cepas pertenecientes a *M. tuberculosis sensu stricto* (5).

Posteriormente, el MTBC adaptado al ser humano se ha dividido en siete linajes filogenéticos que divergieron a partir de un ancestro común y se diversificaron por distintas partes del mundo (33–37) (Figura 3). Aquellas cepas con la deleción TbD1 (consideradas modernas) se dividen en tres linajes (L): L4 (Euro-Americano), distribuido ampliamente por Europa y América, y también por África y Oriente Medio; L2, localizado en los países de Asia Oriental; y L3, cuya distribución se concentra en el este de África y en el centro y sur de Asia (34,36). Las cepas que conservan la región TbD1 se dividen a su vez en cuatro linajes: L1 (Indo-Oceánico), concentrado en el Océano Índico y Filipinas; L5 y L6, que comprenden al patógeno *M. africanum* y están geográficamente restringidos a África; y L7, restringido a Etiopía (34,36,38).

Los miembros de los L2 y L4 son los responsables de la mayoría de casos de TB a nivel mundial, y se ha observado en modelos animales que muestran una progresión más rápida hacia la enfermedad activa y que son más virulentos que los linajes ancestrales (39,40). El origen monofilético de estos linajes (esto es, comparten un mismo ancestro común) y su expansión a través de diferentes partes del mundo sugiere que estas cepas habrían acompañado a los primeros humanos modernos en la migración fuera de África y que diversificaron junto con diferentes poblaciones humanas (36,41).

En cuanto al linaje adaptado a animales, deriva de un ancestro común compartido por *M. africanum* L6. Algunas de estas cepas pueden infectar y transmitirse en otras especies animales diferentes a su hospedador primario. El MTBC adaptado a animales comprende:

M. microti (infecta ratones y musarañas), dassie bacillus (damanes), *M. pinnipedii* (focas y leones marinos), *M. caprae* (cabras), *M. bovis* (ganado bovino), *M. orygis* (órices, gacelas, ciervos, antílopes, búfalos y también humanos), *M. mungi* (mangostas) y *M. suricattae* (suricatos). Todas estas especies animales y el L6 comparten las deleciones RD7, RD8, RD9 y RD10, y es posible que el ancestro común fuera un bacilo con amplio espectro de hospedador. La especialización habría ocurrido en un estado evolutivo posterior (29). Es interesante que el continente africano alberga la mayor diversidad de linajes, tanto de los adaptados a humano como los animales, lo que sustenta la hipótesis del origen africano del MTBC (3,34,36,42).

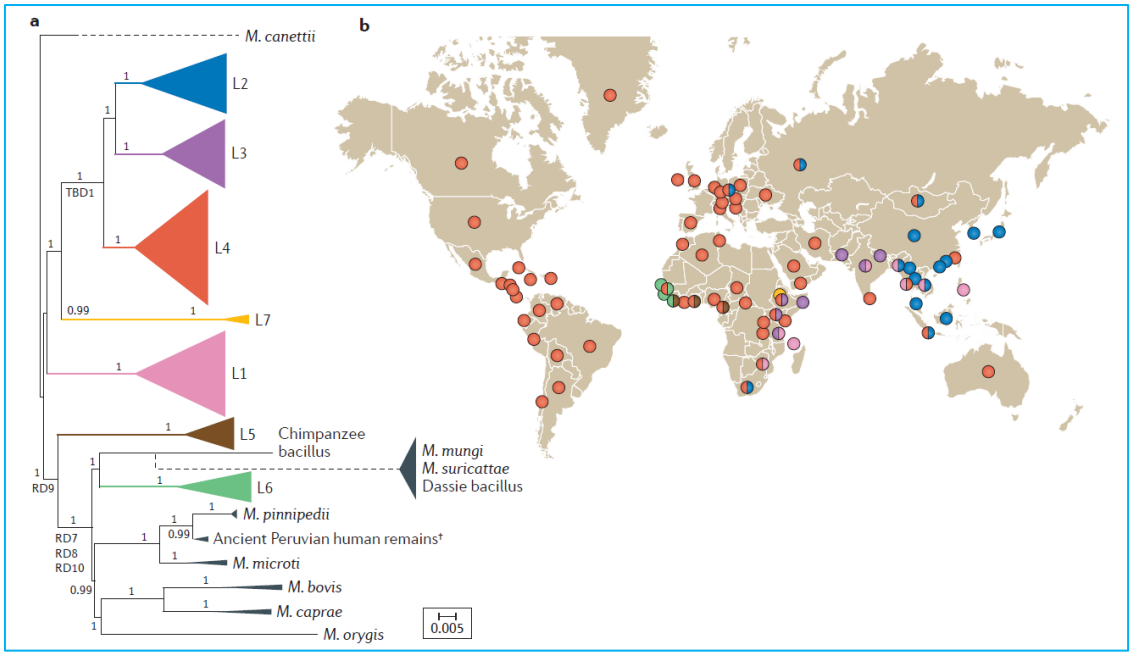


Figura 3. Esquema evolutivo de los diferentes linajes del MTBC y un mapa en el que se indica su distribución global en un código de colores. El L4 es el linaje más extendido a nivel global (43).

1.3.2.1. *M. tuberculosis* L4

El L4 es el linaje más extendido a nivel mundial (40). El estudio más reciente sobre filogenia del L4 lo llevaron a cabo Stucki et al. 2016 (44), y está basado en mutaciones de un solo nucleótido (SNP, por sus siglas en inglés) específicas de cada sub-linaje. Según esta clasificación, el L4 se divide a su vez en 10 sub-linajes: L4.1.1 (familia X según el patrón de Spoligotipo), L4.1.2/Haarlem, L4.1.3/Ghana, L4.2, L4.3/LAM, L4.4, L4.5, L4.6.1/Uganda, L4.6.2/Cameroon y L4.10 (que agrupa los anteriores L4.7, L4.8 y L4.9 descritos por Coll et al., 2014 (45)).

Los L4.1.2/Haarlem, L4.3/LAM y L4.10 se dan en todo el mundo (generalistas) mientras que los L4.1.3/Ghana, L4.5, L4.6.1/Uganda y L4.6.2/Cameroon ocurren en alta frecuencia en regiones específicas de África y Asia (especialistas) pero están ausentes en Europa y

América. La diferente distribución geográfica de estos sub-linajes se puede deber a factores biológicos intrínsecos, factores extrínsecos como las migraciones humanas, o a ambos.

El L4 es el linaje predominante en España, y en concreto en Aragón. En la presente tesis doctoral se incluyen 4 artículos (Publicaciones 4-7) en los que se investigan brotes de TB en Aragón causados por cepas pertenecientes a este linaje.

1.3.2.2. *M. africanum* L5 y L6

M. africanum es el patógeno responsable de la mitad de los casos de TB en el oeste de África (46). Sin embargo, apenas se dan casos fuera de este continente, por lo que parece que es un patógeno restringido a África. Hay tres hipótesis que tratan de explicar este fenómeno: *M. africanum* no es capaz de competir con los linajes modernos de TB (47), extendidos a nivel global; es un patógeno adaptado a la población africana (48,49); o podría haber un reservorio animal, siendo por tanto una enfermedad zoonótica (50–53). *M. africanum* se describió por primera vez en 1968, aislada de un paciente tuberculoso de Senegal (54), y mostró características intermedias entre *M. tuberculosis* y *M. bovis* (55,56). Además, hay estudios con resultados contradictorios en cuanto a si *M. africanum* es menos o igual de virulenta que *M. tuberculosis* (57).

M. africanum incluye dos linajes del MTBC, el L5 y L6, ambos con un mismo ancestro común que tenía delecionada la región RD9. Sin embargo, el L5 se separó mucho antes de la rama evolutiva que el L6, siendo el L6 el último que se separó de la rama de linajes animales. El estudio de filogenias utilizando los SNPs divide el L5 en dos sub-linajes, L5.1 y L5.2. Además, los datos sugieren que el L6 es genéticamente más diverso que el L5 (58).

Es raro encontrar *M. africanum* en países industrializados como España. La mayoría de casos son importados de pacientes que vienen de África (59). Como en Aragón se genotipifican todos los casos de TB, se tienen identificados y registrados todos los casos de *M. africanum* en la comunidad. Esos datos se utilizaron en la presente tesis doctoral para realizar un estudio (Publicación 1) sobre la prevalencia de *M. africanum* en Aragón durante 17 años.

1.4. Genoma de *M. tuberculosis*

1.4.1. Genoma de *M. tuberculosis* H37Rv y otras cepas de referencia

La cepa H37Rv de *M. tuberculosis* se aisló en 1905, y desde entonces se ha utilizado a nivel mundial en la investigación biomédica porque mantiene la virulencia en modelos animales de TB, es sensible a los antibióticos y se puede manipular genéticamente. El genoma de H37Rv se secuenció en 1998: tenía un total de 4411529 pares de bases (pb) y un contenido en GC del 65,6%, constante a lo largo de todo el genoma, lo que indicaba ausencia de islas de patogenicidad adquiridas mediante transferencia horizontal. Además, el genoma era rico en ADN repetitivo, particularmente secuencias de inserción, y también en nuevas familias multigénicas y genes *housekeeping* o constitutivos duplicados. Con respecto a las secuencias de inserción (IS), se encontraron 16 copias de IS*6110* (considerada

más promiscua) y seis copias de IS*1081* (considerada más estable). En total se identificaron más de 32 secuencias de inserción diferentes. También se detectaron dos profagos, phiRv1 y phiRv2, de unas 10 kilobases (kb) de tamaño y organizados de forma muy parecida (60).

Toda la información del genoma de H37Rv está recogida en las plataformas Tuberculist (http://genolist.pasteur.fr/TubercuList/) y Mycobrowser (https://mycobrowser.epfl.ch/). En esta última, más moderna, se recogen los 4173 genes codificados por el genoma de H37Rv. Las funciones de estos genes se agrupan en: virulencia, detoxificación y adaptación (6% de los genes); rutas de información (5,9%); pared celular y procesos celulares (18,9%); ARNs estables (1,7%); secuencias de inserción y fagos (3,6%); PE/PPE (4,1%); intermediarios del metabolismo y la respiración (22,6%); proteínas reguladoras (4,8%); proteínas hipotéticas conservadas (26,5%); metabolismo lipídico (6%); y desconocida (0,4%) (61).

Mycobrowser es un repositorio para datos genómicos y proteómicos de micobacterias patógenas (Figura 4). Así, además de la información de *M. tuberculosis* H37Rv, también se puede encontrar el genoma de *M. leprae*, *M. abscessus* o *M. marinum*, entre otras. Centrándonos en *M. tuberculosis* H37Rv, la información sobre un gen concreto que se puede encontrar en Mycobrowser está dividida en tipo de gen, función, producto del gen, comentarios de interés, categoría funcional, proteómica, las coordenadas en el genoma, la secuencia genética y proteica y si existen mutantes publicados. Además, también se puede acceder al perfil transcriptómico de H37Rv, se puede acceder directamente a la herramienta BLAST (Basic Local Alignment Search Tool) del NCBI (Centro Nacional para la Información Biotecnológica), y el repositorio tiene vinculación con otras bases de datos, como Uniprot o Pubmed, facilitando el acceso rápido a toda la información.

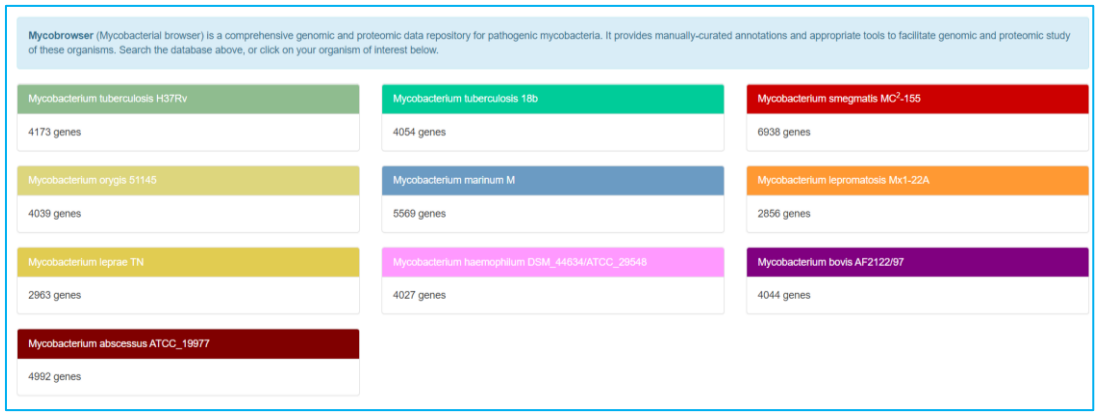


Figura 4. Pantalla de inicio del portal de Mycobrowser. Aquí se muestran todos los genomas de micobacterias patógenas disponibles y el número de genes descritos para cada una.

Otros genomas de referencia que se utilizan con frecuencia son el de *M. bovis* AF2122/97 y el de la cepa *M. tuberculosis* CDC1551. El número de genomas se incrementa cada vez más

gracias a que las técnicas de secuenciación son cada vez más baratas y accesibles. Estos genomas pueden encontrarse en la base de datos del Genbank (https://www.ncbi.nlm.nih.gov/genbank/) (H37Rv: NC_000962.3, *M. bovis* AF2122/97: LT708304.1, CDC1551: NC_002755.2, etc.).

1.4.2. Secuencia de inserción *6110*

IS*6610* es una secuencia de inserción que pertenece a los llamados elementos móviles, es decir, puede «saltar» de una localización cromosómica a otra mediante un mecanismo llamado transposición (62). Este elemento se ha descrito como específico del MTBC y se ha utilizado durante años para diagnosticar y caracterizar a nivel molecular las cepas de *M. tuberculosis* mediante IS*6110*-RFLP (polimorfismos de la longitud del fragmento de restricción, RFLP por sus siglas en inglés) (63–66).

IS*6110* tiene una longitud de 1355 pb y se caracteriza por tener unas repeticiones terminales invertidas de 28 pb (64). Está formada por dos marcos de lectura (open reading frame, orf por sus siglas en inglés) solapantes, *orfA* y *orfB*, que codifican para la transposasa (Figura 5). Recientemente se ha descrito el mecanismo de transposición de IS*6110*, regulado por un cambio en el marco de lectura de -1 pb en la zona solapante de los orfs (67). Además, los productos individuales OrfA y OrfB son capaces de inhibir la transposición (68).

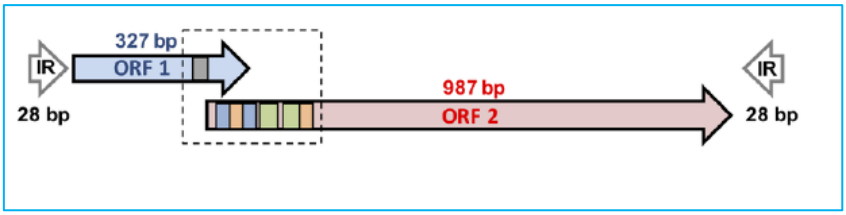


Figura 5. Estrcutura de IS*6110*, con dos ORFs solapantes y 28 pb de las repeticiones invertidas (IR) de los extremos (67). ORF1 = *orfA*; ORF2 = *orfB*.

IS*6110* está conservada entre los miembros del MTBC y se piensa que las diferencias en la frecuencia de transposición no se deben a mutaciones de la secuencia sino a la región genómica en la que está insertada: si está en una zona de baja transcripción, IS*6110* permanece inactiva y apenas transpone, mientras que si está en una zona de alta transcripción la frecuencia de transposición aumenta considerablemente (69,70). Según el número de copias de IS*6110* que se acumulan en el genoma, las cepas se pueden dividir en dos grupos: cepas de bajo número de copias (menos de 5 o 6 copias, en función de los estudios) y cepas de alto número de copias. Se ha propuesto que una acumulación descontrolada de copias de IS*6110* sería deletérea para la bacteria por lo que este número debe ser controlado por mecanismos inhibitorios (62).

No existe ninguna localización específica para la inserción de IS*6110* en el genoma de *M. tuberculosis*, aunque sí algunas zonas en las que es más frecuente encontrar una copia: la

región DR (71), la región de la fosfolipasa C (72), en genes de la familia PPE (73,74), la región intergénica entre *dnaA* y *dnaN* (75), y otras IS, como IS*1547* (76,77). Así mismo, se han descrito algunas localizaciones preferenciales para los miembros de las diferentes familias del MTBC (78,79). La mayoría de los fenómenos de transposición resultan en la integración de IS*6110* en la región codificante de un gen, lo que presumiblemente causará su inactivación (78,80,81). También puede suceder un fenómeno de recombinación entre dos IS*6110* cercanas, lo que provocará la pérdida de los genes situados entre ambas IS*6110*: se han observado deleciones de hasta 20 kb de ADN que contenían 13 genes (82). Finalmente, un tercer efecto de la inserción puede ser que IS*6110* actúe como promotor del gen situado a continuación, ya que se ha demostrado que contiene un promotor externo en el extremo 3' terminal (73,79,83,84).

Las cepas que pertenecen a los linajes modernos del MTBC (L2 y L4) tienen, en general, un número mayor de copias de IS*6110* que las de los linajes ancestrales. Además, se observa que estos linajes son los que tienen una mayor distribución global, es decir, son cepas capaces de infectar altas densidades de población, y se considera que están mejor adaptadas (67). Para explicar esto se ha hipotetizado que IS*6110* podría aumentar el fitness de la bacteria mediante la sobreexpresión de genes de virulencia o la inactivación de determinantes antigénicos (74,84).

En la presente tesis doctoral se han llevado a cabo dos estudios relacionados con IS*6110*: un estudio de mutaciones en la secuencia de IS*6110* para diferentes cepas y localizaciones (Publicación 2), y un estudio transcriptómico para ver cómo las distintas copias de IS*6110* afectan a la transcripción de los genes adyacentes en la cepa MtZ, con todas las copias localizadas, responsable del mayor brote ocurrido en nuestra comunidad (Publicación 6).

1.5. Historia natural y patogénesis

Para que la infección de *M. tuberculosis* tenga lugar, los aerosoles infectados deben llegar hasta el alvéolo pulmonar. No todos los enfermos pueden generar suficiente cantidad de aerosoles capaces de internarse en el alvéolo (85). Además, aquellas personas con mayor probabilidad de sufrir una TB activa son las que están en contacto de manera continuada con una persona que la sufre (86) (Figura 6).

El alvéolo es una estructura formada por células epiteliales (neumocitos de tipo I) de un grosor muy delgado para permitir el intercambio de gases y fuertemente adheridas entre sí para evitar la entrada de plasma. Así se consigue una tensión superficial baja, gracias al surfactante, pero se impide la entrada de anticuerpos. Cada alvéolo tiene asociado un macrófago alveolar (87) que se encarga de mantener limpio el espacio para que el intercambio de gases tenga lugar y también de evitar cualquier desarrollo inflamatorio que afecte la delicada estructura de estas cavidades.

Cuando un bacilo viable de *M. tuberculosis* es fagocitado por el macrófago alveolar, secreta el péptido ESAT-6, esencial para evitar la fusión lisosoma-fagosoma y la apoptosis, lo que le permite entrar en el citoplasma y multiplicarse unas 5-6 veces causando finalmente la

necrosis del macrófago (88,89). Estos bacilos, ahora extracelulares, vuelven a ser fagocitados por otro macrófago que ha venido a sustituir al que se ha necrosado y por otros vecinos, repitiéndose el ciclo hasta alcanzar una cantidad de en torno los 1000 bacilos. Esto provoca que los macrófagos comiencen a secretar quimiocinas suficientes como para activar la respuesta inflamatoria. La inflamación rompe la estanquidad del alvéolo permitiendo la entrada de células polimorfonucleares y un lavado más enérgico de los alveolos afectados mediante el drenaje hacia los nódulos linfáticos. Allí, *M. tuberculosis* infectará a los macrófagos de los nódulos y a las células dendríticas (90). Las células dendríticas procesan *M. tuberculosis* y presentan epítopos del ESAT-6 y del complejo Ag85, antígenos mayoritariamente secretados por el bacilo. La presentación antigénica estimula a los linfocitos T CD4, normalmente a los de tipo Th1, que generarán interferón gamma para activar a los macrófagos infectados. Esta respuesta inmunológica es protectora, porque evita el desarrollo de TB activa en un 90% de los casos (90).

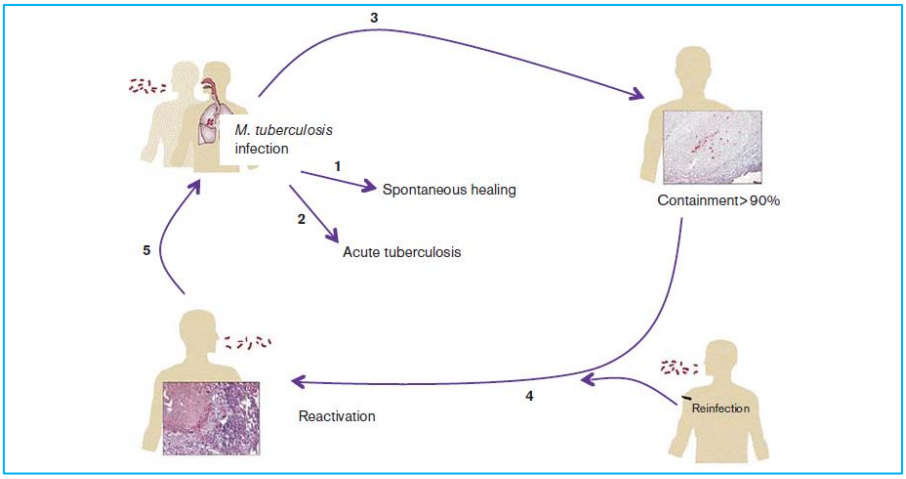


Figura 6. Proceso de la infección natural por *M. tuberculosis*. Cuando una persona se infecta mediante la inhalación de aerosoles que contienen la bacteria, pueden ocurrir tres casuísticas: la curación espontánea, el desarrollo de un episodio agudo de TB (TB activa) o que el sistema inmune sea capaz de contener a la bacteria encerrada en los granulomas, entrando esta en un estado de latencia. Esto último es lo más frecuente, ocurriendo en un 90% de los casos. Si el sistema inmune se ve comprometido de alguna forma, el granuloma no se puede mantener y las bacterias son liberadas, produciéndose la reactivación de la enfermedad, que puede además ocurrir por una nueva reinfección (92).

La infección de *M. tuberculosis* suele terminar ahí, pero puede ocurrir que durante el drenaje a los nódulos linfáticos los bacilos sean liberados hacia los capilares eferentes, llegar hasta la vena cava y ser trasladados de nuevo al pulmón, generándose nuevos focos infecciosos. Además, también puede ocurrir que los bacilos lleguen a los capilares venosos, ser trasportados a la aurícula y ventrículo izquierdo, y desde allí diseminarse de

forma sistémica, ya que *M. tuberculosis* tiene el potencial de colonizar cualquier órgano, produciéndose TB extrapulmonar en un 30% de los casos (90).

Cabe destacar que inicialmente la multiplicación de *M. tuberculosis* es silente, no se genera ningún tipo de respuesta inflamatoria, por lo que los linfocitos no pueden ser atraídos al foco de la infección. Esto significa que la infección y reinfección no pueden evitarse, por tanto, todo objetivo de las vacunas debe ser evitar la TB activa (91).

1.5.1. TB latente

Se estima que alrededor de un cuarto de la población mundial está infectada de forma latente con *M. tuberculosis*. Estas personas no tienen síntomas ni son contagiosas, y la única forma de diagnosticarlas es a través de un test positivo de tuberculina (93,94) o con un test de liberación de interferón gamma (IGRA, por sus siglas en inglés) (Figura 7). Algunos factores de riesgo para desarrollar TB activa son la infección con VIH, las terapias inmunosupresoras, la diabetes, fumar o estar en continua exposición a *M. tuberculosis*.

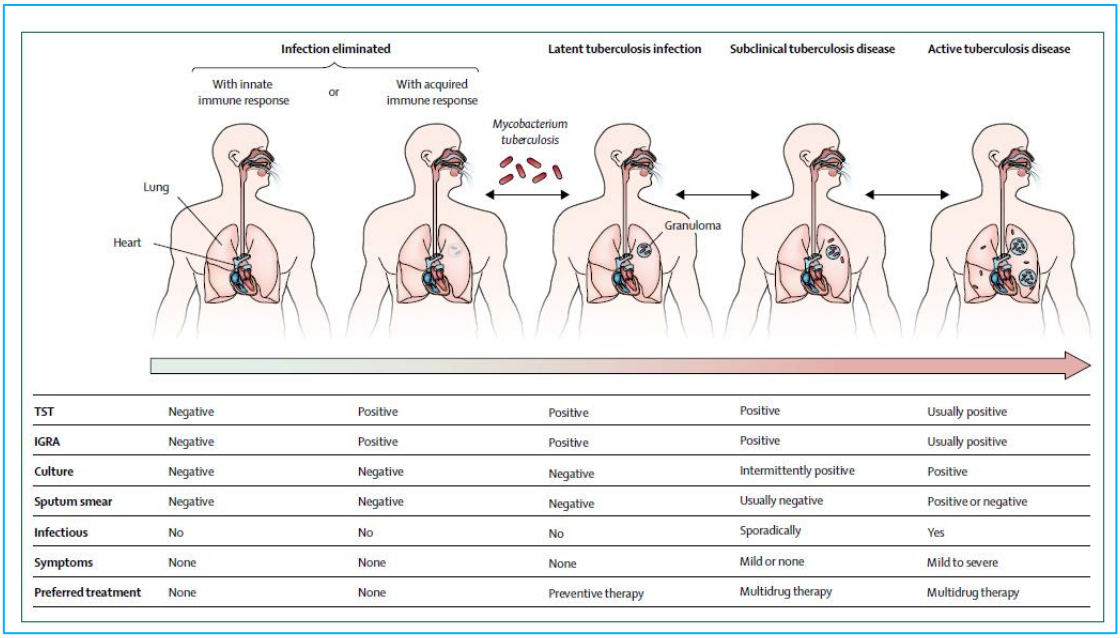
Durante la infección, numerosas células del sistema inmune (neutrófilos, monocitos y células dendríticas) son reclutadas por las señales que emiten los macrófagos infectados (95,96) formando el granuloma, una estructura que aísla las bacterias del resto del cuerpo. En el interior del granuloma los bacilos detienen su crecimiento y mantienen una actividad metabólica mínima, lo que se conoce como dormancia. En este momento se produce un cambio en el metabolismo y la fisiología de las bacterias, ya que los genes que se expresan en este estado son diferentes de los del fenotipo activo (97,98). Este cambio genético parece deberse, o al menos estar bastante influenciado, por el metabolismo del colesterol, pues es la única fuente de carbono presente en el granuloma (97,99).

Las señales que activan la entrada de *M. tuberculosis* en este estado de latencia tienen que ver con el ambiente interno del granuloma: hipoxia, privación de nutrientes y alta concentración de óxido nítrico (93). En 1933 se demostró que cuando se transfería un cultivo de *M. tuberculosis* de un medio rico en nutrientes a una solución salina (PBS), los niveles de respiración celular decrecían y el cultivo entraba en una fase temprana de crecimiento estacionario; cuando estas bacterias se volvían a transferir a un medio rico, se recuperaban los niveles de respiración y crecimiento normal (100). La hipoxia y el ácido nítrico inducen la expresión del regulón DosR, que activa otros 47 genes para comenzar el proceso de dormancia (101).

Si el granuloma no se puede mantener debido a una disfunción del sistema inmune, los bacilos son liberados y la infección progresa a TB activa. A partir de este momento el individuo es contagioso y comienza a experimentar los síntomas característicos de la TB: fatiga, pérdida de peso y tos con esputo sanguinolento (102).

En la presente tesis doctoral se ha llevado a cabo un estudio de casos recurrentes como aproximación a latencia (Publicación 3), en la que se han secuenciado los aislados de 18 pacientes que han tenido al menos dos episodios de TB activa para estudiar las diferencias genéticas entre ambos aislados y contrastar resultados previos de otros estudios que

sugieren que durante los dos primeros años de latencia la tasa de mutación de *M. tuberculosis* es mayor que en los años posteriores.



TST	Negative	Positive	Positive	Positive	Usually positive
IGRA	Negative	Positive	Positive	Positive	Usually positive
Culture	Negative	Negative	Negative	Intermittently positive	Positive
Sputum smear	Negative	Negative	Negative	Usually negative	Positive or negative
Infectious	No	No	No	Sporadically	Yes
Symptoms	None	None	None	Mild or none	Mild to severe
Preferred treatment	None	None	Preventive therapy	Multidrug therapy	Multidrug therapy

Figura 7. Espectro de la infección por *M. tuberculosis* y de la enfermedad. Para cada estadio clínico se indica el resultado de la prueba de la tuberculina (TST), del ensayo de interferón (IGRA), del cultivo y de la baciloscopia del esputo. Además, se indica si el paciente es infeccioso, si se presentan síntomas y el tratamiento a aplicar, si procede (103).

1.6. Diagnóstico de la TB

1.6.1. TB activa

El método estándar para diagnóstico clínico de TB es la tomografía computarizada (TC) de tórax, que permite también monitorizar el efecto del tratamiento, si bien en muchos casos se utiliza la radiografía simple. En la TB primaria, que se da principalmente en niños y adolescentes, se observan mayoritariamente infiltrados inflamatorios acompañados de linfoadenopatía. Además, en torno a un cuarto de los pacientes con TB primaria tienen efusión pleural, siendo rara la formación de cavidades. Los pacientes inmunodeprimidos, los ancianos y los bebés tienen un riesgo elevado de sufrir una TB miliar. En este caso, la TC permite observar muy claramente nódulos pulmonares múltiples. En la TB post-primaria se forman cavidades con necrosis y destrucción del tejido (104). Actualmente se está desarrollando la radiología digital y la interpretación guiada por ordenador, haciéndola más fácil de usar (105).

Las técnicas de diagnóstico bacteriológico incluyen la microscopía del esputo (o de otra muestra de interés), los cultivos líquidos y los kits rápidos basados en biología molecular.

1.6.1.1. Microscopía del esputo

Debido a la pared celular rica en lípidos de *M. tuberculosis*, los colorantes que se adhieren a ella resisten la decoloración mediada por los agentes que contienen ácidos. La tinción más utilizada para observar las muestras con el microscopio óptico es la Ziehl-Neelsen (Figura 8A). En lugares con más recursos se utilizan también tinciones fluorescentes, como la de Auramina (Figura 8B), que es un 10% más sensible y permite un análisis más rápido de las muestras (106). La OMS recomienda además el reemplazo de las fuentes de luz fluorescentes por microscopía fluorescente LED, ya que esto reduce el coste, la energía, los requerimientos de mantenimiento y ya no es necesario un cuarto oscuro para mantener la sensibilidad (107).

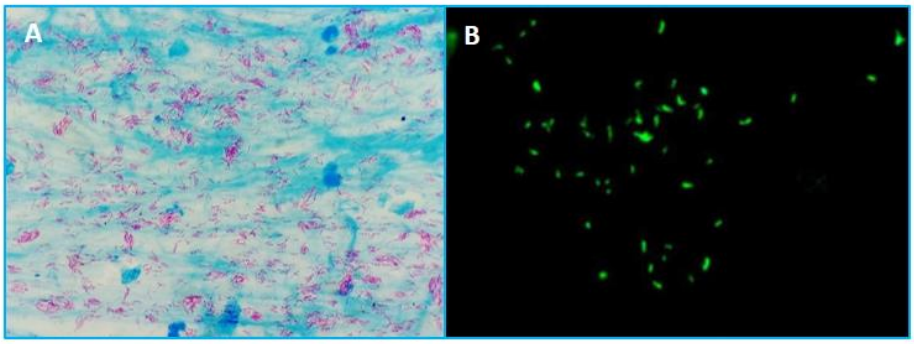


Figura 8. A: Imagen de microscopia de *M. tuberculosis* tras realizar una tinción Ziehl-Neelsen. **B.** Imagen de microscopía de *M. tuberculosis* tras realizar una tinción fluorescente con Auramina.

1.6.1.2. Cultivos líquidos comerciales

El cultivo de las micobacterias en agar sólido (en medio Lowenstein-Jensen) o en medio líquido, como por ejemplo el MGIT (Mycobacterial Growth Indicator Tube, MGIT, de Becton Dickinson) continúan siendo el estándar dorado para el diagnóstico de TB. Antes de sembrar o inocular las muestras clínicas en estos medios, se procede a una descontaminación con NaOH para eliminar los hongos y bacterias de rápido crecimiento, y que no afecta a la viabilidad de *M. tuberculosis*. Algunas limitaciones para la implementación de esta técnica a nivel global son la necesidad de infraestructuras y mantenimiento para asegurar las medidas adecuadas de bioseguridad, una fuente de energía ininterrumpida, entrenamiento del personal, un rápido transporte de las muestras al laboratorio, y su elevado coste (108).

El MGIT se utiliza también para los test de sensibilidad a los antibióticos ofreciendo buenos resultados para isoniazida, rifampicina, fluoroquinolonas, aminoglicósidos y polipéptidos. Sin embargo, es menos fiable para etambutol y pirazinamida, así como antibióticos de segunda línea (109). Además, se recomienda utilizar sistemas automatizados para la detección de las resistencias. Una limitación de los actuales sistemas

comerciales es que solo se incluye una (o como mucho dos) concentración crítica para cada antibiótico, por lo que el resultado obtenido es cualitativo más que semicuantitativo.

1.6.1.3. Detección rápida de *M. tuberculosis*

Existen kits que pueden utilizase para detectar *M. tuberculosis* en una muestra, para diferenciarla de otras especies de micobacterias y para diagnosticar resistencias. Como ejemplo de diagnóstico molecular, el Xpert MTB/RIF se desarrolló y difundió por todo el mundo para diagnosticar rápidamente las cepas MDR de TB. Estos kits de detección rápida están basados en biología molecular. El esputo, o la muestra clínica pertinente, se licua e inactiva con un reactivo que contiene NaOH e isopropanol que viene en el kit, y posteriormente se introduce en el cartucho, donde la muestra es filtrada automáticamente (para capturar los bacilos) y sonicada (para liberar el ADN bacteriano). El proceso continúa con una PCR semi-anidada que tiene como diana una región de 81 pb del gen *rpoB*, en la que se concentran más del 95% de las mutaciones asociadas con resistencia a rifampicina. Dos de las sondas moleculares diseñadas indican la presencia de *M. tuberculosis* en la muestra, mientras que las otras tres están relacionadas con la resistencia a rifampicina (110). La sensibilidad de la técnica es del 89% y la especificidad es del 99% para pacientes con TB pulmonar (111). Una limitación importante es que este kit, y todos los basados en esto, no diferencia entre bacilos vivos o muertos, por lo que la muestra puede mantenerse positiva después de completar el tratamiento, y por tanto no se debe usar para monitorizar la respuesta al mismo (112).

Otros kits rápidos usados para detectar resistencias son los ensayos de sonda lineal (LPAs, por sus siglas en inglés), como por ejemplo GenoType MTBDR*plus* de Hain Lifescience. Este test identifica resistencias a isoniazida y rifampicina mediante la detección de fragmentos amplificados por PCR que se unen a las sondas diseñadas, que contienen las mutaciones más frecuentes que confieren resistencia a estos fármacos. Las ventajas que presentan estos kits frente a las pruebas de sensibilidad en medio líquido son que es más rápido, disminuye el riesgo y aumenta el rendimiento (113). Para el GenoType MTBDR*plus* se ha demostrado una sensibilidad del 98% en la detección de resistencia a rifampicina y una sensibilidad variable (entre el 77 y el 90%) para la isoniazida. La especificidad en ambos antibióticos fue del 99% (114). La principal limitación de estos kits es que requieren una manipulación de los amplicones obtenidos en la PCR, lo que puede provocar contaminaciones cruzadas entre distintas muestras.

También existen estos kits para detectar resistencias a los antibióticos de segunda línea, como el MTBDR*sl* (Hain Lifescience) para detectar resistencias a fármacos inyectables y fluoroquinolonas, y que testa los genes *gyrA*, *gyrB*, *rrs* y *eis*. Este kit mostró una sensibilidad de 83% para la detección de resistencia a fluoroquinolonas y un 77% para los fármacos inyectables. La especificidad fue mayor del 98% en ambos tipos de fármacos (115).

1.6.2. TB latente

Para diagnóstico de la TB latente se utiliza el Test Cutáneo de Tuberculina y los ensayos IGRA, ambos basados en aproximaciones inmunológicas. Sin embargo, ninguno de ellos

puede diferenciar entre enfermedad activa o latente, ni tampoco identifican a aquellos pacientes que tienen un mayor riesgo de progresar a enfermedad activa (116).

La prueba de la tuberculina se realiza inyectando 0,1 ml de un derivado de tuberculina en el antebrazo de manera intradérmica. Dependiendo de la respuesta inmune del organismo frente a la tuberculina inyectada, se forma una induración en la piel de entre 6 y 10 mm de diámetro. Tras 48-72 h, se mide el diámetro de la roncha y se interpreta el resultado de acuerdo con los criterios estandarizados (117). La vacunación previa con BCG puede producir falsos positivos, ya que esta prueba solo mide si el sistema inmune de la persona testada se ha enfrentado alguna vez a los antígenos (en este caso la tuberculina) de *M. tuberculosis*.

Para el IGRA se requiere una muestra de sangre y se mide la cantidad de interferón gamma liberado por los linfocitos T cuando esa sangre se pone en contacto con antígenos purificados de *M. tuberculosis* (como ESAT-6 y CFP-10), diferentes a los de BCG, por lo que sí que permite diferenciar entre la vacuna y una posible infección (116).

1.7. Tratamiento de la TB y resistencias

El tratamiento estándar de una TB pulmonar dura seis meses e implica cuatro antibióticos diferentes: los dos primeros meses se administra isoniazida (inhibe la síntesis de los ácidos micólicos de la pared bacteriana), rifampicina (inhibe la síntesis de ARN), etambutol (inhibe la síntesis de arabinogalactano de la pared bacteriana) y pirazinamida (inhibe el transporte a través de la membrana plasmática al alterar el potencial energético de la misma), y los últimos cuatro meses se continúa con isoniazida y rifampicina (118). Las formas extrapulmonares y diseminadas requieren un tratamiento de mayor duración: la TB ósea o articular puede tener un tratamiento de 9 meses, y una TB que afecte al sistema nervioso central puede tener un tratamiento de hasta 12 meses.

Aquellos pacientes que presentan mono-resistencia a rifampicina o isoniazida deben seguir un tratamiento que sustituya el fármaco al que la bacteria es resistente por una fluoroquinolona (inhiben la síntesis de ADN), como moxifloxacino o levofloxacino. En estos casos, el tratamiento también se alarga considerablemente, hasta 9 meses para los casos resistentes a isoniazida y de 18 a 20 meses en la resistencia a rifampicina.

En los casos de TB MDR y XDR, la situación es muy compleja y requiere un tratamiento personalizado. En general se recomienda que el régimen incluya cinco fármacos durante la fase intensiva y cuatro en la fase de continuación, con una duración de entre 15 y 21 meses en los casos de TB MDR, y de 14 a 24 meses en los casos de TB pre-XDR/XDR. La fase intensiva, que se caracteriza por una disminución significativa de la carga bacteriana, debe tener una duración de entre 5 y 7 meses. El diseño del tratamiento debería contemplar los siguientes pasos: prescripción de una fluoroquinolona de última generación (levofloxacino, moxifloxacino) seguida de la administración de bedaquilina (inhibe la ATP sintetasa) y linezolid (inhibe la síntesis proteica), así como otros fármacos de alta eficacia como clofazimina (inhibe el crecimiento bacteriano mediante la unión al

ADN) o cicloserina (inhibe la síntesis de la pared bacteriana). Si apareciera resistencia a alguno de estos fármacos y por tanto una imposibilidad de dar al paciente cinco fármacos diferentes, habría que incluir un antibiótico inyectable (amikacina o estreptomicina). Si fuera necesario o si se prefiriese un régimen oral, los fármacos inyectables deberían ser reemplazados por delamanid (inhibe la síntesis de ácidos micólicos), pirazinamida o etambutol (siempre que se haya confirmado la sensibilidad) (119).

1.8. Epidemiología molecular

El objetivo de la epidemiología molecular es comparar el material genético de dos o más aislados. Si las cepas están relacionadas, es decir, proceden de un ancestro común, el grado de similitud genotípica y fenotípica será mucho mayor que entre cepas no relacionadas de la misma especie escogidas al azar. Las cepas de *M. tuberculosis* tienen poca variabilidad genética entre sí, ya que no existe transferencia horizontal de genes. Sin embargo, existen diferencias a nivel de SNPs, de secuencias largas (inserción/deleción) o de secuencias repetidas. Todas estas variaciones se pueden detectar mediante el uso de determinados marcadores, algunos de los cuales detectan variaciones que tuvieron lugar en momentos concretos del tiempo y que permiten clasificar las cepas de *M. tuberculosis* en linajes; otros detectan cambios más recientes, lo que permite tipificar las cepas para identificar cadenas de transmisión (120).

Los SNPs se pueden usar como marcadores para estudios filogenéticos, ya que su reloj evolutivo es lento. Los SNPs no sinónimos, que representan las dos terceras partes de la variabilidad de *M. tuberculosis*, están sometidos a presión selectiva y pueden influir en la transmisibilidad, la resistencia a antibióticos, la virulencia, la respuesta inmune y la presentación del cuadro clínico. El análisis de grandes regiones del genoma delecionadas (RD, por sus siglas en inglés) también ha permitido el estudio evolutivo del complejo tuberculosis y la identificación de las sub-especies del mismo.

Para tipificar las cepas se utilizan diferentes técnicas que se explicarán más adelante. Todas ellas, con mayor o menor poder discriminatorio, tienen como objetivos detectar brotes o epidemias, identificar infecciones mixtas o complejas, resolver contaminaciones cruzadas y estudiar dinámicas de transmisión.

Poco a poco, gracias al abaratamiento de las técnicas, especialmente la secuenciación genómica, y al desarrollo de sistemas de análisis bioinformáticos más asequibles, se va produciendo la transición de la epidemiología molecular a la epidemiología genómica.

1.8.1. Técnicas de genotipificación tradicional

Una de las técnicas más utilizadas durante años para genotipificar las cepas del MTBC es el IS*6110*-RFLP, ya que el alto grado de polimorfismo de esta IS entre las distintas cepas del complejo, tanto en número como en localización, lo hacen un marcador muy útil. Esta técnica tiene un buen poder discriminatorio y su desarrollo y validación estuvieron apoyados por consenso internacional (63,121). El IS*6110*-RFLP se basa en la purificación de ADN, digestión con una enzima de restricción (*Pvu*II), separación de los fragmentos

obtenidos mediante electroforesis e hibridación con una sonda de ADN complementaria a la región 3' de la secuencia de la IS*6110* (Figura 9). Se genera de esta forma un patrón de bandas que es característico en el número y tamaño de las bandas para cada cepa del MTBC. En la práctica, dos o más aislados con un patrón de bandas idéntico (siempre para cepas de alto número de copias de IS*6110*) se consideran parte de un mismo clúster de transmisión. Una de las limitaciones de esta técnica se da en las cepas de bajo número de copias, en las que puede haber una sobreestimación de los agrupamientos y la técnica no resulta tan fiable (122). Otra limitación es que necesita de 1-2 microgramos de ADN purificado, lo que requiere un largo periodo de incubación para tener bien crecido el cultivo de *M. tuberculosis* (de 6 a 8 semanas), y además, la técnica en sí es laboriosa (123).

La necesidad de obtener resultados más rápidos provocó el desarrollo de nuevas técnicas basadas en PCR. La más utilizada fue el *Spoligotyping*, basada en el estudio de la región CRISPR (Repeticiones Palindrómicas Cortas Agrupadas y Regularmente Espaciadas, CRISPR por las siglas en inglés) o DR (124). En esta técnica se utilizan 43 espaciadores distribuidos entre repeticiones de ADN de la región CRISPR (Figura 9). Dependiendo del número de espaciadores presentes en cada cepa, variable entre cepas, se obtienen diferentes patrones o e*spoligotipos* que permiten diferenciar los miembros del MTBC. Esta técnica es barata y rápida, pero su poder discriminatorio es menor que el del IS*6110*-RFLP. Por ello, actualmente se utiliza más para estudios evolutivos y filogenéticos, y para la asignación de linajes, familias y subfamilias en los protocolos de vigilancia global (125).

Tras varios intentos de mejora, se consiguió desarrollar una nueva y prometedora técnica llamada MIRU-VNTR (Mycobaterial interspersed repetitive unit – variable number tandem-repeat, en inglés), que se basa en el análisis del número de repeticiones en tándem en una selección de dianas en el cromosoma de *M. tuberculosis* (Figura 9). El primer diseño analizaba 12 loci (126,127), que después se aumentó a 15 loci (128) y finalmente a 24, considerándose que esta versión es la que ofrece el mayor poder discriminatorio (129). La mayoría de estudios concuerdan en que el MIRU-VNTR es tan bueno como el RFLP en cuanto a poder discriminatorio, y lo supera en términos de rapidez, es más fácil de realizar y también se facilita el intercambio de información entre laboratorios al estar los resultados basados en un código numérico (130–132).

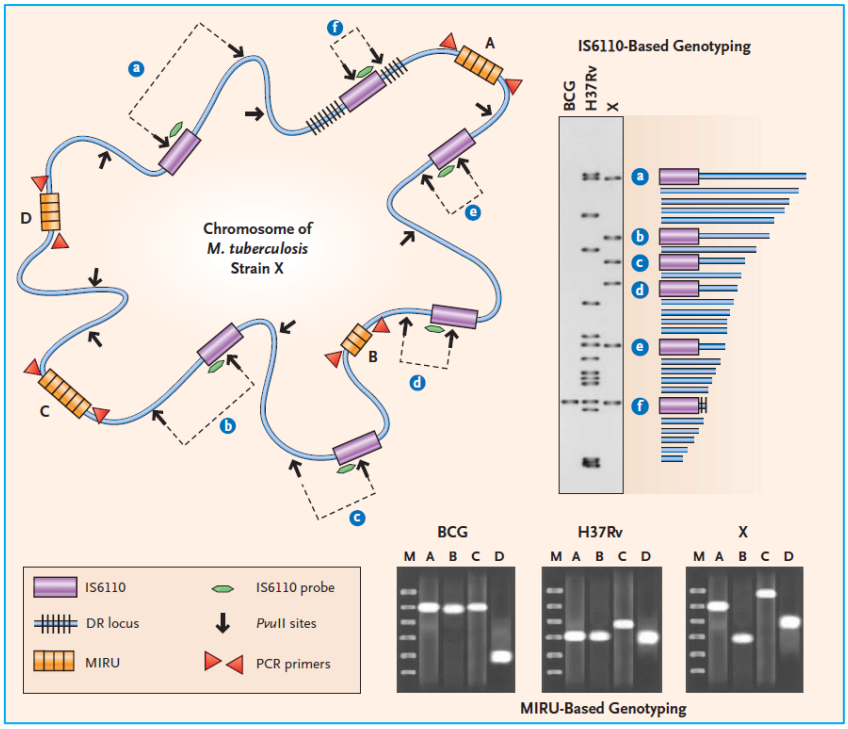


Figura 9. Esquema del cromosoma de *M. tuberculosis* indicando las diferentes dianas de las técnicas de genotipificación tradicional. La IS*6110* se utiliza para el IS*6110*-RFLP, las repeticiones de la región DR se usan en el *Spoligotyping* y las unidades repetidas (MIRU) en la técnica MIRU-VNTR (133).

SITVIT2 (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/index.jsp) es una base de datos internacional de marcadores genotípicos (e*spoligotipos*, 12-loci, 15-loci y 24-loci MIRU-VNTR**)** de *M. tuberculosis* proporcionada por el Instituto Pasteur de Guadalupe, con más de 110000 aislados registrados. Contiene una gran colección de datos epidemiológicos, demográficos y genéticos que permiten una visión exhaustiva de la distribución del MTBC a nivel global. Además, también ofrece información sobre las correlaciones estadísticas que existen entre la información filogenética, clínica, demográfica y epidemiológica. También dispone de aplicaciones integradas que permiten dibujar mapas, gráficos y tablas frente a diferentes parámetros y variables, disponibles para aislados individuales, así como herramientas para generar mapas geográficos de una cepa clínica a nivel nacional, regional y local (134).

1.8.2. Plataformas de secuenciación de genoma completo: ADN y ARN

Para conseguir un mayor poder discriminatorio que asegure la máxima precisión en la definición de los clústeres de transmisión se debe acceder a una información más amplia del cromosoma bacteriano. Para ello, la identificación definitiva debe ser la secuenciación de genoma completo (WGS, por sus siglas en inglés). El abaratamiento de la técnica y el desarrollo de plataformas más accesibles han permitido proponer su utilización de forma rutinaria en los laboratorios de epidemiología molecular para estudiar la transmisión y

facilitar así la transición, en un futuro próximo, hacia una epidemiología basada en la genómica (123).

La técnica de secuenciación que abrió el camino para todas las demás fue la secuenciación Sanger, que se basa en el uso de unos desoxirribonucleótidos (dNTPs) que carecen del grupo hidroxilo del 3' necesario para la extensión de las cadenas de ADN, los dideoxinucleótidos (135,136). Mediante reacciones en las que se combinan dNTPs normales y una pequeña concentración de dideoxirribonucleótidos marcados radiactivamente se obtienen cadenas de ADN de todas las longitudes posibles, ya que cuando la ADN polimerasa incorpora a la cadena un dideoxinucleótido, la reacción de polimerización se detiene. Esto se hace en cuatro tubos independientes, uno para cada base del ADN, y posteriormente se cargan las cadenas obtenidas en un gel de poliacrilamida y se hace una electroforesis para que los fragmentos se separen por tamaños. Leyendo el gel desde abajo hacia arriba se consigue la secuencia de la cadena de ADN diana. Esta técnica se ha ido mejorando, sustituyendo los nucleótidos radiactivos por una detección fluorométrica, lo que permite introducir los cuatro dideoxinucleótidos en un mismo tubo y llevar a cabo una sola reacción. Además, los geles de poliacrilamida se han sustituido por electroforesis capilar haciendo el sistema más robusto cuanto mayor sea el número de capilares del secuenciador. Estas mejoras han contribuido al desarrollo de los secuenciadores automáticos (137), lo que se conoce como secuenciación de segunda generación. El primero de ellos fue la pirosecuenciación, basada en luminiscencia (138,139). Este método utiliza varias enzimas: la ATP sulforilasa, que convierte el pirofosfato liberado tras la incorporación del nucleótido a la cadena de ADN en ATP, y la luciferasa, que utiliza ese ATP para una reacción que produce un destello de luz. La luz producida es proporcional a la cantidad de pirofosfato del medio, de manera que si se incorporan dos bases iguales a la cadena, la señal será de doble intensidad, aunque deja de ser fiable cuando el número de bases iguales es superior a cuatro. Antes de añadir al medio un nuevo nucleótido se realiza un lavado para eliminar el anterior, lo que encarece los costes porque las enzimas también son eliminadas y hay que volver a añadirlas. La preparación de las librerías de ADN también resultó novedosa en su momento: al ADN fragmentado se le unen unos adaptadores complementarios a otros adaptadores presentes en unas perlas, de manera que cada fragmento de ADN se va a unir a una perla. A continuación tiene lugar una PCR en emulsión: en cada perla se amplifica el fragmento unido previamente, obteniéndose varias copias del mismo fragmento en cada perla. Luego, las perlas se depositan en los pocillos de la placa, idealmente una perla por pocillo, y comienza la pirosecuenciación, consiguiéndose lecturas de entre 400 y 500 pb.

Posteriormente se desarrolló el método Solexa, adquirido por Illumina (140), y que hoy en día se conoce por el nombre de la compañía. En vez de realizar una PCR en emulsión basada en perlas, las moléculas de ADN, con sus adaptadores unidos durante la preparación de las librerías, se unen a los adaptadores fijados en el fondo de un pocillo de la celda de flujo (idealmente un fragmento de ADN en cada pocillo) donde tiene lugar una PCR en fase sólida que se conoce como amplificación en puente y que generará grupos vecinos de poblaciones clonales de cada una de las hebras de ADN (141,142). La

secuenciación en sí misma utiliza dNTPs fluorescentes con terminador reversible, que no se pueden unir directamente al ADN porque tienen el grupo hidroxilo del 3' ocupado por el fluoróforo, por tanto debe ser eliminado antes de que continúe la polimerización. Antes de la eliminación, el fluoróforo se lee tras ser excitado con el láser apropiado, emitiendo cada dNTP un color diferente (Figura 10). Esta técnica tiene la ventaja de obtener secuencias *paired-end*, esto es, se secuencia tanto la hebra molde como la complementaria. El secuenciador original ha evolucionado a nuevos modelos con mayor rendimiento, como por ejemplo el HiSeq, que permite lecturas de mayor longitud y profundidad, o el MiSeq, que tiene un menor rendimiento, pero menor coste, mayor rapidez y lecturas más largas (143,144).

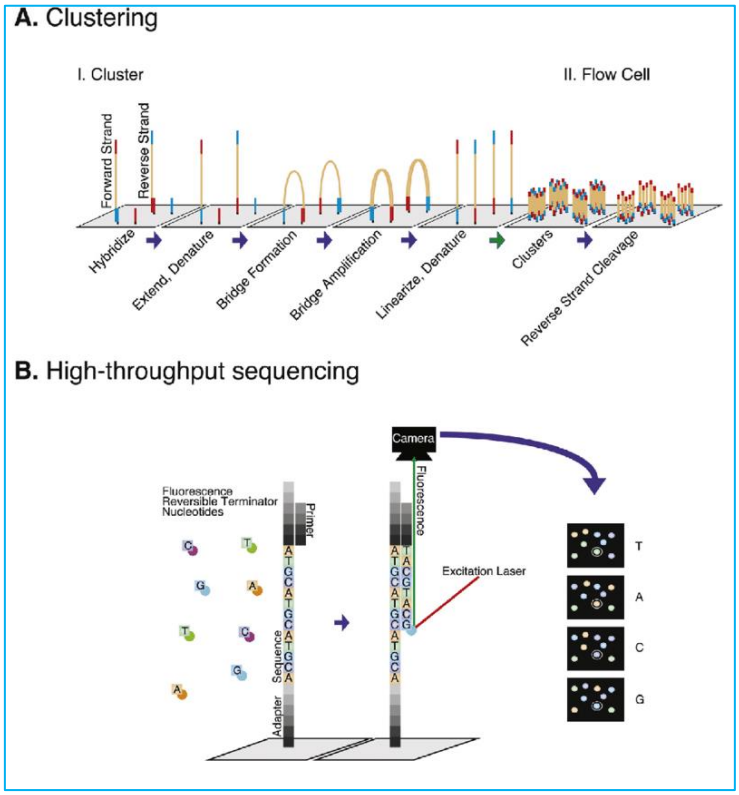


Figura 10. Esquema del proceso de secuenciación utilizado por Illumina, con la amplificación en puente y el uso de dNTPs fluorescentes con terminador reversible (145).

Otra plataforma de secuenciación es Ion Torrent, de Thermo Fisher Scientific, que fue revolucionaria en su momento porque no utiliza luminiscencia ni fluorescencia, sino cambios de pH (146). La preparación de las librerías es similar a la pirosecuenciación, con la unión de adaptadores al ADN y la PCR en emulsión basada en perlas. Estas perlas se colocan posteriormente en los pocillos de un microchip, idealmente una perla por pocillo, y se van añadiendo los dNTPs de uno en uno, lavando entre uno y el siguiente. Cuando la ADN polimerasa incorpora un dNTP a la cadena de ADN, se libera un protón (H+), que producirá un cambio de pH en el medio, medible gracias al poder conductor del silicio del

microchip (Figura 11). Esta técnica es muy rápida, pero, igual que en la pirosecuenciación, tiene poca fiabilidad cuando las secuencias son homopoliméricas.

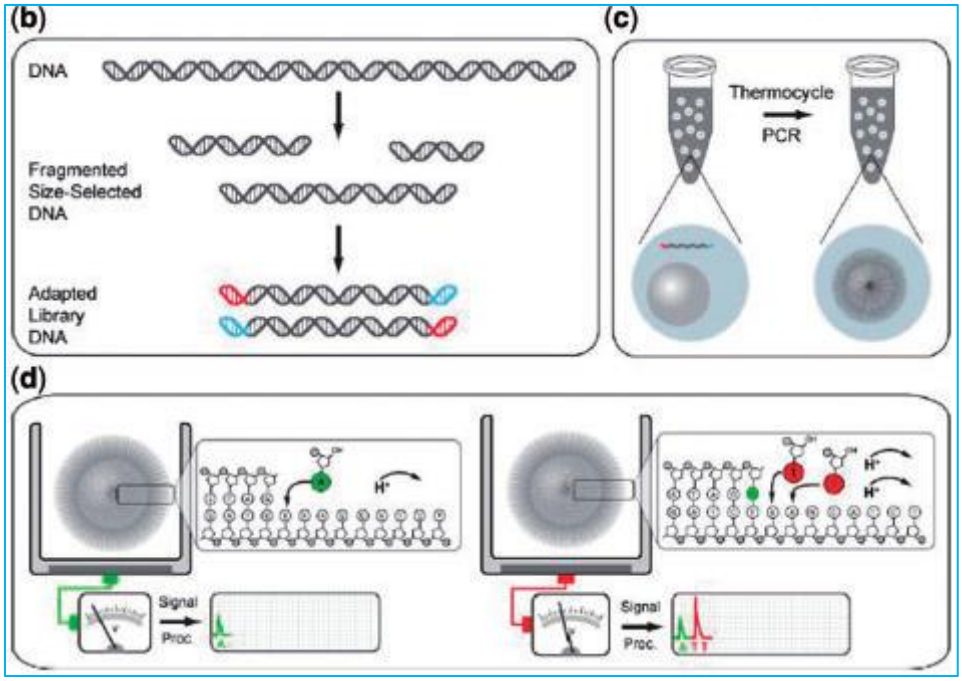


Figura 11. Esquema del proceso de secuenciación de la tecnología Ion Torrent, con la PCR en emulsión y la medida de los cambios en el pH producidos por la liberación de los H+ durante la reacción de polimerización (147).

Actualmente ya están disponibles las técnicas de secuenciación de tercera generación, que se basan en la secuencia de una sola molécula, haciendo innecesaria la amplificación previa por PCR de todas las tecnologías anteriores. La primera de ellas fue Helicos BioScience, que funciona igual que Illumina pero sin la amplificación en puente (148). Posteriormente se han desarrollado plataformas de tiempo real de molécula única, como la de Pacific Biosciences (149), capaces de producir lecturas muy largas, de hasta 10 kb, muy útiles para los ensamblajes *de novo*. Recientemente ha emergido la secuenciación a través de un nanoporo (150), siendo la más importante en el mercado la de Oxford Nanopore (151). Esta tecnología se basa en la capacidad de una molécula de ADN de atravesar una membrana lipídica a través de canales iónicos, como la α-hemolisina, mediante electroforesis. Para mejorar la especificidad, la α-hemolisina puede ser sustituida por canales artificiales obtenidos por ingeniería de proteínas o con tecnología de nanomateriales (152,153). Estas plataformas de tercera generación son más caras y están orientadas a obtener lecturas de un tamaño mayor.

En la presente tesis doctoral se han utilizado tecnologías de segunda generación, como la secuenciación capilar para secuenciar fragmentos obtenidos mediante PCR, o Ion Torrent e Illumina para secuenciar genomas completos de *M. tuberculosis*. Además, el gran avance

que se ha producido en las tecnologías de secuenciación ha permitido el desarrollo de estrategias más dirigidas como el AmpliSeq, donde el investigador puede diseñar a la carta los genes que quiere estudiar, y solo se amplifican y secuencian esos, estrategia utilizada en la presente tesis doctoral.

Un factor muy importante en la WGS es la gran cantidad de datos que se generan durante la secuenciación. El secuenciador extrae un archivo fastQ, en el que se encuentran, sin ensamblar, todas las lecturas obtenidas. A partir de aquí puede hacerse un ensamblaje *de novo*, que consiste en alinear todas las secuencias para obtener un genoma completo ensamblado sin usar un genoma de referencia, lo que requiere un poder computacional elevado; o bien mapear las lecturas frente a un genoma de referencia, la opción utilizada en esta tesis doctoral. La desventaja del mapeo frente a una referencia es que aquellas regiones genómicas que no estén en la cepa de referencia, pero sí en la secuenciada, se perderán al no poder alinearse frente a la referencia. Las secuencias repetidas, como IS*6110*, tampoco se puede estudiar fácilmente, ya que las lecturas de la cepa secuenciada mapearán con todas las IS*6110* del genoma de referencia, sin indicarnos dónde se encuentran realmente en la cepa de interés, y por tanto es necesario diseñar un programa de búsqueda activa. En esta tesis se ha trabajado en esto.

Tras el mapeo se pueden obtener otros tipos de archivos, como el Binary Aligned Map (BAM) y Variant Call Format (VCF). El archivo BAM contiene todas las lecturas obtenidas mapeadas frente al genoma de referencia, lo que permite visualizar las diferencias encontradas; y el VCF contiene los SNPs encontrados frente al genoma de referencia. Estos archivos se pueden cargar en programas de visualización de genomas, como el Integrative Genomics Viewer (IGV, del Broad Institute) (154).

Además, para el estudio de los genomas de *M. tuberculosis*, se ha utilizado en la presente tesis doctoral el software Bionumerics (v7.6, Applied Maths, Kortrijk, Belgium), donde pueden almacenarse, estudiarse y compararse todos los genomas cargados. Se trata de una plataforma para almacenar y analizar datos biológicos. Además de para *M. tuberculosis*, también se puede utilizar para otros microorganismos. Entre sus numerosas aplicaciones se encuentran: tipado de geles de electroforesis de campo pulsado, predicción de resistencias, identificación de bacterias por MALDI-TOF, tipado mediante MIRU-VNTR, *spoligotipying* y RFLP, análisis del MLST (Multi-Locus Sequence Typing), análisis de SNPs a partir de secuencias obtenidas por WGS y el estudio filogenético. También se han utilizado otras herramientas online para el estudio de los genomas, como la plataforma PhyResSe, en la que se cargan directamente los archivos fastQ obtenidos tras la WGS (tanto de Ion Torrent como de Illumina) y hace un estudio de calidad, resistencias y linajes de las secuencias obtenidas (http://www.phyresse.org/); también Genewise, desarrollada por el Instituto de Bioinformática Europeo, que permite hacer alineamientos entre secuencias de ADN y proteínas, identificando cambios de aminoácido (https://www.ebi.ac.uk/Tools/psa/genewise/); y PROVEAN, del Instituto Craig Venter, que predice si un cambio de aminoácido o un INDEL (eventos de inserción/deleción) tiene un efecto en la función biológica de una proteína (http://provean.jcvi.org/) (155).

El ARN se puede secuenciar utilizando las mismas tecnologías que para el ADN, con algunos pasos previos como la depleción del ARN ribosómico. Tras secuenciar el ARN se obtiene el perfil transcriptómico de la célula, es decir, se pueden estudiar los niveles de expresión de todos los genes en cualquier momento y bajo cualquier condición. En la presente tesis doctoral se analizó el perfil transcriptómico de tres cepas de *M. tuberculosis* bajo condiciones diferentes de fases de crecimiento, con el objetivo de encontrar diferencias en los niveles de expresión de las IS*6110*. Para visualizar los datos obtenidos se puede utilizar el Integrated Genome Browser (IGB), un programa con una interfaz similar al IGV, y que permite visualizar los transcriptomas (156).

1.8.3. Uso de la secuenciación genómica en la tipificación del MTBC

El estudio de la filogenia de *M. tuberculosis* utilizando WGS ha permitido establecer diferentes clasificaciones para poder asignar los linajes y familias a las cepas del MTBC, basadas en SNPs específicos de cada uno de estos linajes y familias (44,45). Además, la epidemiología de la TB utiliza WGS para cuatro aspectos: identificar la cadena de transmisión, diferenciar entre recaída y reinfección, determinar la diversidad intra-hospedador (microevolución o coinfección) e identificar resistencias primarias frente a resistencias secundarias (157).

Para estudiar la transmisión se utiliza el umbral fijado por Walker et al. (158): se considera que dos aislados pertenecen al mismo clúster si entre ellos no hay más de 12 SNPs, y se considera que la transmisión es reciente si el número de SNPs es ≤ 5. La WGS permite distinguir cepas que serían idénticas utilizando cualquier otro método de genotipificado, y eso puede ayudar a inferir la dirección de la transmisión. Las tres aproximaciones propuestas son la acumulación de SNPs, inferencia estadística bayesiana y las redes. La más simple, la basada en la acumulación de SNPs, asume que las cepas adquieren nuevos SNPs a lo largo del tiempo y que retienen los ya existentes, y por tanto la transmisión siempre es hacia el caso con los nuevos SNPs. Esta aproximación se puede combinar con las historias clínicas de los pacientes y los datos de contacto para hacerla más robusta (159–161). Cabe destacar que incluso con el uso de la WGS existe una incertidumbre considerable acerca de la interpretación de la transmisión. Además, los errores de secuenciación pueden ser malinterpretados como SNPs y eso tendría un gran impacto en la inferencia de la cadena de transmisión (162).

Con respecto a las recaídas (implican un fracaso en el tratamiento) y las reinfecciones (sugieren una transmisión en desarrollo y una falta de inmunidad a la nueva cepa infectiva) (163,164), es necesario cuantificar las diferencias genómicas entre el primer aislado y los aislados de los episodios recurrentes. Los estudios llevados a cabo en este campo se basan en el número de SNPs encontrados entre los aislados. En cuanto a episodios de recaída, los SNPs encontrados entre los aislados están entre cero y ocho. Sin embargo, pueden ser más de 1000 cuando hablamos de reinfecciones con cepas que pertenecen a diferentes linajes (163,165). La limitación fundamental de la WGS en este aspecto es que es posible reinfectarse con una cepa genéticamente idéntica (157).

Para identificar co-infección y/o microevolución se utilizan los SNPs detectados como *hetorocigotos*, esto es, que en una misma posición se encuentra la lectura de dos bases con la misma probabilidad (160,166–169). Estos SNPs *heterocigotos* también sirven para resolver direcciones de transmisión: su presencia puede sugerir transmisión desde un paciente con el alelo de referencia seguido de microevolución en el segundo paciente, o microevolución en el primero surgiendo un alelo alternativo y transmitiéndose después al segundo, donde este SNP se queda fijado (170,171).

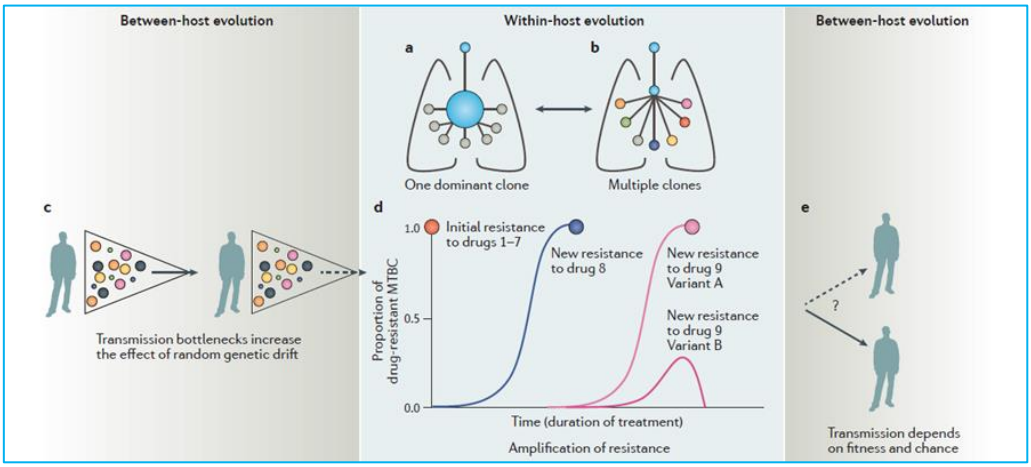


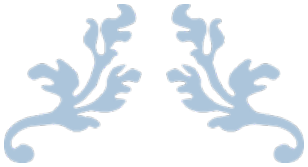
Figura 12. Papel de la selección natural y de la deriva genética en la evolución y la emergencia de resistencias a antibióticos. Los cuellos de botella que se producen durante la transmisión aumentan el efecto de la deriva genética aleatoria. Dentro del mismo hospedador pueden coexistir varios clones de la cepa original, cada uno con unas resistencias diferentes: uno de ellos puede dominar sobre los demás, o todos pueden coexistir en proporciones similares. Dependiendo de las diferentes presiones selectivas que estén actuando en cada momento, la dinámica de los clones puede alterarse y eso puede provocar un cambio en el perfil de resistencias que se había trazado de la cepa original (43).

Finalmente, la WGS incorporada de forma rutinaria en un laboratorio de microbiología clínica permite estudiar simultáneamente todos los genes relacionados con las resistencias de una forma automatizada, y obtener el perfil de resistencias de una forma más rápida que los métodos de cultivo tradicional. Los estudios llevados a cabo en este campo han intentado discernir si la cepa resistente se está transmitiendo (resistencia primaria) o si la resistencia está surgiendo de forma separada en distintos individuos (resistencia secundaria o adquirida) (157) (Figura 12). La resistencia a rifampicina está causada por mutaciones en el gen *rpoB*; la resistencia a isoniazida por mutaciones en *katG*, *inhaA* y *kasA*; la resistencia a etambutol por mutaciones en *embB* y *ubiA*; la resistencia a pirazinamida por mutaciones en el gen *pncA*; la resistencia a estreptomicina puede estar causada por mutaciones en los genes *rpsL*, *rrs* y *gidB*; la resistencia a fluoroquinolonas está causada principalmente por mutaciones en *gyrA*; la resistencia a fármacos inyectables como la

capreomicina, amikacina y kanamicina está causada por mutaciones en los genes *rrs* y *eis*; la resistencia a etionamida viene dada por mutaciones combinadas en los genes *ethA*, *mshA*, *ndh*, *inhA* y en el promotor de *inhA*; la resistencia a bedaquilina se produce por mutaciones en los genes *Rv0678*, *atpE* y *pepQ*; finalmente, la resistencia a linezolid se produce por mutaciones en *rplC* y *rrl* (172). Estas mutaciones están en continua revisión, y se van incluyendo más conforme van apareciendo evidencias clínicas de que confieren un fenotipo de resistencia (173).

En la presente tesis doctoral se incluyen tres trabajos (Publicaciones 4, 5 y 6) que estudian tres brotes ocurridos en Aragón causados por cepas con diferentes características (ara7, ara50 y ara217), en los que se analizaron las características moleculares de las cepas y se trató de esclarecer la cadena de transmisión utilizando WGS. Así mismo, también se incluye un trabajo (Publicación 7) en el que se hace un estudio general de las cepas causantes de los 26 brotes más grandes que han tenido lugar en la comunidad utilizando WGS. En todos ellos se han utilizado, en mayor o menor medida, las aplicaciones de la WGS en epidemiología molecular descritas en este apartado.

Además, también se incluye en la presente tesis doctoral un estudio (Trabajo 8) en el que se ha utilizado la tecnología de secuenciación masiva AmpliSeq para obtener el genotipo y las resistencias de las cepas directamente de muestras clínicas recopiladas en el Hospital Universitario Miguel Servet a lo largo del tiempo de realización de esta tesis.



Objetivos

2. Objetivos

Objetivo principal

El objetivo principal de esta tesis doctoral es introducir la tecnología de secuenciación masiva para el estudio de la TB en Aragón.

Objetivos específicos

1. **Estudiar la situación en Aragón de la TB producida por *M. africanum*** (Publicación 1)
   1.1. Describir *M. africanum* desde el punto de vista epidemiológico y genotípico para aumentar el conocimiento sobre este linaje del MTBC.

   1.2. Analizar las secuencias de inserción IS*6110* de *M. africanum* L6 (bajo número de copias) para que nos permita su identificación rápida.

2. **Estudiar la variabilidad de la secuencia IS*6110*** (Publicación 2)
   2.1. Explorar la variabilidad genética de la secuencia de inserción IS*6110* intra-cepa e inter-cepa.
   2.2. Analizar si alguna mutación podría relacionarse con la habilidad de transposición.

3. **Estudiar la TB recurrente en Aragón** (Publicación 3)
   3.1. Estudiar la frecuencia de reactivación y reinfección en la población aragonesa.

   3.2. Estudiar los factores de riesgo asociados a una reactivación más temprana de la enfermedad.

   3.3. Estudiar la correlación entre el número de SNPs y el tiempo de latencia, y la tasa de mutación frente al tiempo de latencia, tanto para tiempos de generación fijos como variables.

4. **Estudiar tres brotes de comportamiento epidemiológico diferente mediante secuenciación masiva** (Publicaciones 4, 5 y 6)
   4.1. Analizar factores de riesgo y posibles contactos del caso con otras personas infectadas con TB para determinar cadenas de transmisión.

   4.2. Secuenciar el genoma de un número representativo de aislados de cada brote para llevar a cabo un estudio genómico: asignación de un linaje y familia del MTBC, estudio de SNPs y construcción de un dendrograma para estudiar la cadena de transmisión y evolución de las cepas, y estudio de SNPs en factores de virulencia que pudieran explicar la propagación de estas cepas en la población aragonesa.

   4.3. Estudiar la transcriptómica de tres aislados de la cepa MtZ con diferente localización de copias de IS*6110* con el principal objetivo de analizar las diferencias en la transcripción de los genes adyacentes. Estudiar la secreción de

PE_PGRS de la cepa MtZ con el objetivo de comprobar si comparte con las cepas de la familia Beijing un fenotipo hipervirulento (Publicación 6).

5. **Caracterizar las cepas de los 26 brotes más grandes de Aragón** (Publicación 7)
   5.1. Explorar las características que hacen a estas cepas más virulentas y/o transmisibles.

   5.2. Disponer de los genomas para futuros protocolos de vigilancia de la epidemiología de la TB en Aragón.

6. **Desarrollar un método de diagnóstico rápido de resistencias en muestras clínicas** (Trabajo 8)
   6.1. Poner a punto un protocolo de extracción de ADN directamente de muestra clínica para llevar a cabo la detección de mutaciones mediante la tecnología AmpliSeq.

   6.2. Comprobar la sensibilidad de los resultados obtenidos con el AmpliSeq en relación a la asignación de familias/linajes del MTBC y en la detección de resistencias, tanto para muestras cuyo ADN se ha extraído de cultivo como para las que se ha extraído directamente de la muestra clínica.

Resultados y Discusión

3. Resultados y Discusión

Publicación 1: Analysis of *Mycobacterium africanum* in the last 17 years in Aragon identifies a specific location of IS*6110* in Lineage 6.

Publicación 2: In-depth Analysis of IS*6110* Genomic Variability in the *Mycobacterium tuberculosis* Complex.

Publicación 3: Estimation of the mutation rate of *Mycobacterium tuberculosis* in cases with recurrent tuberculosis using whole genome sequencing.

Publicación 4: Investigation of a rapidly spreading tuberculosis outbreak using whole-genome sequencing.

Publicación 5: A whole-genome sequencing study of an X-family tuberculosis outbreak focus on transmission chain along 25 years.

Publicación 6: The MtZ Strain: Molecular Characteristics and Outbreak Investigation of the Most Successful *Mycobacterium tuberculosis* Strain in Aragon Using Whole-Genome Sequencing.

Publicación 7: Analysis of the twenty-six largest outbreaks of tuberculosis in Aragon using whole-genome sequencing for surveillance purposes.

Trabajo 8: AmpliSeq technology for rapid lineage and drug-resistance identification in clinical samples of *Mycobacterium tuberculosis*.

Publicación 1

scientific reports

Check for updates

OPEN

Analysis of *Mycobacterium africanum* in the last 17 years in Aragon identifies a specific location of IS*6110* in Lineage 6

Jessica Comín[1,2], María Luisa Monforte[3], Sofía Samper[1,2,4,6]✉, Aragonese Working Group on Molecular Epidemiology of Tuberculosis (EPIMOLA)* & Isabel Otal[2,4,5]

The purpose of this study was to increase our knowledge about *Mycobacterium africanum* and report the incidence and characteristics of tuberculosis (TB) due to their lineages in Aragon, Spain, over the period 2003–2019. The study includes all the cases in our region, where all the *M. tuberculosis* complex isolates are systematically characterised. We detected 31 cases of *M. africanum* among 2598 cases of TB in the period studied. TB caused by *M. africanum* is rare (1.19%) in our population, and it affects mainly men of economically productive age coming from West African countries. Among the isolates, Lineage (L) 6 was more frequent than L5. The genotyping of these strains identified five clusters and 13 strains with a unique pattern. The isolates' characterisation identified a copy of IS*6110* within the *moaX* gene, which turned out to be specific for L6. It will allow the differentiation of this lineage from the rest of MTBC with a simple PCR reaction. It remains to be established whether this polymorphism may limit *M. africanum* transmission. Furthermore, a mutation in the *mutT2* promoter was found as specific for L6 strains, which could be related to the high variability found for L6 compared to L5.

Tuberculosis (TB) is still an important cause of death, especially in developing countries[1]. All members of the *Mycobacterium tuberculosis* complex (MTBC) can cause TB, with *M. tuberculosis* being the most important. However, the strains from *M. africanum* lineages (L) are responsible for almost half of the TB cases in West Africa[2]. *M. africanum* was first described in 1968 from TB patients in Senegal[3], showing intermediate characteristics between *M. tuberculosis* and *M. bovis*[4,5]. There are three hypotheses being considered of why *M. africanum* is almost restricted to West Africa: it is not able to compete with modern *M. tuberculosis* lineages[6]; it is adapted to the African population[7,8]; and there could be an animal reservoir, being a zoonotic disease[5,9–12].

Regarding phylogeny, both *M. africanum* L5 and L6 have a common ancestor with the region of difference (RD) 9 deleted[13,14]. Besides, L5 split from the common phylogenetic branch before L6, and the latter also has deleted RD7, RD8, and RD10 regions. Recent studies subdivided L5 into two sub-lineages, L5.1 and L5.2, discriminated by RD711[15].

The prevalence of *M. africanum* is rare in industrialised countries, such as in Spain; a retrospective study over 10 years (2000–2010) was published[16]. They analysed 36 cases due to *M. africanum* and concluded that most of them were immigrants from Africa, and only four cases were Spaniards. However, not all cases were searched exhaustively, considering that they did not systematically genotype the isolates.

The insertion sequence (IS) *6110*, specific to MTBC, has been used to genotype these strains since 1993[17]. Some studies analysed the location of the IS*6110* copies trying to clarify its role in the bacteria's physiology[19,20] or revealing that some of the IS*6110* locations are characteristic of some specific families[21,22]. However, to our knowledge, the location of IS*6110* in the genome of *M. africanum* strains has not been studied until now.

In this work, we used an epidemiological and molecular perspective to investigate the presence of *M. africanum* for the last 17 years in Aragon, Spain. The molecular analysis of these strains allowed us to identify specific polymorphisms described for the first time in these lineages.

[1]Instituto Aragonés de Ciencias de la Salud, Zaragoza, Spain. [2]Fundación IIS Aragón, Zaragoza, Spain. [3]Hospital Royo Villanova, Zaragoza, Spain. [4]CIBER de Enfermedades Respiratorias, Zaragoza, Spain. [5]Universidad de Zaragoza, Zaragoza, Spain. [6]Laboratorio de Investigación Molecular-UIT, Hospital Universitario Miguel Servet, Pº Isabel la Católica 1-3, planta calle, 50009 Zaragoza, Aragón, Spain. *A list of authors and their affiliations appears at the end of the paper. ✉email: ssamper.iacs@aragon.es

	N = 31	%
Sex		
Male	24	77.41
Female	7	22.59
Age group		
15–24	6	19.35
25–34	14	45.16
35–44	8	28.80
45–64	2	6.45
Unknown	1	3.22
Origin		
Immigrant	28[a]	90.32
Autochthonous	3	9.67
Location of the disease		
Pulmonar only	16	51.61
Extrapulmonary	15	48.38
Municipality		
Urban	20	64.51
Rural	11	35.48

Table 1. Patients' characteristics whose isolates were identified as *M. africanum* in Aragon from 2003 to 2019. [a]27 immigrants came from West African countries and one came from Bulgaria.

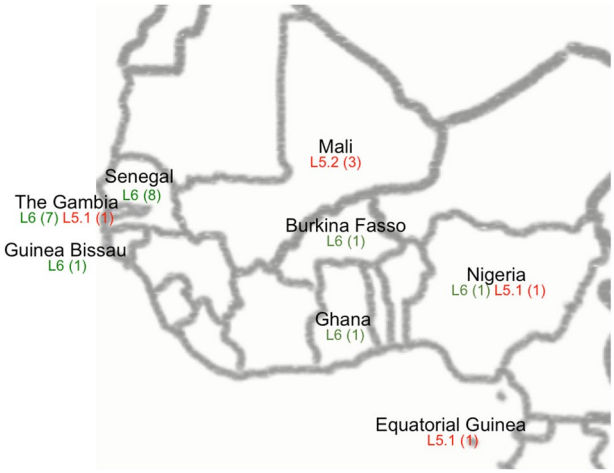
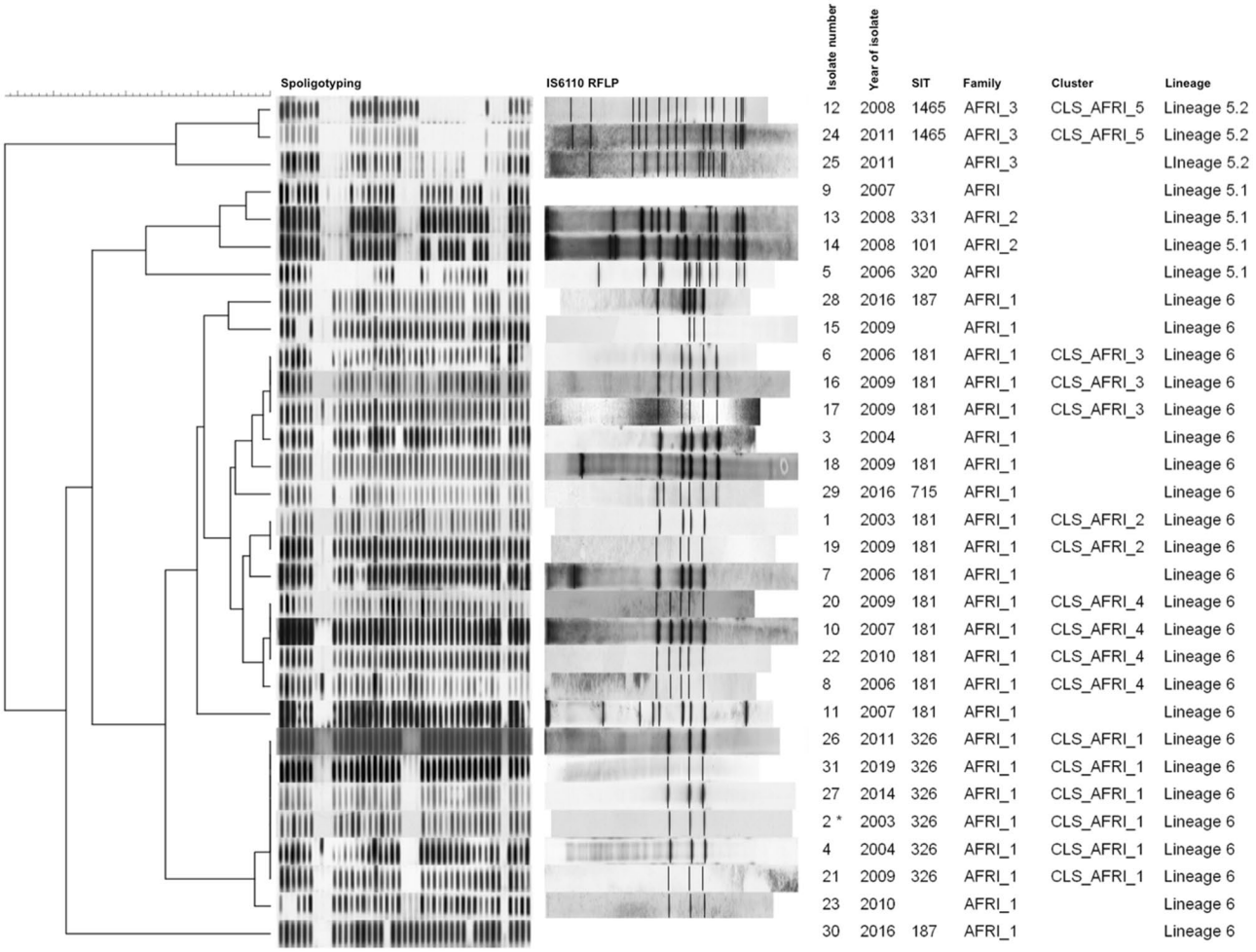


Figure 1. Map drawn with Adobe Photoshop CS6 (www.adobe.com) of the West African countries of origin for 26 of the cases. African countries where the patients came from are marked. The red colour indicates an L5 case of TB, and the green colour an L6 case. The number of cases is presented in brackets.

Results

A retrospective descriptive study of the TB cases caused by isolates identified as *M. africanum* in the Autonomous Community of Aragon was carried out. All TB cases with positive culture in Aragon between 2003 and 2019 were genotyped. Also, spoligotyping supplied the ability to distinguish among the different members of the complex, such as *M. tuberculosis, M. bovis*, and *M. africanum*. Out of 2598 MTBC cases of TB over this 17-year period, 31 cases (1.19%) have been caused by an African spoligo-family strain.

Social and clinical characterisation of the TB cases. The characteristics of the cases due to *M. africanum* are detailed in Table 1. Of the 31 patients studied, 77.41% were male, the age range was between 18 and 62 years, with the largest in the 25–34-year age-group (45.16%). No cases were detected in the youngest and the eldest age groups. Regarding their origin, 27 patients were born in West African countries (87.09%), three were Spaniards (9.67%), and one patient was born in Bulgaria. The African countries of origin are detailed in Fig. 1. According to the geographical location, the number of patients who lived in an urban area was superior to those who lived in a rural area (20, 64.51% vs 11, 35.48%). At least 12 of the cases lived in our country for less than



Isolate number	Year of isolate	SIT	Family	Cluster	Lineage
12	2008	1465	AFRI_3	CLS_AFRI_5	Lineage 5.2
24	2011	1465	AFRI_3	CLS_AFRI_5	Lineage 5.2
25	2011		AFRI_3		LIneage 5.2
9	2007		AFRI		Lineage 5.1
13	2008	331	AFRI_2		Lineage 5.1
14	2008	101	AFRI_2		Lineage 5.1
5	2006	320	AFRI		Lineage 5.1
28	2016	187	AFRI_1		Lineage 6
15	2009		AFRI_1		Lineage 6
6	2006	181	AFRI_1	CLS_AFRI_3	Lineage 6
16	2009	181	AFRI_1	CLS_AFRI_3	Lineage 6
17	2009	181	AFRI_1	CLS_AFRI_3	Lineage 6
3	2004		AFRI_1		Lineage 6
18	2009	181	AFRI_1		Lineage 6
29	2016	715	AFRI_1		Lineage 6
1	2003	181	AFRI_1	CLS_AFRI_2	Lineage 6
19	2009	181	AFRI_1	CLS_AFRI_2	Lineage 6
7	2006	181	AFRI_1		Lineage 6
20	2009	181	AFRI_1	CLS_AFRI_4	Lineage 6
10	2007	181	AFRI_1	CLS_AFRI_4	Lineage 6
22	2010	181	AFRI_1	CLS_AFRI_4	Lineage 6
8	2006	181	AFRI_1	CLS_AFRI_4	Lineage 6
11	2007	181	AFRI_1		Lineage 6
26	2011	326	AFRI_1	CLS_AFRI_1	Lineage 6
31	2019	326	AFRI_1	CLS_AFRI_1	Lineage 6
27	2014	326	AFRI_1	CLS_AFRI_1	Lineage 6
2 *	2003	326	AFRI_1	CLS_AFRI_1	Lineage 6
4	2004	326	AFRI_1	CLS_AFRI_1	Lineage 6
21	2009	326	AFRI_1	CLS_AFRI_1	Lineage 6
23	2010		AFRI_1		Lineage 6
30	2016	187	AFRI_1		Lineage 6

Figure 2. Dendrogram based on spoligotype patterns. The data shown are spoligotypes, IS*6110*-RFLP types, isolate number, year of isolation, SIT, family under SITVIT definition, and lineage of each *M. africanum* isolate in our population from 2003 to 2019. For two isolates, RFLP-type was not available, both showing a unique spoligotype belonged to AFRI and AFRI1 families. *Isolate 2 in CLS_AFRI_1 has a different location of one copy of IS*6110* despite sharing the RFLP pattern with the other five isolates included in its cluster.

5 years when TB was diagnosed. Attending to the location of the disease, the samples studied were: 16 sputum, four bone biopsies, six abscesses from different locations, two lymphadenopathies, and three other specimens. Near 50% of the cases (15) presented extrapulmonary disease (Table 1). For all cases, TB bacillus were susceptible to the treatment.

Genotypic characterisation of *M. africanum* isolates. The molecular analysis based on IS6110-RFLP and Spoligotyping of the *M. africanum* isolates showed five different clusters, including from two to six cases, and 13 isolates with a unique pattern (Fig. 2). Spoligotyping showed 13 different patterns, three were detected more than once (SIT 181 in 13 isolates, SIT 326 in six isolates, and SIT 1465 in two isolates) and distributed in the AFRI_1, AFRI_2, AFRI_3 or AFRI families according to the SITVIT definition[23]. To confirm the *M. africanum* lineages, the specific differential regions TbD1, RD9, RD702, and RD711 were analysed. TbD1 was present, and RD9 absent in all the isolates. Based on the study of RD702 and RD711, we could classify them into the two existent lineages of *M. africanum*. Twenty-four isolates belonged to L6 and seven to L5. There was a total concordance in the classifications obtained by spoligotyping and the RD analysis. The isolates classified as AFRI1 had the RD702 region deleted, and therefore corresponded to L6 isolates[2]. AFRI, AFRI_2, and AFRI_3 strains had RD702 present and, therefore, were considered L5. Two AFRI and two AFRI_2 spoligotype isolates had deleted RD711 and were sub-classified as L5.1. Meanwhile, the three AFRI_3 isolates had RD711 present, classifying them as L5.2[15]. The IS6110-RFLP showed a low number of IS6110 copies (≤ 6) in the L6 isolates, except one which showed eight copies. However, the six L5 isolates with available RFLP-pattern presented more than 10 IS6110 copies (Fig. 2).

Following the strain characterisation and in the context of a new assay performed in our laboratory to rapidly identify resistances and lineages, 32 isolates of our DNA collection, including different MTBC lineages, were analysed using AmpliSeq-based methodology. Two of the 32 isolates belonged to L6 (isolates 15 and 27) and three

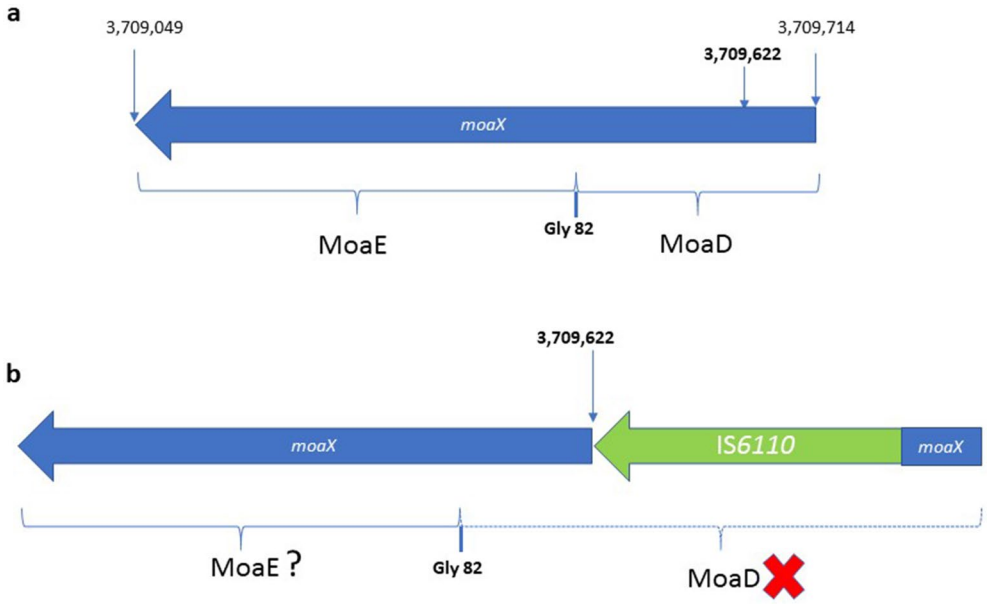


Figure 3. Schematic representation of *moaX* (*Rv3323c*) gene, coding for MoaX protein. The cleavage by Gly82 residue is required for the functionality of this MPT synthase. (**a**) Schema of H37Rv. (**b**) Schema of *M. africanum* L6 with IS*6110* inserted in *moaX* gene. The effect of IS*6110* is unknown for MoaE, but MoaD is going to be unfunctional. Numbers are referred to the position of the nucleotide in the genome of *M. tuberculosis* H37Rv.

to L5.1 (isolates 5, 13, and 14). The sequence of the amplicons obtained showed five specific SNPs for L6 isolates located in *rpoB* (1163c/t[24] and 1917a/c mutations, non-synonymous SNPs), *inhA* (233t/c, non-synonymous SNP), *katG* (609c/t, synonymous SNP), and *Rv0309* (474g/a, synonymous SNP) genes. These SNPs were reviewed and confirmed in NCBI L6 complete genomes. In the three L5 isolates studied, a SNP was present in *gyrA* (2265c/t, synonymous SNP). Finally, one specific SNP was detected in *leuB* (550t/c, non-synonymous SNP) in the isolates of both lineages L5.1 and L6 and absent in the rest of the isolates of the MTBC studied by AmpliSeq.

Specific IS*6110* location in L6 strains. We studied the location of IS*6110* in three L6 strains (isolates 2, 15, and 21) using ligation-mediated PCR (LM-PCR) within a study of MTBC strains with a low copy number of this IS. In addition to the copy located in the DR area, three locations were detected in *Rv0963c*, *lipX:mshB* and *moaX* genes. In the three strains, one of the IS*6110* copies was located in the *moaX* gene and at identical point for all three cases (Fig. 3). Based on the results obtained, the primers moaXr (ccagtcgacgcggttgggg) and moaXd (atcgggtcattaccggcggc) were designed to verify the point of insertion of IS*6110*. The expected PCR products were 2128 bp if IS*6110* was present and 788 bp if IS*6110* was absent from the site of amplification. We sequenced the amplified fragment noting that IS*6110* was inserted at nucleotide 3709622, referred to H37Rv reference genome, flanked by three bp direct repeats (gac), as a consequence of the transposition, and located 90 nucleotides from the beginning of the *moaX* gene (*Rv3323c*) and in its same direction (Fig. 3). Further analysis showed that this IS*6110* copy was present in all our collection strains of *M. africanum* belonging to the L6 but never in our L5 strains. In addition, we observed that this location was absent in 42 isolates of low copy number studied by our group. We have also analysed this insertion point in the strains belonging to *M. africanum* L6, whose genomes are available in the NCBI (CP010334.1 and FR878060.1), verifying the presence of IS*6110* in the *moaX* gene (Fig. 4). On the other hand, to investigate the intergenic IS*6110* insertion in *lipX:mshB* as a possible specific location for *M. africanum*, we amplified the region with the primers LipX-F (gccgtttccccaatcgaatc) and LipX-R (gctcaggctctcatcgtcg). The expected fragment was 264 bp if the IS was absent and 1591 bp if it was present. The PCR results revealed the insertion in five out of nine L6 isolates tested and never in L5 isolates, which means it was not specific but frequent in L6 strains. IS*6110* was inserted at nucleotide 1300194, flanked by two bp direct repeats (tt), in all the isolates at the same point, including those in the NCBI database. However, the location of IS*6110* in *Rv0963c* was not detected in any other of the *M. africanum* isolates analysed.

Nitrate reduction activity absent in *M. africanum* strains. Due to the location of IS*6110* interrupting the *moaX* gene, which codes for the enzymes involved in the synthesis of molybdenum cofactor (MoCo), necessary for the activity of the nitrate reductase (NR) enzyme[25], we wanted to investigate whether this fact would be reflected in a difference in NR activity between the L5 and L6 strains. We analysed the reduction of nitrates of the *M. africanum* isolate 5 (L5) and isolate 11 (L6), and H37Rv and BCG as positive and negative controls, respectively. However, both L5 and L6 strains showed a negative result of NR activity. Both positive and negative controls were in line with expectations. Then, we analysed the sequences of *narG* and *narI* genes, which did not present any mutation to explain the L5 strain result. Nevertheless, the analysis of the *narGHIJ* promoter

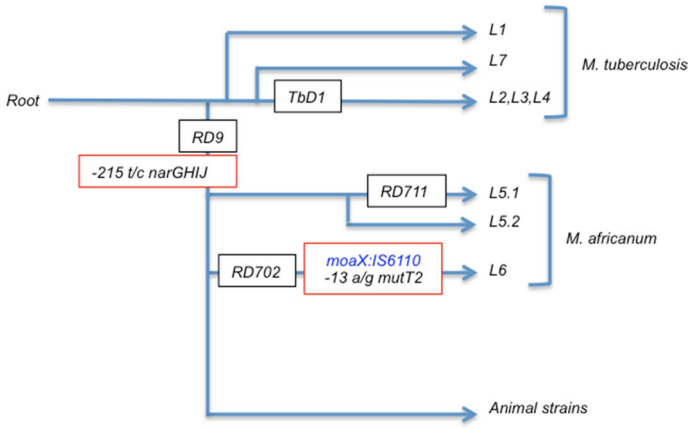


Figure 4. Partial evolutionary scenario of MTBC detailing *M. africanum* lineages. IS*6110* is involved in the natural evolution of MTBC, and the seemingly random transposition may have contributed to the differentiation of MTBC. Mutations and specific insertion of IS*6110* in the *moaX* gene found in this work are red-framed.

in four L5 isolates and two L6 isolates showed an identical mutation in − 215 (t/c), which was also present in the NCBI complete genomes of *M. africanum* (L6) and *M. bovis*. Additionally, the study of the sequenced promoter region of the *narGHIJ* operon showed a mutation in the − 13 (a/g) *mutT2* gene, upstream *narG* gene, in the two L6 isolates, which was absent in the four L5 isolates tested. This was also observed in the *M. africanum* L6 genomes included in the NCBI database.

Discussion

This work was carried out to understand the epidemiological situation in Aragon, Spain, related to TB cases caused by *M. africanum*. This study used the data set on TB cases linked to the genotypes of the clinical isolates. The findings from this study indicate that *M. africanum* is a rare cause of TB in our region and represents 1.19% of the cases with available genotype data reported during the 2003–2019 period. A previous study regarding this causal agent[16] reviewed the percentage detected in other countries, as Brazil, Australia and Portugal, where it represented less than 1% of the TB isolates. In their work, the authors collected information on 36 TB cases of *M. africanum* over a 10-year period in Spain. Nineteen of these isolates were from our region and therefore included in this study. The fact that they did not systematically identify the *M. africanum* isolates leads us to believe that there was an underestimation of the TB cases caused by L5 and L6 in our country. Nevertheless, our study was exhaustively conducted since 2003, identifying all *M. africanum* cases. We consider that the incidence of *M. africanum* in our country should be low given the results observed in this study, even though in higher African migration areas it could be slightly different. We hypothesised that the African lineages that are rare in our population are not adapted to transmit.

The descriptive analysis of the TB cases caused by *M. africanum* showed that most were male (77.41%) and in the 25–34 age group (45.16%). All of them were in the labour force, which could be related to being the most abundant age group among immigrants. It was more likely to occur in foreign-born people coming from West African countries (87%), being that only three cases (9.67%) were of Spanish-born people. In Esteban's study[16] performed in Spain, few Spaniards (7%), in contrast to immigrants, presented TB caused by *M. africanum*. The slight difference obtained by us may be because we did an exhaustive genotyping of all the cases. In our study, the patients came from different rural areas in a higher percentage (35%) than the common TB caused by *M. tuberculosis*, which occurred around 80% in urban areas[26]. These associations suggest that the epidemiology of *M. africanum* in our region is driven primarily by the migration of people from West Africa. The TB in Spaniards suggests that transmission of *M. africanum* might occur in Spain, but the possibility of TB acquisition during a trip (e.g., to West Africa) cannot be excluded, as one of the Spaniards presented a unique genotype strain. It would be of interest to continue the study in the coming years to check if the *M. africanum* strains of this work are maintained or are displaced by other MTBC strains. In previous reports, a lower transmission of *M. africanum* in comparison to L4 was observed. Nevertheless, the proportion of L5 and L6 is maintained over time, suggesting that other factors may be responsible for its continued presence in Africa[27].

The presentation of the disease was in half of the cases restricted to pulmonary location. The extrapulmonary type of the disease (48.38%) was identified in a higher percentage than for the TB case notifications in Spain in 2017 (27.5%)[28]. Some studies showed a high proportion of extrapulmonary TB caused by L5 strains, suggesting that these strains might show a different ability to cause pulmonary disease than *M. tuberculosis* sensu stricto strains[15]. Curiously, while the extra respiratory location of the TB was high among the African cases, the three Spaniards presented only respiratory disease. The differential HLA distribution among the Mali population has been studied, and it was concluded that it might be at least partially responsible for the geographical restriction of *M. africanum* infections to West Africa[29]. The possibility that HLA could also affect the clinical presentation of the

disease would explain these differences. Other studies support the hypothesis that *M. africanum* has a low degree of virulence that may be related to dissemination, rather than lung damage, during the early stages of infection[30].

We could detect a higher percentage of L6 strains (77.41%) than L5 (22.58%), including both L5.1 and L5.2 sub-lineage cases. Other studies in Mali and Gambia also showed a higher identification of L6 among their cases[2,31]. A systematic review of current knowledge about MTBC strain diversity and geographical distribution in African regions showed a different prevalence of *M. africanum* in West African countries. It represented 8% in Nigeria, 19.75% in Ghana, 20% of the isolates in Burkina Fasso, 3.3% in Guinea, 47.10% in Guinea Bissau, and 38.4% in the Gambia[32]. Previous results published about Gambia reflected that half of their isolates were *M. africanum*, and nearly all of them belonged to L6 with SIT 181 as the most prevalent pattern. According to our results, seven out of eight *M. africanum* isolates of patients coming from Gambia were L6 isolates, and five of them had SIT 181[2]. Traore et al.[31] determined 27.8% of the cases as *M. africanum* in Mali, and almost all of them (94.2% of the strains) were MAF2 (L6). However, the three cases detected from Mali in our study were L5.2.

For all cases, TB bacillus were susceptible to the treatment, although the Ampliseq method applied in seven *M. africanum* isolates detected some mutations in genes related to resistance. These polymorphisms could be specific evolutionary characteristics of the respective lineages. These results indicated that we must be cautious when reporting resistant genotypes, such as the mutations found in this study, which do not confer a resistance phenotype. Nevertheless, they could be assessed as specific for L6 (*rpoB, inhA*, and *katG* genes) and L5 (*gyrA gene*).

RFLP showed a substantial difference in the number of IS*6110* copies between the L5 and L6 strains. L6 strains carried a lower number of copies in contrast to L5 strains. Spoligotyping and IS*6110*-RFLP allowed us to detect five clusters, including 17 cases. Although each technique has low discriminatory power separately, especially among low copy number strains, it increases when considered together. On the other hand, the location of some of the IS*6110* insertion points adds differentiation capacity to the RFLP, as indicated in other publications where it has been described that RFLP analysis can underestimate the real copy number for the IS*6110* element[33,34]. In this work, the isolates 2 and 21 present three bands that seem identical when observing their RFLP pattern. However, they share two locations (DR region and *moaX*) but isolate 21 has an IS*6110* in the *Rv0963* gene, which does not share isolate 2 (Fig. 2). This indicates that it can happen in some cases that the coincidence of a band in the RFLP pattern does not imply that the IS*6110* insertion point is the same. The explanation for this would be that a small difference between the lengths of the restriction fragment generated for two different locations of IS*6110* is not appreciable in the RFLP pattern. Despite this, transmission was not considered in this study as it could overestimate the recent transmission rate.

In the context of a study of the IS*6110* location in low copy number strains, we discovered an insertion within the *moaX* gene for the L6 strains analysed, and later we verified its presence in all L6 strains but never in L5 or other MTBC families studied. A previous work studied the insertion points of IS*6110* in high-copy clinical isolates, specifically focusing on the Beijing genotype and revealed that its location in *moaX* gene was not characteristic of Beijing family[22]. Also, we found that in a previous work where the locations of IS*6110* were studied in 579 MTBC strains representatives of the major lineages circulating in Europe and Latin America, the location of IS*6110* in *moaX* was not detected in any case[21]. In all L6 strains included in our collection and the strains whose genomes are available online, the insertion point was the same. Altogether, it strongly suggests that this location is specific for L6, allowing us to differentiate this lineage from the rest of the strains of the MTBC. Within the scheme of the evolutionary stage of the tubercle bacillus, proposed by Brosch et al.[13], we suggest the transposition of IS*6110* into the *moaX* gene when L6 is separated from the rest of the lineages (Fig. 4). Besides, the location in *lipX:mshB* was frequent in L6 strains. These results agree with previous observations, indicating that each family has preferential insertion sites[21,22,35], which is probably related to their evolutionary relationship.

The *moaX* gene encodes a molybdopterin (MPT) synthase with *moaD* and *moaE* activity that contributes to molybdenum cofactor (MoCo) synthesis in MTBC[25]. It has been shown that there is functional interchangeability between the MPT synthase subunits of *M. tuberculosis,* and in the case of MoaX, post-translational cleavage at the Gly82 residue is required for the functionality of this enzyme[36]. According to that, the IS*6110* inserted in *moaX* gene of L6 strains is interrupting the MoaD subunit (Fig. 3). It has been described that some mutants in genes involved in molybdopterin biosynthesis had lost their ability to resist phagosome acidification[37]. In most molybdenum-containing enzymes, the metal is coordinated to the dithiolene group of MPT to form MoCo. Enzymes that utilise MoCo harness the redox properties of molybdenum to catalyse redox reactions in carbon, nitrogen, and sulfur metabolism and to reduce terminal electron acceptors for anaerobic respiration[25]. One of these enzymes is NarG, a membrane-bound respiratory NR, suggesting a potentially important role for MoCo in the metabolism of *M. tuberculosis* in vivo. In an anaerobic environment, many bacteria can use nitrate as a final electron acceptor. Historically, *M. tuberculosis* has been differentiated from *M. bovis* because only *M. tuberculosis* can reduce significant amounts of nitrate (NO_3^-) to nitrite (NO_2^-). NR activity occurs at a low level during the aerobic growth of *M. tuberculosis* and increases significantly upon entry into the microaerobic stage. When we discovered the IS*6110* insertion in *moaX* for L6 strains, we expected to find differences in NR activity between L5 and L6, but none showed NR activity. This indicates that the disruption of the MoaD subunit from MoaX in the L6 strains is not the only one responsible for the lack of activity observed in vitro. This result supports the hypothesis that homologous genes could compensate for any adaptive disadvantage of the bacteria due to the natural knockouts created by IS*6110* insertion or other mutations[25]. Looking for another explanation for this result, we analysed the operon *narGHIJ* implicated in NR activity. The first mutation described[38] that prevented NR activity was – 215 (t/c) SNP in the promoter of *narGHIJ* operon for *M. bovis*. *M. africanum* L5 and L6 have this mutation, but also *M. canetti*, which has NR activity[39]. There is another region responsible for NR activity, the *narK2* operon. A mutation in – 10 promoter elements of the *narK2* operon reduced NR activity in BCG[40,41]. We found this mutation in the L6 strains available in NCBI but not in *M. canetti*, which had the same genotype as H37Rv. It seems that the presence of both mutations could explain the lack of NR activity we observed for *M.*

africanum L5 and L6. However, in latent anaerobiosis, BCG overexpressed the *narX* gene, a fused NR[42]. Thus, a similar enzyme could play this role for *M. africanum*.

Surprisingly, the search for mutations in the *narGHIJ* promoter led us to the location of a mutation in − 13 (a/g) *mutT2* gene, upstream of this operon, in the L6 strains analysed and in the NCBI complete genomes of *M. africanum* (L6), but not in L5 strains analysed nor in other TB genomes available in NCBI. This gene was studied in the Beijing lineage as a possible cause of a major number of SNPs related to resistance[43]. It has been observed that L6 has a higher variability in its genome in comparison to L5, which could be related to a higher mutation rate[44,45]. MutT2 is involved in DNA repair, therefore the mutation detected in the *mutT2* promoter could increase the polymorphisms in L6 strains[46].

A possible limitation of this work is that the number of strains studied was low. Nevertheless, all the isolates have been exhaustively and systematically characterised in a continuous period of 17 years. Consequently, the results objectively reflect the incidence of *M. africanum* in our region. On the other hand, genotyping methods do not discriminate enough to analyse transmission, so that whole-genome sequencing of the isolates would be more informative.

In summary, the results of this study indicate that TB caused by *M. africanum* is rare in Aragon, and the majority of the cases were in immigrants from West Africa. L6 was more prevalent, with few cases of L5. As far as we know, this is the first time that IS*6110* locations have been determined in *M. africanum* strains, which has allowed us to detect the presence of a copy of IS*6110* in the *moaX* gene in all L6 strains. Further studies on the implication of interruption of MPT synthase subunit-encoding genes in the physiology of L6 strains and its possible relationship with lower virulence would be of interest. The analysis of this location showed that it is a specific characteristic of the L6 strains, which allows us to distinguish this lineage of *M. africanum* from the rest of MTBC in a simple and fast way, using a PCR-based test.

Material and methods

Origin of clinical isolates. In Aragon, a north region in Spain, all MTBC isolates are genotyped for surveillance purposes routinely since 2004, but 2003 isolates are also registered in the context of a previous study. In this work, we selected all patients with a microbiological diagnosis of TB caused by *M. africanum* between 1 Jan 2003 and 31 Dec 2019. The demographic (age, sex, country of birth, years since entry to Spain) and clinical (location of disease, sputum smear status, and previous diagnosis of TB) characteristics of the patients were retrospectively reviewed.

Genotyping. Genomic DNA was isolated using the cetyltrimethylammonium bromide (CTAB) method[47]. DNA was frozen at − 80 °C and used in the different molecular techniques in this study. All strains were systematically genotyped by restriction fragment length polymorphism (RFLP) based on IS*6110* and Spoligotyping. RFLP was performed as described by van Embden et al.[17]. Spoligotyping used a commercial membrane (Mapmygenome India Limited) to hybridise with the amplicons of the direct repeats region of each isolate. The procedure was previously described[48]. The genetic patterns were analysed by Bionumerics v7.6 software (Applied Maths, Kortrijk, Belgium) and introduced into the Database of the University of Zaragoza. TB cases caused by *M. africanum* were selected retrospectively by their spoligotype, a specific intermediate pattern between those of *M. tuberculosis* and *M. bovis,* according to the SITVIT definition[23]. Isolates were considered in cluster if they carried an identical IS*6110*-RFLP pattern and the same spoligotype if they had less than five copies of IS*6110*.

Study of differential regions. Subsequently, the presence or absence of the differential regions RD9, TbD1, RD702, and RD711 were analysed and used to classify the isolates into the different lineages of *M. africanum* L5, its sub-linages L5.1 and L5.2, and L6[15]. The PCR were performed using the following primers: TBD1fla1-F (ctacctcatcttccggtcca) and TBD1fla1-R (catagatcccggacatggtg) 2637/484 pb; RD9-flankF (gtgtaggtcagccccatcc) and RD9-flankR (gcccaacagctcgacatc) 2484/72 pb[13]; RD702-F (ccgcaacttcgagtaccttt) and RD702-R (gttgggttgctggtccat), and RD711-F (ggccgccctgctcaagaacct) and RD711-R (cctaggccggcgacgaagtg)[14].

Study of single polymorphisms. A panel of primers focused on genes related to resistance in MTBC, and SNPs for lineage differentiation was analysed by AmpliSeq-based methodology using next-generation sequencing**.** This panel of primers was designed to amplify the *gyrA* gene from 7302 to 9818 (2516 pb), *rpoB* from 759,807 to 763,325 (3518 pb), *rpsL* from 781,560 to 781,934 (374 pb), *inhA* promoter from 1,673,303 to 1,673,440 (137 pb), *inhA* from 1,674,102 to 1,675,011 (909 pb), *katG* from 2,153,889 to 2,156,211 (2322 pb), *pncA* 2,288,681 to 2,289,341 (660 pb), *eis* from 2,714,124 to 2,715,432 (1308 pb), and *embB* from 4,246,514 to 4,249,810 (3296 pb). In addition, other hotspots to identify the linages, specifically to identify *M. africanum* L5 (SNP in point 1377185, *Rv1234*) and L6 (SNP in point 378404, *Rv0309*), were amplified and analysed. Besides, the polymorphisms previously described for *M. bovis* in the *narGHIJ* operon were analysed by amplification of the different regions. Primers used were the following: mutT2F-2 (tccggatgatgatttacctcc) and mutT2R-2 (tccgccgggtcggggac)[43]; narG-Fw (gcccagctttgacaccatcg) and NarG-Rv (gcccagatgacgtttcgccag); NarI-Fw (tggctaccactcggaatgac) and NarI-Rv (acgatgtagggccggaacag). The detailed points are referred to as NC_000962_3.

Location of IS*6110* insertion sites. To study IS*6110* insertion sites, a ligation mediated PCR was used as described by Prod'hom et al.[49] to amplify one or both ends of each copy of IS*6110* and its flanking sequence. Briefly, genomic DNA was digested with *Sal*I enzyme and ligated to a linker containing a *Sal*I restriction site. The resulting template was then digested by *Sal*I. PCR was performed using ISA1 or ISA3, specific primers for IS*6110* and directed outwards from this element[50], and the linker primer Salgd. The template was initially denatured by incubation at 95 °C for 9 min and amplified by 35 cycles of PCR (95 °C for 30 s, 70 °C, and 72 °C for 90 s)

followed by a final extension at 72 °C for 10 min. Amplified products were separated by standard horizontal gel electrophoresis in a 1.5% agarose gel in tris–borate-EDTA buffer (90 mM tris, 90 mM boric acid, 2 mM EDTA) and stained with ethidium bromide. PCR products were purified, using GFX PCR DNA and Gel Band Purification Kit (Amersham Pharmacia Biotech) followed by ExoSAP-IT PCR Product Cleanup Reagent (Affymetrix), sequenced and analysed for homology with Tuberculist (http://genolist.pasteur.fr/TubercuList).

Enzymatic assay of NR. The NR activity test was performed with actively growing cultures, which were inoculated directly into phosphate buffer supplemented with nitrate and incubated for 2 h at 37 °C. The mycobacteria were cultured on 7H10 agar supplemented with 0.2% glycerol and 10% albumin/dextrose/catalase (ADC). One L5 strain and one L6 were inoculated into phosphate buffer supplemented with 10 mM nitrate. Following 2 h of incubation at 37 °C, naphthylamide and sulfanilic acid reagents were added, and the colour was then observed[51].

Computer analysis. The sequences generated were aligned and compared with the sequences of *M. tuberculosis* H37Rv (http://genolist.pasteur.fr/TubercuList) and *M. africanum* complete genomes, NC_015758.1 and CP010334.1 (http://blast.ncbi.nlm.nih.gov), using the Basic Local Alignment Search Tool (BLAST).

Ethics declarations. The permission to take informed consent was formally waived by the Comité de Ética de la Investigación de la Comunidad Autónoma de Aragón (CEICA), Spain, CI.PI18/068. No human tissues were used in the study. Once received the bacterial isolate, it was coded (NSTRAIN). The epidemiological data of the cases were sent by fax and were anonymised keeping only the code given to track the analysis of the clinical characteristics, to follow the Helsinki ethical principles for medical research involving human data. The experiment protocol followed was revised and approved by the CEICA and is in line with the Declaration of Helsinki, as revised in 2013.

Data availability

Sequences data reported in the present study were deposited in GenBank under accession numbers MW987573MW987574 and MW987575.

Received: 24 February 2021; Accepted: 23 April 2021
Published online: 14 May 2021

References

1. World Health Organization. Global Tuberculosis Report 2020. Licence: CC BY-NC-SA 3.0 IGO. Geneva. https://doi.org/10.3201/eid1910.121023 (2020).
2. Gehre, F. *et al.* Immunogenic *Mycobacterium africanum* strains associated with ongoing transmission in the Gambia. *Emerg. Infect. Dis.* **19**, 1599–1605. https://doi.org/10.3201/eid1910.121023 (2013).
3. Castets, M., Boisvert, H., Grumbach, F., Brunel, M. & Rist, N. Les bacilles tuberculeux de type Africain: Note préliminaire. *Rev. Tuberc. Pneumol.* **32**, 179–184 (1968).
4. Meyer, L. & David, H. L. Evaluation de l'active urease et de l'activite β-glucosidase pour l'identification pratique des mycobacteries. *Ann. Microbiol.* **130B**, 323–332 (1979).
5. Thorel, M. F. Isolation of *Mycobacterium africanum* from monkeys. *Tubercle* **61**, 101–104. https://doi.org/10.1016/0041-3879(80)90018-5 (1980).
6. Bold, T. D. *et al.* Impaired fitness of *Mycobacterium africanum* despite secretion of ESAT-6. *J. Infect. Dis.* **205**, 984–990. https://doi.org/10.1093/infdis/jir883 (2012).
7. Asante-Poku, A. *et al.* Molecular epidemiology of *Mycobacterium africanum* in Ghana. *BMC Infect. Dis.* **16**, 385. https://doi.org/10.1186/s12879-016-1725-6 (2016).
8. Asante-Poku, A. *et al. Mycobacterium africanum* is associated with patient ethnicity in Ghana. *PLoS Negl. Trop. Dis.* **9**, e3370. https://doi.org/10.1371/journal.pntd.0003370 (2015).
9. Coscolla, M. *et al.* Novel *Mycobacterium tuberculosis* complex isolate from a wild chimpanzee. *Emerg. Infect. Dis.* **19**, 969–976. https://doi.org/10.3201/eid1906.121012 (2013).
10. Rahim, Z. *et al.* Characterization of *Mycobacterium africanum* subtype I among cows in a dairy farm in Bangladesh using spoligotyping. *Southeast Asian J. Trop. Med. Public Health.* **38**, 706–713 (2007).
11. Alfredsen, S. & Saxegaard, F. An outbreak of tuberculosis in pigs and cattle caused by *Mycobacterium africanum*. *Vet. Rec.* **131**, 51–53. https://doi.org/10.1136/vr.131.3.51 (1992).
12. Gudan, A. *et al.* Disseminated tuberculosis in hyrax (*Procavia capensis*) caused by *Mycobacterium africanum*. *J. Zoo Wildl. Med.* **39**, 386–391. https://doi.org/10.1638/06-041.1 (2008).
13. Brosch, R. *et al.* A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl. Acad. Sci. USA* **99**, 3684–3689. https://doi.org/10.1073/pnas.052548299 (2002).
14. Gagneux, S. *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **103**, 2869–2873. https://doi.org/10.1073/pnas.0511240103 (2006).
15. Ates, L. S. *et al.* Unexpected genomic and phenotypic diversity of *Mycobacterium africanum* lineage 5 affects drug resistance, protein secretion, and immunogenicity. *Genome Biol. Evol.* **10**, 1858–1874. https://doi.org/10.1093/gbe/evy145 (2018).
16. Isea-Peña, M. C. *et al. Mycobacterium africanum*, an emerging disease in high-income countries?. *Int. J. Tuberc. Lung Dis.* **16**, 1400–1404. https://doi.org/10.5588/ijtld.12.0142 (2012).
17. Van Embden, J. D. A. *et al.* Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: Recommendations for a standardized methodology. *J. Clin. Microbiol.* **31**, 406–409 (1993).
18. Soto, C. Y. *et al.* IS6110 mediates increased transcription of the phoP virulence gene in a multidrug-resistant clinical isolate responsible for tuberculosis outbreaks. *J. Clin. Microbiol.* **42**, 212–219. https://doi.org/10.1128/JCM.42.1.212-219.2004 (2004).
19. Alonso, H. *et al.* Deciphering the role of IS6110 in a highly transmissible *Mycobacterium tuberculosis* Beijing strain, GC1237. *Tuberculosis* **91**, 117–126. https://doi.org/10.1016/j.tube.2010.12.007 (2011).
20. Gonzalo-Asensio, J. *et al.* New insights into the transposition mechanisms of IS6110 and its dynamic distribution between *Mycobacterium tuberculosis* complex lineages. *PLoS Genet.* **14**, e1007282. https://doi.org/10.1371/journal.pgen.1007282 (2018).

21. Reyes, A. *et al.* IS-seq: A novel high throughput survey of in vivo IS6110 transposition in multiple *Mycobacterium tuberculosis* genomes. *BMC Genom.* **13**, 249. https://doi.org/10.1186/1471-2164-13-249 (2012).
22. Alonso, H., Samper, S., Martín, C. & Otal, I. Mapping IS6110 in high-copy number *Mycobacterium tuberculosis* strains shows specific insertion points in the Beijing genotype. *BMC Genom.* **14**, 422. https://doi.org/10.1186/1471-2164-14-422 (2013).
23. Demay, C. *et al.* SITVITWEB—A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infect. Genet. Evol.* **12**, 755–766. https://doi.org/10.1016/j.meegid.2012.02.004 (2012).
24. Huard, R. C. *et al.* Novel genetic polymorphisms that further delineate the phylogeny of the *Mycobacterium tuberculosis* complex. *J. Bacteriol.* **188**, 4271–4287. https://doi.org/10.1128/JB.01783-05 (2006).
25. Williams, M. J., Kana, B. D. & Mizrahi, V. Functional analysis of molybdopterin biosynthesis in mycobacteria identifies a fused molybdopterin synthase in *Mycobacterium tuberculosis*. *J. Bacteriol.* **193**, 98–106. https://doi.org/10.1128/JB.00774-10 (2011).
26. López-Calleja, A. I. *et al.* Genotyping of *Mycobacterium tuberculosis* over two periods: A changing scenario for tuberculosis transmission. *Int. J. Tuberc. Lung Dis.* **11**, 1080–1086 (2007).
27. Asare, P. *et al.* Reduced transmission of *Mycobacterium africanum* compared to *Mycobacterium tuberculosis* in urban West Africa. *Int. J. Infect. Dis.* **73**, 30–42. https://doi.org/10.1016/j.ijid.2018.05.014 (2018).
28. WHO. *WHO|Global tuberculosis report 2019*. Geneva: World Health Organization; Licence: CC BY-NC-SA 3.0 IGO (2019). https://doi.org/10.1037/0033-2909.I26.1.78.
29. Kone, A. *et al.* Differential HLA allele frequency in *Mycobacterium africanum* vs *Mycobacterium tuberculosis* in Mali. *HLA.* **93**, 24–31. https://doi.org/10.1111/tan.13448 (2019).
30. Cá, B. *et al.* Experimental evidence for limited in vivo virulence of *Mycobacterium africanum*. *Front. Microbiol.* **10**, 2102. https://doi.org/10.3389/fmicb.2019.02102 (2019).
31. Traore, B. *et al.* Molecular strain typing of *Mycobacterium tuberculosis* complex in Bamako, Mali. *Int. J. Tuberc. Lung Dis.* **16**, 911–916. https://doi.org/10.5588/ijtld.11.0397 (2012).
32. Kone, B. *et al.* Exploring the usefulness of molecular epidemiology of tuberculosis in Africa: A systematic review. *Int. J. Mol. Epidemiol. Genet.* **11**, 1–15 (2020).
33. Otal, I. *et al.* Mapping of IS6110 insertion sites in *Mycobacterium bovis* isolates in relation to adaptation from the animal to human host. *Vet. Microbiol.* **135**, 406. https://doi.org/10.1016/j.vetmic.2007.11.038 (2008).
34. Millán-Lou, M. I. *et al.* In Vivo IS6110 profile changes in a *Mycobacterium tuberculosis* strain as determined by tracking over 14 years. *J. Clin. Microbiol.* **53**, 2359–2361. https://doi.org/10.1128/JCM.00607-15 (2015).
35. Comín, J. *et al.* A whole-genome sequencing study of an X-family tuberculosis outbreak focus on transmission chain along 25 years. *Tuberculosis* **126**, 102022. https://doi.org/10.1016/j.tube.2020.102022 (2021).
36. Narrandes, N. C., Machowski, E. E., Mizrahi, V. & Kana, B. D. Cleavage of the moaX-encoded fused molybdopterin synthase from *Mycobacterium tuberculosis* is necessary for activity. *BMC Microbiol.* **15**, 22. https://doi.org/10.1186/s12866-015-0355-2 (2015).
37. Brodin, P. *et al.* High content phenotypic cell-based visual screen identifies *Mycobacterium tuberculosis* acyltrehalose-containing glycolipids involved in phagosome remodeling. *PLoS Pathog.* **6**, e1001100 (2010).
38. Stermann, M., Sedlacek, L., Maass, S. & Bange, F. C. A promoter mutation causes differential nitrate reductase activity of *Mycobacterium tuberculosis* and *Mycobacterium bovis*. *J. Bacteriol.* **186**, 2856–2861. https://doi.org/10.1128/JB.186.9.2856-2861.2004 (2004).
39. Goh, K. S., Rastogi, N., Berchel, M., Huard, R. C. & Sola, C. Molecular evolutionary history of tubercle bacilli assessed by study of the polymorphic nucleotide within the nitrate reductase (narGHJI) operon promoter. *J. Clin. Microbiol.* **43**, 4010–4014. https://doi.org/10.1128/JCM.43.8.4010-4014.2005 (2005).
40. Honaker, R. W. *et al. Mycobacterium bovis* BCG vaccine strains lack narK2 and narX induction and exhibit altered phenotypes during dormancy. *Infect. Immun.* **76**, 2587–2593. https://doi.org/10.1128/IAI.01235-07 (2008).
41. Chauhan, S., Singh, A. & Tyagi, J. S. A single-nucleotide mutation in the-10 promoter region inactivates the narK2X promoter in *Mycobacterium bovis* and *Mycobacterium bovis* BCG and has an application in diagnosis. *FEMS Microbiol. Lett.* **303**, 190–196. https://doi.org/10.1111/j.1574-6968.2009.01876.x (2010).
42. Hutter, B. & Dick, T. Up-regulation of narX, encoding a putative 'fused nitrate reductase' in anaerobic dormant *Mycobacterium bovis* BCG. *FEMS Microbiol. Lett.* **178**, 63–69. https://doi.org/10.1016/S0378-1097(99)00348-1 (1999).
43. Ebrahimi-Rad, M. *et al.* Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerg. Infect. Dis.* **9**, 838–845. https://doi.org/10.3201/eid0907.020803 (2003).
44. Otchere, I. D. *et al.* Comparative genomics of *Mycobacterium africanum* Lineage 5 and Lineage 6 from Ghana suggests distinct ecological niches. *Sci. Rep.* **8**, 11269. https://doi.org/10.1038/s41598-018-29620-2 (2018).
45. Coscolla, M. *et al.* Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. *Microb. Genom.* https://doi.org/10.1099/mgen.0.000477 (2021).
46. Dos Vultos, T., Blázquez, J., Rauzier, J., Matic, I. & Gicquel, B. Identification of nudix hydrolase family members with an antimutator role in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. *J. Bacteriol.* **188**, 3159–3161. https://doi.org/10.1128/JB.188.8.3159-3161.2006 (2006).
47. van Soolingen, D., de Haas, P. E., Hermans, P. W. & van Embden, J. D. DNA fingerprinting of *Mycobacterium tuberculosis*. *Methods Enzymol.* **235**, 196–205 (1994).
48. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907–914. https://doi.org/10.1128/jcm.35.4.907-914.1997 (1997).
49. Prod'hom, G. *et al.* A reliable amplification technique for the characterization of genomic DNA sequences flanking insertion sequences. *FEMS Microbiol. Lett.* **158**, 75–81. https://doi.org/10.1016/S0378-1097(97)00503-X (1998).
50. Mendiola, M. V., Martin, C., Otal, I. & Gicquel, B. Analysis of the regions responsible for IS6110 RFLP in a single *Mycobacterium tuberculosis* strain. *Res. Microbiol.* **143**, 767–772. https://doi.org/10.1016/0923-2508(92)90104-V (1992).
51. Organización Panamericana de la salud. Manual para el diagnóstico bacteriológico de la tuberculosis. Normas y guía técnica. Parte II Cultivo. Washington, D.C. https://iris.paho.org/handle/10665.2/18616 (2008).

Acknowledgements

Authors would like to acknowledge the use of Servicio General de Apoyo a la Investigación-SAI, Universidad de Zaragoza (Servicio de Análisis Microbiológico), and Servicios Científico Técnicos, IACS (Servicio de Secuenciación y Genómica Funcional and Servicio de Biocomputación). We would like to thank the Aragon Health Department for their constant support in this work.

Author contributions

I.O., and S.S., participated similarly in the coordination of the work. J.C. performed the laboratory techniques to confirm results. They three wrote the manuscript. M.L.M. analysed the epidemiological data. J.V., J.S. and L.T. worked in the mycobacterial laboratories and are participants in the Working Group on Tuberculosis in Aragon

(EPIMOLA), they diagnosed, and completed minimal information on the TB cases and performed the genotyping. M.J.I., D.I., C.L. and C.M. work in the molecular epidemiology studies of TB in Aragon.

Funding
This work was supported by the Carlos III Health Institute in the context of two consecutive Grants FIS18/0336 and J.C. was awarded a scholarship by the Government of Aragon/European Social Fund, "Building Europe from Aragon".

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at www.nature.com/reprints.

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

Aragonese Working Group on Molecular Epidemiology of Tuberculosis (EPIMOLA)

María José Iglesias[5], Daniel Ibarz[5], Jesús Viñuelas[7], Luis Torres[8], Juan Sahagún[9], María Carmen Lafoz[5] & María Carmen Malo[10]

[7]Hospital Universitario Miguel Servet, Zaragoza, Spain. [8]Hospital General Universitario San Jorge, Huesca, Spain. [9]Hospital Universitario Lozano Blesa, Zaragoza, Spain. [10]Salud Pública, Gobierno de Aragón, Zaragoza, Spain.

Publicación 2

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 February 2022
doi: 10.3389/fmicb.2022.767912

In-depth Analysis of IS*6110* Genomic Variability in the *Mycobacterium tuberculosis* Complex

Jessica Comín[1,2]*, *Isabel Otal*[2,3,4†] and *Sofía Samper*[1,2,4†]

[1]Unidad de Investigación Traslacional, Hospital Universitario Miguel Servet, Instituto Aragonés de Ciencias de la Salud, Zaragoza, Spain, [2]Fundación IIS Aragón, Zaragoza, Spain, [3]Facultad de Medicina, Universidad de Zaragoza, Zaragoza, Spain, [4]CIBER de Enfermedades Respiratorias, Madrid, Spain

The insertion sequence (IS) *6110* is a repetitive mobile element specific for the *Mycobacterium tuberculosis* complex (MTBC) used for years to diagnose and genotype this pathogen. It contains the overlapping reading frames *orfA* and *orfB* that encode a transposase. Its genetic variability is difficult to study because multiple copies are present in the genome. IS*6110* is randomly located, nevertheless some preferential locations have been reported, which could be related to the behaviour of the strains. The aim of this work was to determine the intra- and inter-strain genetic conservation of this element in the MTBC. For this purpose, we analysed 158 sequences of IS*6110* copies from 55 strains. Eighty-four copies were from 17 strains for which we knew all the locations in their genome. In addition, we studied 74 IS*6110* copies in 38 different MTBC strains in which the location was characteristic of different families including Haarlem, LAM, S, and L6 strains. We observed mutation in 13.3% of the copies studied and we found 10 IS*6110* variants in 21 copies belonging to 16 strains. The high copy number strains showed 6.2% of their IS*6110* copies mutated, in contrast with the 31.1% in the low-copy-number strains. The apparently more ancient copy localised in the DR region was that with more variant copies, probably because this was the most studied location. Notably, all Haarlem and X family strains studied have an IS*6110* in *Rv0403c*, suggesting a common origin for both families. Nevertheless, we detected a variant specific for the X family that would have occurred in this location after the phylogenetic separation. This variant does not prevent transposition although it may occur at a lower frequency, as X strains remain with low copy number (LCN) of IS*6110*.

Keywords: tuberculosis, IS*6110*, *Mycobacterium tuberculosis* complex, IS*6110* genomic variability, tuberculosis evolution

OPEN ACCESS

Edited by:
Daniel Yero,
Universidad Autónoma de Barcelona, Spain

Reviewed by:
Brosch Roland,
Université Louis-Pasteur, France
Maria Laura Boschiroli,
Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail (ANSES), France

***Correspondence:**
Jessica Comín
jcomin.iacs@aragon.es

*These authors have contributed equally to this work and share last authorship

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

Received: *31 August 2021*
Accepted: *04 February 2022*
Published: *24 February 2022*

Citation:
Comín J, Otal I and Samper S (2022) In-depth Analysis of IS6110 Genomic Variability in the Mycobacterium tuberculosis Complex. Front. Microbiol. 13:767912. doi: 10.3389/fmicb.2022.767912

INTRODUCTION

The insertion sequence (IS) *6110* has been described as a repetitive mobile element specific for the *Mycobacterium tuberculosis* complex (MTBC; Thierry et al., 1990). It has been used for years to diagnose and to characterise at the molecular level *M. tuberculosis* strains by the IS*6110*-restriction fragment length polymorphism (RFLP) technique (Thierry et al., 1990;

Brisson-Noel et al., 1991; Otal et al., 1991). It is 1,355 bp and is flanked by inverted repeats (IR) of 28 bp (Thierry et al., 1990). When transposition occurs, it creates a duplication of 2–4 bp in the insertion site (Mendiola et al., 1992; Dale, 1995). IS*6110* contains two overlapping reading frames called *orfA* and *orfB*. While the product orfAB is a transposase, the individual products orfA and orfB inhibit transposition, so the proportion of each product is important for the autoregulation of transposition (Sekine et al., 1997). Transcription of this element is regulated by a-1 frameshift due to a ribosomal slippery sequence and by a pseudoknot in the mRNA; these processes lead to a transposition based on a copy-out-paste-in mechanism (Gonzalo-Asensio et al., 2018).

The global *M. tuberculosis* population can be divided into several phylogeographic lineages (Ls) and families. The genetic diversity found in circulating strains plays an important role in disease outcome. The number of IS*6110* copies in the genome is variable and seems to be related to the evolution of the strains. Modern Ls have more IS*6110* copies than ancient Ls, and they are considered to be better adapted to high-density populations (Gonzalo-Asensio et al., 2018). It seems that if the IS*6110* is inserted in a transcriptionally active region of the genome, the number of copies increases, so the variable number of copies present in different strains could be due to the genomic region in which they are inserted, although a large number of IS*6110* copies may be deleterious for the strain because they can remove parts of the genome (Wall et al., 1999).

While IS*6110* supposedly transposes randomly, there are some preferential location sites such as the CRISPR region, also known as the DR region in *M. tuberculosis*; it was probably the first IS copy because it is present in almost all MTBC isolates (Hermans et al., 1991). Other hot spots described are: the *plcD* gene (Vera-Cabrera et al., 2001), members of the PPE family genes (Sampson et al., 1999; Beggs et al., 2000), the intergenic region *dnaA*:*dnaN* (Supply et al., 2006), and other ISs (Fang and Forbes, 1997; Fang et al., 1999). In addition, characteristic locations of IS*6110* in different Ls have been observed (Roychowdhury et al., 2015). The function of IS*6110* remains unknown although some general effects have been described. The first is the disruption of the gene in which it has been inserted, with a possible deleterious effect (Sampson et al., 1999; Beggs et al., 2000; Warren et al., 2000; Yesilkaya et al., 2005). The second is the recombination and the deletion or inversion of the DNA in between, a phenomenon that can occur when there are two IS relatively close to each other (Sampson et al., 2003). A third effect is that IS*6110* can act as a promoter of the downstream gene (Beggs et al., 2000; Safi et al., 2004; Soto et al., 2004).

For the MTBC, the rate of point mutations has been determined to be 1×10^{-9} events per site per generation. However, it has been demonstrated that the rate of point mutations for the IS*6110* is about 7.9×10^{-5}, higher compared with the rest of the genome, indicating that IS*6110* is under selective selection (Tanaka, 2004). There are few studies focused on the sequence genetic variability of this element. Dale et al. (1997) amplified and sequenced IS*6110* of four *M. tuberculosis* strains, each carrying one copy, and one of the copies carried by the Beijing W strain, which has 17 copies. They compared the sequences with the Bacillus Calmette and Guérin (BCG) copy published and found they are identical. Nevertheless, they only checked one copy of each strain. Conversely, Thabet et al. (2015) found several different haplotypes when they randomly cloned and sequenced only one of the IS*6110* copies of some characteristic strains from their region.

In this work, we aimed to determine whether IS*6110* is a conserved gene, as Dale et al. (1997) suggested or, on the contrary, it is relatively common to find this IS mutated, as Thabet et al. (2015) suggested. In addition to the clinical relevance, since it is widely used for TB diagnosis, we thought mutations of the IS*6110* could affect its transposition ability. Therefore, we planned to find out the genetic variability of the sequence. For this purpose, we applied two different approaches. The first was the study of all the IS*6110* copies present in the same strain to determine whether they had an identical sequence. For the second approach, a sample collection was used to study IS*6110* copies in different MTBC families for which the location had been previously published (Reyes et al., 2012; Comín et al., 2021) to elucidate how conserved IS*6110* is among families for the same location.

MATERIALS AND METHODS

Collection of Samples

Study of IS*6110* in Available Completed Genomes

We selected the available MTBC complete genomes from the NBCI database. First, we checked the entire sequence of the 16 IS*6110* copies of H37Rv (NC_000962.3) using the Tuberculist BLAST tool.[1] Second, we looked for the IS*6110* copies of CDC1551 (AE000516.2), *Mycobacterium africanum* (CP010334.1 and FR878060.1), *Mycobacterium bovis* (LT708304.1), and *M. bovis* BCG (CP033311.1, CP014566.1, and AM408590.1) loaded in the NCBI database. We checked the insertion points using Tuberculist. All the points detailed in the results refer to the H37Rv genome, unless otherwise indicated (**Table 1**).

Strains Used for Sequencing IS*6110*

All the DNAs used in this work were extracted from clinical *M. tuberculosis* isolates previously genotyped by IS*6110*-RFLP and spoligotyping in our laboratory, and stored at −20°C. Strains previously characterised and with all their IS*6110* copies located were selected to amplify all the copies within the genome. These strains were: *M. tuberculosis Zaragoza* (*MtZ*, L4; Isabel Millan-Lou et al., 2013), GC1237 (L2; Alonso et al., 2011), and seven low copy number (LCN) of IS*6110* strains belonging to different lineages. These LCN strains were HMS 2407 and HMS 2382 (L6), HCU 3445 (L4), HMS 2484, HMS 2445, HCU 3717 (Comín et al., 2021**;** L4 strains belonging to X family), and *M. bovis* B (hypervirulent strain; Sagasti et al., 2016; **Table 1**). For the extension of the study (**Table 2**), we used

[1]http://genolist.pasteur.fr/TubercuList/

TABLE 1 | Summary of IS*6110* copies analysed in the strains for which the location of all the IS copies within their genomes is known, detailing the IS copies mutated by comparison with the reference IS in *Mycobacterium tuberculosis* H37Rv.

Strain	Lineage (Family)	CN strain	IS analysed (*N* = 84)	Mutated copies location	Mutation detected	Non-mutated copies location	Reference
H37Rv	L4.9	HCN	16	-	-	all wt	NC_000962
CDC1551	L4.1.1.3	LCN	4	*Rv0403c,cut,ppe46*	*orfB Gly215Ser**	*DR region*	AE000516.2
M. africanum 25	L6	LCN	5	*Rv3750c*	*orfB gap in 110n.*	all wt but one	*CP010334.1*
M. africanum GM041182	L6	HCN	7	DR region	orfA syn 33 aa	all wt but one	FR878060.1
M. bovis AF2122/97	L Animal	LCN	1	-	-	all wt	LT708304.1
M. bovis BCG_ S49	L Animal	LCN	1	-	-	all wt	CP033311.1
M. bovis BCC_ Tokyo	L Animal	LCN	2	-	-	all wt	CP014566.1
M. bovis BCG_ Pasteur	L Animal	LCN	1	-	-	all wt	AM408590.1
M. tuberculosis MtZ	L4.8	HCN	12	-	-	all wt	Isabel Millan-Lou et al., 2013
M. tuberculosis GC1237	L2	HCN	18	-	-	all wt	Alonso et al., 2011
M. bovis B	L Animal	LCN	2	-	-	DR region, Rv0756c:phoP	Sagasti et al., 2016
HMS 2382	L6	LCN	3	DR region	orfB Asp2Gly	moaX,Rv0963c	Comín et al., 2021
HMS 2407	L6	LCN	3	-	-	lipX:mshB,moaX, DR region	Comín et al., 2021
HCU 3445	L4	LCN	1	-	-	DR region	
HMS 2485	L4.1.1.3	LCN	2	*Rv0403c*	*orfB Gly215Ser**	DR region	
HCU 3717*	L4.1.1.3	LCN	4	DR region (Otal et al., 1991), *Rv0403c,* and *MT1802:cut1*	*orfB Gly215Ser**		Comín et al., 2021
HMS 2445	L4.1.1.3	LCN	2	*Rv0403c*	*orfB Gly215Ser**	DR region	

*Even though HCU3717 strain only has four out of five IS copies analysed, it has been included in this section because the fifth copy is located in ppe46 gene (not successfully amplified). CN, Copy number; HCN, high copy number; and LCN, low copy number. *Mutation associated to X family.*

32 L4 strains distributed in different families: 10 LAM strains (HMS 18005, HSJ 241, HMS 18045, HMS 18017, HMS 18025, HMS 18010, HMS 18047, HMS 18048, HMS 18018, and ara217, a LAM9 strain studied for producing an outbreak in our region; Comin et al., 2020), 12 Haarlem strains (HSJ 238, HMS 18009, HMS 18021, HMS 18037, HCU 3729, HMS 18031, HMS 18022, HMS 18007, HMS 18046, HSJ 234, HMS 18001, and HMS 18002), one S strain (HMS 18019), and seven T strains (HCU 3718, HMS 18041, HMS 18035, HMS 18042, HMS 18014, HMS 18040, and HMS 18044). In addition, we also included one L3 strain belonging to the CAS1_DELHI family (HMS 18038) and seven *M. africanum* L6 strains (HCU 2828, HCU 3775, HMS 14017, HMS 2000, HMS 1942, HSJ 66, and HMS 1693; **Table 2**).

Analysis of the IS*6110* DNA Sequences

Amplification and Sequencing of the Region Including the IS*6110*

For the PCR, MyTaq DNA polymerase (Lot No: MT-7171128, Bioline) and 5x MyTaq Reaction Buffer (Lot No: MTB-T17212A, Bioline) were used. We performed an initial denaturation at 95°C for 1 min. Then we performed 35 cycles: denaturation at 95°C for 15 s, annealing at the primer temperature for 15 s, and extension at 72°C for 30 s. The specific primers used for each point of insertion are shown in **Supplementary Tables S1–S4**. We used a commercial kit (GFX PCR DNA and Gel Band Purification Kit, GE Healthcare, Lot: 16834032) to clean the samples before sequencing. We used the same kit when we needed to extract the DNA from a gel band. The samples were sequenced using capillary electrophoresis with the specific primers flanking the different IS copies and the internal IS primers: IS7 (TTCGGACCACCAGCACCTAACC), IS9 (GCTTTGCCGCGGGTGG), and IScom (ATGTCAGGT GGTTCATCGAGGAGG) in a 3500XL Genetic Analyser (Applied Biosystem).

Study of the Sequences

The sequences obtained were studied using the BioEdit program (Informer Technologies, Inc.), a biological sequence alignment editor. We also used Tuberculist[1] and Bovilist,[2] where the annotation of *M. tuberculosis* H37Rv and *M. bovis* AF2122/97 can be found. We also used the BLAST tool from the NCBI[3] to study CDC1551, two *M. africanum* strains, and three *M. bovis* BCG strains. Integrative Genomics Viewer (IGV; Robinson et al., 2011) software,

[2]http://genolist.pasteur.fr/BoviList/
[3]https://www.ncbi.nlm.nih.gov/

TABLE 2 | Summary of the IS*6110* sequences localised in preferential sites published for determined *Mycobacterium tuberculosis* families (Reyes et al., 2012; Comín et al., 2021).

Strain	Lineage	Family by SIT	CN STRAIN	Number of IS*6110* */ studied (*N* = 74)	Mutated copies location	Mutation detected	Non-mutated copies location
HSJ 238	L4.1.2.1	Haarlem3	HCN	11/3	-	-	*Rv1754c,Rv0403c,Rv0963c*
HMS 18009	L4.1.2.1	Haarlem3	HCN	10/2	*DR region*	*orfB Asp2Gly*	*Rv0403c*
HMS18021	L4.1.2.1	Haarlem3	HCN	12/2	-	-	*Rv0403c,Rv0963c*
HMS 18037	L4.1.2.1	Haarlem3	HCN	11/4	*Rv0963c*	*IR*	*Rv1754c,Rv2336,* and *Rv0403c*
HCU 3729	L4.1.2.1	Haarlem3	HCN	7/5	-	-	*Rv1754c,Rv2336,Rv0403c,Rv0963c,* and *DR region*
HMS 18031	L4.1.2.1	Haarlem3	HCN	8/2	*Rv0963c*	*orfB syn aa174*	*Rv0403c*
HMS 18022	L4.1.2.1	Haarlem1	LCN	6/4	-	-	*Rv1754c,Rv2336,Rv0403c,* and *Rv0963c*
HMS 18007	L4.1.2.1	Haarlem1	HCN	10/1	-	-	*DR region*
HMS 18046	L4.1.2.1	Haarlem1	HCN	9/4	*DR region*	*orfB Asp2Gly*	*Rv0963c,Rv1754c,* and *Rv0403c*
HSJ 234	L4.1.2.1	Haarlem3	HCN	11/3	*Rv0963c*	*orfB syn103*	*Rv2336,Rv0403c*
HMS 18001	L4.1.2.1	Haarlem1	HCN	8/4	-	-	*Rv1754c,Rv0403c,Rv0963c,* and *DR region*
HMS 18002	L4.1.2.1	Haarlem1	HCN	10/4	-	-	*Rv1754c,Rv2336,Rv0963c,* and *DR region*
HMS 18005	L4.3	LAM3	LCN	5/1	-	-	*Rv3113*
HSJ 241	L4.3	LAM3	HCN	15/2	-	-	*Rv3113,DR region*
HMS 18045	L4.3	LAM3	HCN	15/2	-	-	*Rv3113,DR region*
HMS 18017	L4.3	LAM12	HCN	14/2	-	-	*lpqQ:Rv0836c,DR region*
HMS 18025	L4.3	LAM9	HCN	11/2	*lpqQ:Rv0836c*	*orfB Gly123Arg*	*DR region*
HMS 18010	L4.3	LAM9	HCN	14/2	-	-	*Rv1754c,DR region*
HMS18047	L4.3	LAM4	HCN	14/1	-	-	*DR region*
HMS 18048	L4.3	LAM9	HCN	10/3	-	-	*Rv1754c,lpqQ:Rv0836c,* and *DR region*
HMS 18018	L4.3	LAM3	HCN	12/3	-	-	*Rv1754c,MT3426:MT3427,* and *lpqQ:Rv0836c*
Ara217**	L4.3	LAM9	HCN	13/2	-	-	*MT3426:MT3430,Rv1754c:cut1*
HMS 18019	L4.4.1.1	S	HCN	11/1	-	-	*pks9*
HCU 3718	L4	T1	HCN	9/1	-	-	*DR region*
HMS 18035	L4	T2	LCN	5/1	-	-	*DR region*
HMS 18040	L4	T5	HCN	8/1	-	-	*DR region*
HMS 18042	L4	T2	HCN	7/1	-	-	*DR region*
HMS 18041	L4	T1	HCN	9/1	-	-	*DR region*
HMS 18044	L4	T5_MAD2	LCN	6/1	-	-	*DR region*
HMS 18014	L4	T5_MAD2	LCN	6/1	*DR region*	*orfA Thr67Iso*	-
HMS 18038	L3	CAS1_DELHI	HCN	16/1	-	-	*DR region*
HCU 2828	L6	*AFRI_1*	HCN	8/1	-	-	*lipX:mshB*
HCU 3775	L6	*AFRI_1*	LCN	3/1	-	-	*lipX:mshB*
HMS 1693	L6	*AFRI_1*	LCN	3/1	*lipX:mshB*	*IR*	-
HMS 14017	L6	*AFRI_1*	LCN	3/1	*lipX:mshB*	*IR*	-
HMS 1942	L6	*AFRI_1*	LCN	unk/1	-	-	*moaX*
HMS 2000	L6	*AFRI_1*	LCN	5/1	-	-	*moaX*
HSJ 66	L6	*AFRI_1*	LCN	3/1	-	-	*moaX*

*CN, Copy number; HCN, high copy number; LCN, low copy number; and SIT, Spoligo International Type. *Total number of IS6110 based in the IS6110 RFLP pattern.*
***Outbreak strain (Comín et al., 2020).*

which allows the study of the whole genome in an interactive way, was also used.

RESULTS

We aimed to determine the genetic variability of the IS*6110* sequence. With this purpose, a large number of copies were analysed. The first approach of the study focused on the comparison of all the IS copies present in a strain to determine whether different copies coexist in the same genome. Initially, we analysed the strains whose genomes were available in the databases. Furthermore, we studied several strains whose IS*6110* locations were known as result of our previous studies. For the second approach, we studied IS*6110* in specific locations defined for different families, including Haarlem, LAM, S, and L6 strains (Reyes et al., 2012; Comín et al., 2021) to determine its inter- and intra-strain variability using a sample collection. In addition, CAS_DELHI and T strains were included to study the IS copy located in the DR region. As a result, 158 IS*6110*

sequences from 55 different strains were analysed (**Figure 1**). Twenty-one of the copies, belonging to 16 strains, had a mutation (13.3%). Among the high copy number strains, seven out of the 113 copies studied (6.2%) were mutated while 14 out of the 45 copies (31.1%) showed a mutation in the LCN strains.

Review of the Different IS Copies in the NCBI and Tuberculist Databases

To investigate the IS*6110* sequence variability, we checked MTBC strains whose genomes are available in the NCBI and Tuberculist databases: *M. tuberculosis* H37Rv and CDC1551, *M. africanum* 25 and GM041182, *M. bovis* AF2122/97, and three different BCG strains. The results are summarised in **Table 1**.

Initially, the analysis of the 16 IS*6110* copies registered in Tuberculist belonging to the H37Rv strain showed that all the sequences are identical to each other. Therefore, we chose the first copy in its genome, from points 889,020 to 890,375 (*Rv0795* and *Rv0796*), as the reference wild-type IS for the successive comparisons.

Searching the IS*6110* copies in the different genomes available in the NCBI database showed several IS copy variants. The CDC1551 strain has four IS*6110* copies, located in *Rv403c* [point 483,296, reverse direction (−)], *cut1* (1,989,058, −), DR region (3,120,523:3121879, −), and *ppe46* genes [3,377,327, forward direction (+)]. Only the copy in the DR region is wild type, while the other three IS*6110* copies share a mutation in the first base of codon 215, within *orfB* of the IS sequence (G/A, Gly/Ser).

We performed the same procedure for *M. africanum* 25. We found five copies of IS*6110* in its genome, inserted in *lipX:mshB* (1,300,194, −), *Rv1765c* (1,998,416, +), DR region (3,120,523:3,121,879, −), *moaX* (3,709,622, −), and *Rv3750c* genes (4,198,431, +). Only the copy in *Rv3750c* showed a gap in nucleotide 110 of *orfB*. *Mycobacterium africanum* GM041182 has seven IS*6110* copies. Five of them are at the same location and direction as those in strain 25. Nevertheless, the one located in the DR region has a synonymous mutation in the last base of codon 33 (A/G), within *orfA* of the IS sequence. The previous gap found in strain 25 was not detected. The sixth and seventh copies are inserted in *pks8* (1,882,012, +) and *Rv3128c:Rv3129* genes (3,494,393, −), both presenting the wild-type sequence.

The single IS*6110* copy of *M. bovis* AF2122/97, inserted in the DR region (3,120,523–3,121,879, −), is wild type. We chose three BCG strains, CP033311.1, CP014566.1, and AM408590.1. All of them have the copy of the DR region inserted at the same point and direction as *M. bovis* AF2122/97. A second copy of CP014566.1 is upstream the *phoP* gene (851,590, −). All the copies are wild type.

Analysis of the IS*6110* in Clinical Strains With All Their Copies in Known Locations

Using the information obtained in previous studies by our group, we were able to successfully amplify and sequence 47 IS*6110* copies from the genomes of three outbreak strains previously characterised (MtZ, GC1237, and *M. bovis* B) and six LCN strains (HMS 2382, HMS 2407, HCU 3445, HMS 2485, HMS 2445, and HMS 3717). The specific characteristics of these strains can be found in **Table 1**. The 12 IS*6110* copies of the *MtZ* strain (locations can be found in **Supplementary Table S1**), the 18 copies of the GC1237 strain (**Supplementary Table S2**), and the two copies of *M. bovis* B present the wild-type sequence. Among the analysed sequences of the LCN strains, we found two different mutations. *Mycobacterium africanum* HMS 2382 has a change in the overlapping region of *orfA* and *orfB* in its IS*6110* copy within the DR region (−). This mutation (A/G) affects the second amino acid (Asp/Gly) after the ribosomal slippery. We also found a mutation in the IS*6110* copies inserted at *Rv0403c* (−) in HMS 2485, HMS 2445, and HCU 3717 strains, all belonging to X family. This Gly/Ser (G/A) mutation coincides with that found in CDC1551 IS copies (which also belongs to X family). This mutation was also found in three other IS copies of HCU 3717, located in *MT1802:cut1* and the DR region (two copies). We were not successful in the amplification of its fifth copy, located in *ppe46*, but if it was also mutated, HCU 3717 would be the unique strain with all its IS copies mutated. All the other copies from the LCN strains analysed are wild type. The exact point of the mutations found is detailed in the sequence of IS*6110* in **Figure 2**.

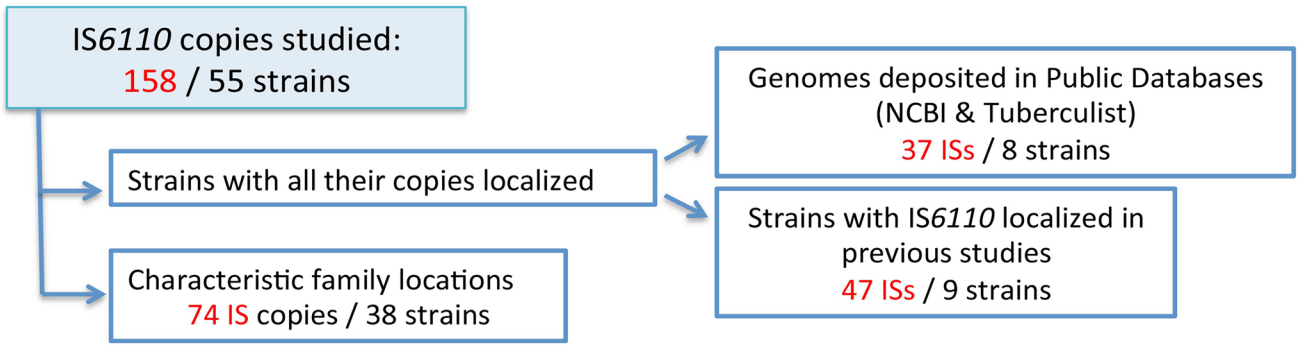


FIGURE 1 | Summary of the different IS*6110* copies studied in this work.

Analysis of the IS*6110* Sequences in Characteristic Locations in MTBC Families

To extend the study to other MTBC families and to more strains, we relied on the specific IS*6110* locations for Haarlem, LAM, S, and L6 strains detailed in other publications (Reyes et al., 2012; Comín et al., 2021). For the hot spot DR location, we included additional T and CAS_DELHI isolates. The results are shown in **Table 2**.

According to Reyes et al. (2012), 96% of Haarlem strains have an IS inserted in *Rv0403c* (483,296, −), *Rv2336* (2,610,861, +), and *Rv1754c* genes (1,986,622, +), and 89% have a copy in *Rv0963c* (1,075,948, −). We succeeded in the amplification of 10 IS*6110* copies in *Rv0403c* in different strains, five copies in *Rv2336* and seven copies in *Rv1754c*. All of them are wild type. However, we found two different synonymous mutations for the IS copy in *Rv0963c* in two of the 10 strains analysed. HMS 18031 has a mutation in the last base of codon 174 and HSJ 234 has a mutation in the last base of codon 103, both in *orfB*. The rest of the copies are identical to the reference IS sequence.

Regarding the LAM family, it has been reported (Reyes et al., 2012) that more than 96% strains have an IS inserted in *Rv1754c* (1,986,623, −), more than 95% have it inserted in *lpqQ:Rv0836c* (932,202, −), and more than 90% have it inserted in *Rv3113* (3,480,371, +). Based on these findings, we analysed four sequences from the IS inserted in *Rv1754c*, four sequences from the IS inserted in *lpqQ:Rv0836c*, and three sequences from the IS inserted in *Rv3113*. All the sequences are wild type, except the one in *lpqQ:Rv0836c* of HMS 18025, which has a mutation in the first base of codon 123 after the ribosomal slippery sequence (*orfB*, Gly/Arg). Furthermore, we studied one more location in *MT3426:MT3430* in the context of a LAM strain outbreak investigation (ara217; Comin et al., 2020), successfully amplified for this strain. We also studied this location for the other LAM strains but it was only successfully amplified for one of them. Both present the wild-type genotype.

According to Reyes et al. (2012), more than 85% of S strains have an IS*6110* copy in the *pks9* gene (1,889,066, −). We studied the IS*6110* copy inserted in this location in one S family strain (HMS 18019). This copy is identical to the reference IS sequence.

A recent work (Comín et al., 2021) determined that 100% of the L6 strains have a copy of IS*6110* in the *moaX* gene (3,709,622, −). We analysed five copies for this position, all presenting the wild-type genotype. In the same work, the authors reported that it is common to find a copy between the *lipX* and *mshB* genes (1,300,194, −) for L6. We successfully amplified four copies for this location, each showing the wild-type *orfA* and *orfB* genes.

Almost all MTBC strains have an IS*6110* inserted in the DR region. We amplified the IS of this region for 30 strains. We found three different mutations in five strains (**Figure 2**). HMS 2382 (L6), HMS 18009, and HMS 18046 (both belonging to the Haarlem family) share a mutation in the second base

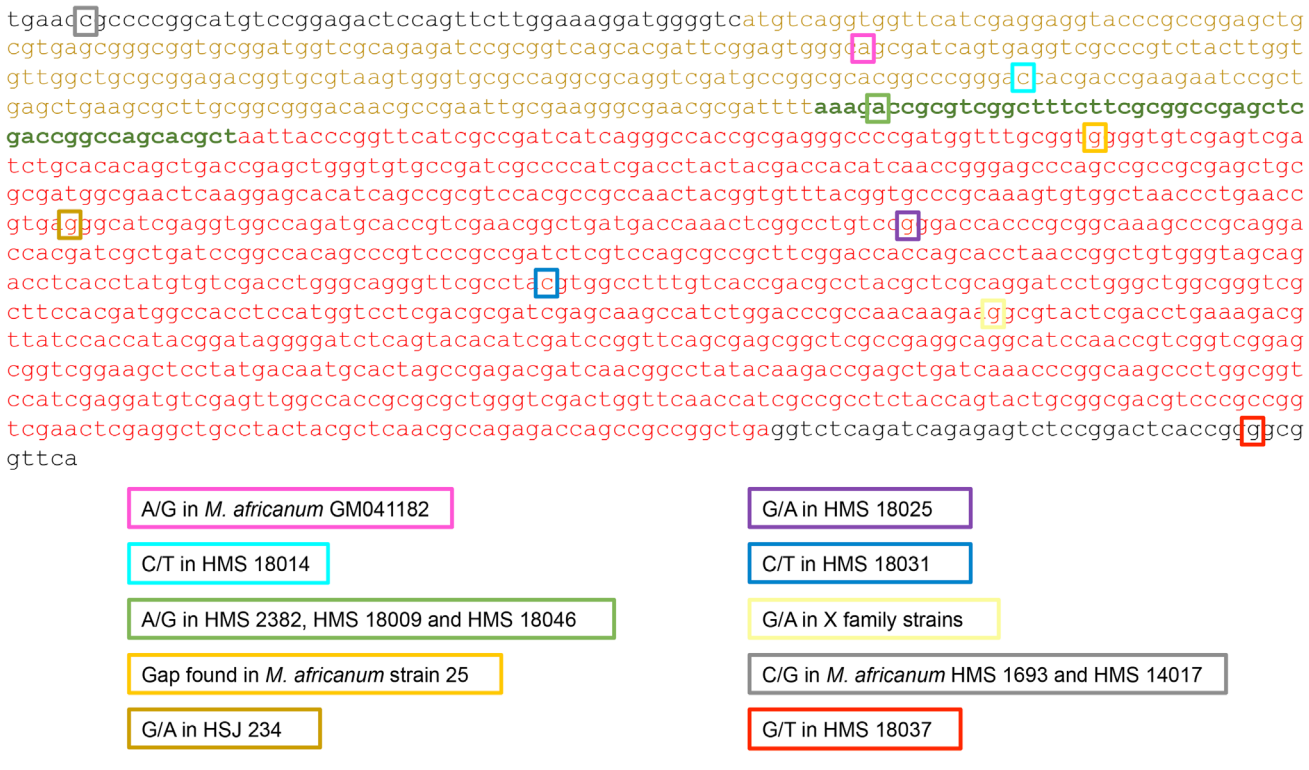


FIGURE 2 | Sequence of IS*6110*, highlighting in different coloured squares the points, the nucleotide changes, and the different strains where the 10 mutations were found. *orfA* is written in yellow, *orfB* is in red, and the overlapping region is in green. The black letters at the ends represent the inverted repeats.

of codon 2 after the ribosomal slippery (overlapping region, Asp2Gly). HMS 18014 (T5_MAD2 family) has a mutation in the second base of codon 67 (*orfA*, Thr67Iso). As we have commented above, HCU 3717 (X family) has its two IS copies inserted in the DR region mutated (*orfB*, Gly215Ser). The rest of the strains studied have the wild-type IS.

Among all the copies studied, we found two different mutations for three strains in the IR of the IS*6110* (**Figure 2**): a SNP in base 6 of the IR flanking *orfA*, shared by HMS 1693 and HMS 14017 strains (L6) in the copy inserted in *lipX:mshB*; and a different SNP in base 9 of the IR flanking the *orfB*, in the copy inserted in *Rv0963c* of the HMS 18037 Haarlem strain.

Study of IS*6110* Copies in *ppe* Genes

PPE regions are hot spots for the IS*6110* insertion (Sampson et al., 1999; Beggs et al., 2000), generating a high variability in the genome. In addition, these genes are difficult to amplify because the high number of repetitive regions and a high GC content. Nevertheless, we were successful in the amplification of two IS copies located in *ppe38* and *ppe71* for the MtZ strain (−) and one copy located between *ppe38/ppe71* (+) in Beijing GC1237 (**Figure 3**). This IS in the Beijing GC1237 strain produced a deletion of *esx* genes and the truncation of both *ppe* genes. We were also successful in the amplification of an IS*6110* copy located downstream of the *ppe46* gene, between the *pe27a* and *esxR* genes. In addition, we amplified another IS*6110* inserted within *ppe34*. We observed a different organisation of this gene compared with the H37Rv strain, with short repeated fragments of *ppe34* after the IS, a feature that may be typical of the Beijing family (H37Rv does not have these repetitions), as using the BLAST tool of the NCBI; Beijing genomes were the full matches. As we stated before, these copies do not have mutations.

DISCUSSION

This work deepens the knowledge of the IS*6110* sequence in the MTBC. IS*6110* has been used for a long time for diagnosis and genotyping, but very little is known about its genetic variability, mainly due to the difficulties in determining its complete sequence for each copy location in the genome. The limited data that have been published regarding this issue have shown dissimilar results (Dale et al., 1997; Thabet et al., 2015).

We have analysed the sequences of both *orfA* and *orfB* genes that constitute the transposase in several strains for which we knew its exact location in the genome using specific primers for the amplification of each copy. So far, the only data reported on the sequence of IS*6110* were obtained using internal primers that interchangeably amplify all the IS copies of the genome (Thabet et al., 2015), or the number of analysed copies was really low (Dale et al., 1997). In addition, as only one copy of each strain was analysed, the intra-strain variability was not studied. Therefore, the novelty of this work, as far as we know, is that for the first time the IS*6110* sequence corresponding to each location has been identified and sequenced, and likewise, the sequences of the different copies in the same strain or among families have been compared. Furthermore, IS*6110* cannot be studied directly using the most frequent whole genome sequencing (WGS) technologies, as is the case of Illumina, due to repetitive sequences are not correctly aligned; thus, this study complements this handicap.

We studied 158 IS*6110* sequences from 55 different strains belonging to L2 (Beijing), L3 (CAS_DELHI), L4 (Haarlem, LAM, T, S, and X), L6, and the animal branch, including strains with a high and a LCN of IS*6110*. There were variations of the wild-type sequence in 21 (13.3%) of the copies belonging to 16 strains (29%). We observed a higher percentage of mutated copies among the LCN strains (31.1 vs. 6.2% in high copy number), which could be related with a less transposition ability. To support this idea, a higher number of copies with a widely strain spectrum should be studied.

We found 10 different mutations (**Figure 2**). Five of them are located in the *orfB* of the transposase vs. two in the *orfA*, findings that seem to indicate that *orfB* is more susceptible to mutation. A possible explanation could be that the *orfB* is three times longer than *orfA*. One mutation is in the overlapping region affecting both ORFs. In addition, two other mutations are in the IR, which we do not expect to be involved in the

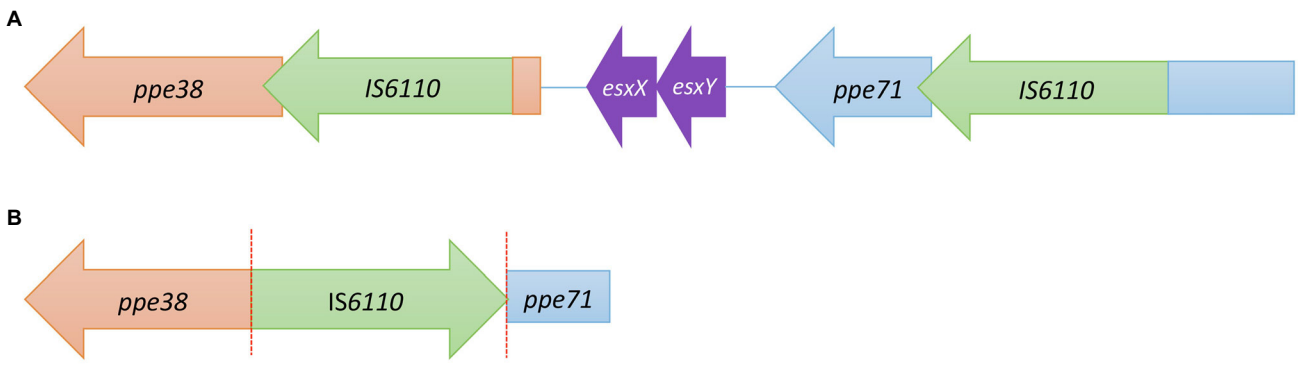


FIGURE 3 | (A) Configuration of the *ppe38* locus in the MtZ strain. **(B)** Configuration of the *ppe38* locus in the Beijing GC1237 strain. The red lines indicate a truncation of the gene.

transposase function as they are outside the coding region. Thabet et al. (2015) also found more polymorphic sites for *orfB* than for *orfA*. They suggested that non-synonymous mutations tend to be eliminated because the ancestral IS*6110* copy inherited by the MTBC would be functionally optimal. However, we found five out of eight mutations in the coding *orfA* and *orfB* that are non-synonymous. The fact that some of them are conserved in different strains supports the idea that those non-synonymous mutations are neutral in these cases, as Thabet et al. (2015) suggested. They described different haplotypes for *orfA* and *orfB*, reporting the wild type as the most frequent, similarly to our findings. On the other hand, our results differed from Thabet et al. (2015) in other aspects, including the description of some haplotypes with 1–3 SNPs. This caught our attention because all the variant copies we found have just one SNP. The high variability found by these authors could be partially caused by the different procedures applied to obtain the IS*6110* DNA sequence. We used specific primers for each location, what allowed us to sequence directly, for each copy, the DNA obtained from PCR. In contrast, other authors used IS internal primers to amplify, clone and sequence one of the copies of each strain without knowing what copy had been amplified or if the sequence obtained was the sum of several. One of the mutations they found (haplotype 20 of *orfB*) is Gly215Ser (*orfB*), the same we found in the X family. Since their collection did not include any X strains, we conclude that this mutation is not restricted to the X family. None of the other mutations obtained matched their results. Our results completely differed from Dale et al. (1997), who did not find any mutation when comparing five IS copies of different *M. tuberculosis* strains, so they affirmed that the IS has been totally conserved among the MTBC. Surely, the low number of copies they studied caused these differences.

As far as we know, this is the first time that all the IS copies of a strain have been studied. In some cases, we observed that it exists variability among the IS*6110* copies of the same strain, finding wild-type and mutated copies at the same time. In addition, we did not find copies with different mutations in the same strain.

Our expertise in genotyping *M. tuberculosis* by IS*6110*-RFLP supports the fact, previously described (Gonzalo-Asensio et al., 2018), that modern lineages (L2, L3, and L4) have accumulated more copies of IS*6110* in their genomes than ancient lineages (L1, L5, and L6). Modern lineages are better adapted to high-density populations, so it is possible that the number of ISs is related to this success (Gonzalo-Asensio et al., 2018). This eventuality agrees with the obtained results because the two most successful strains in our population, used in this work, are high copy number strains (MtZ and GC1237) and, curiously, all their copies (12 and 18, respectively) are wild type. Remarkably, all the strains of the X family (L4) studied, such as CDC1551; share a characteristic variant in at least one of their IS*6110* copies. According to these results, it is possible that this mutation could affect the transposition ability of IS*6110*, reducing its transposition frequency. Thereby, there would be a greater probability that the wild-type copies transpose, leading to a higher copy number of IS*6110* in the genome. This phenomenon could partially explain why the X family strains, belonging to L4, are LCN strains despite being a successful clinical family. Nevertheless, the mutation found in the X family strains did not prevent the transposition, as there are more copies with the same SNP in the same strain. Dale et al. (1997) suggested that the differences observed in the copy number are not related to structural changes in the transposase, but rather to the genomic region in which it is inserted. In view of our results, focusing especially on the X strains, we think that both facts are not mutually exclusive and more than one factor must be affecting its transposition.

It is notable that the Haarlem and X family strains studied have an IS*6110* inserted in *Rv0403c* at the same point and direction (483,296, −), in addition to an identical 3 bp created duplication. This finding suggests a common origin for both families, supporting all the phylogenetic trees in which the X and Haarlem families are closely related (Stucki et al., 2016). All the X strains studied in this work have a mutated copy located in *Rv0403c*, unlike the Haarlem strains, which lack a mutation. The evolutionary process that is inferred by these results is that the common ancestor of these two families had the IS*6110* copies located in the DR region and in the *Rv0403c* gene. Subsequently, the mutation in the copy located in *Rv0403c* of the X strains occurred after the phylogenetic separation between the Haarlem and X families, and this mutated copy was the one that transposed and led to the rest of the copies in this family, because all IS*6110* copies have the same mutation, except the one located in the DR region. The exception is HCU 3717, which has its two DR copies also mutated, suggesting a different evolutionary process.

Researchers demonstrated that the synthesis of the transposase is regulated post-transcriptionally by the ribosome (Gonzalo-Asensio et al., 2018). *orfA* finishes on Leu92 (UUA), but the ribosome can make a −1 frameshift and starts *orfB* with Lys1 (AAA). Considering this, we conclude that the mutation in the IS sequences of the X family strains has changed glycine 215 for serine (*orfB*). Glycine is the smallest amino acid and the only one that is not chiral. Furthermore, it could provide more flexibility to the protein region. On the contrary, serine is a polar amino acid, uncharged, and is bigger than glycine, a factor that diminishes the flexibility of the protein chain. We believe it is possible that this amino acid change may affect the protein structure but, according to our results, it does not prevent the transposition ability.

Regarding *M. africanum*, in both NCBI strains one of its IS*6110* copies is mutated. While the mutation found in the GM041182 strain is synonymous, strain 25 has a gap in the sequence, which completely alters the reading frame and probably disrupts the transposase. Among the LCN strains we sequenced, isolated in our region, *M. africanum* HMS 2382 has a non-synonymous mutation in one of its three IS*6110* copies. It changes aspartic acid for glycine, an important change because aspartic acid has negative charge and is larger than glycine. Although the Haarlem family is a completely different lineage (L4) than *M. africanum* (L6), we found that two Haarlem strains (HMS 18009 and HMS 18046) share the same mutation as HMS 2382, in the same nucleotide and in the same IS

copy (DR region). In addition, IS*6110* has been inserted at exactly the same point in the three strains. As both L6 and L4 are separate phylogenetic branches, it is not likely that they acquired the mutation from a common ancestor; therefore, the mutation could be random. There is one mutation in the IR of the IS*6110* for HMS 1693 and HMS 14017 (L6). These two strains seem to be from the same cluster according to their RFLP patterns, a finding that would be supported by this shared mutation. As it is out of the coding region, we do not know if it affects the transposase somehow.

Among the findings in the other MTBC families studied is the mutation in the IS*6110* copy of the DR region in HMS 18014 (T family), which changes threonine to isoleucine. Both amino acids have a similar size, but they are in different classification groups. Threonine has a polar lateral chain without a charge, whereas isoleucine has a hydrophobic lateral chain, so the interactions will be different. Notably, the three Haarlem strains studies have a different mutation in the copy located in *Rv0963c*, one with a mutation in the IR of the IS (HMS 18037) and two with a synonymous mutation in the coding region of the transposase (HSJ 234 and HMS 18031). These mutations seem to be random as they are not the same or conserved among strains. The chance of three different mutations in the same location could be because a high number of these copies were sequenced. HMS 18025 (LAM) has a mutation in the copy inserted in *lpqQ:Rv0836c*, with a change from glycine to arginine. This is an important structural change because arginine is larger than glycine and has a positive charge. More studies are needed to determine how these mutations affect the transposition ability.

In total, we studied 38 IS copies located in the DR region. We successfully amplified this copy in 30 strains of our collection and examined eight more from the NCBI strains. Among them, we have found three different variants in five strains, described above. This is the copy in which we have found the most mutations, a phenomenon that can be explained because this is the region for which more IS copies have been studied, even though the percentage of mutation is constant with regard to the total mutations found. As almost all MTBC strains have an IS copy in this region, it would be reasonable to think that this copy could have been the first IS*6110* that started to transpose and thus has increased its number within the genome (McEvoy et al., 2007). If this is true, we have to assume that mutations in this copy occurred after the transposition, as the other studied copies do not have any mutations. Again, the exception is HCU 3717 (X family), in which all its IS*6110* copies, including the two of the DR region, seems to be mutated (in the absence of knowing the genotype of the copy located in *ppe46*).

Ates et al. (2018a) showed that modern Beijing families have different deletions involving the *ppe38* locus, suggesting that these changes may contribute to a higher growth rate *in vivo* and lung inflammation due to the block of PE_PGRS secretion, rendering some Beijing strains hypervirulent. The most virulent deletion described in their work was the one that disrupts the *ppe38* and *ppe71* genes with the loss of the *esx* genes between *ppe38* and *ppe71*. This is the deletion we found in the Beijing GC1237 strain (**Figure 3**). We found that this IS copy has no direct repeats, which suggests that there has been a recombination event between two IS*6110* copies with the removal of the genes between them (Brosch et al., 1999; Sampson et al., 2003). Regarding the MtZ strain, there is an IS*6110* within the *ppe38* gene and another one in the *ppe71* gene. We do not know how the insertion affects both genes. Ates et al. (2018b) demonstrated that a single copy of these *ppe* genes is enough to support PE_PGRS secretion; therefore, MtZ could have this secretion affected. Several experiments are ongoing to determine whether the secretion of PE_PGRS in the MtZ strain is altered. McEvoy et al. (2009) found IS*6110* insertion events in the *ppe38* locus in other MTBC families. As we also found this in the MtZ strain, we conclude that it is not something specific to the Beijing family. As we have stated previously, none of the IS*6110* copies amplified that are inserted in *ppe* genes have a mutation.

In summary, we have studied the genetic variability of IS*6110*. We have analysed the DNA sequence of the different IS*6110* copies in the same strain of *M. tuberculosis* as well as in different MTBC strains. We have observed mutation in 13.3% of the copies studied. Gly215Ser (*orfB*) mutation seems to be characteristic of the X family. In general, the high copy number strains analysed carry wild-type copies whereas several LCN strains, such as the X family and L6, have one or more of their copies mutated. Some strains share the location as well as mutations in these copies, allowing us in these cases to establish the transposition events that have occurred over time and indicating a close relationship in the evolution of these strains. The detailed study of each of the copies has also allowed us to provide information regarding the evolution of some MTBC families. Many publications have studied the variability among strains of the MTBC caused by changes in the number and location of IS*6110*. To these, we add the changes produced over time in the sequence itself as one more factor that will probably affect its evolution.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/; MZ574181–MZ574188.

AUTHOR CONTRIBUTIONS

SS and IO: conceptualised the work. JC: carried out the laboratory experiments and curated the data. JC, IO, and SS: wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Carlos III Health Institute in the context of a Grant (FIS18/0336), and JC was awarded

a scholarship by the Government of Aragon/European Social Fund, "Building Europe from Aragon."

ACKNOWLEDGMENTS

The authors would like to acknowledge the use of Servicio General de Apoyo a la Investigación-SAI, Universidad de Zaragoza (Servicio de Análisis Microbiológico), and Servicios Científico Técnicos, IACS (Servicio de Secuenciación y Genómica Funcional and Servicio de Biocomputación). We would like to thank the EPIMOLA group for supplying the genotyped bacterial DNA used in this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.767912/full#supplementary-material

REFERENCES

Alonso, H., Aguilo, J. I., Samper, S., Caminero, J. A., Campos-Herrero, M. I., Gicquel, B., et al. (2011). Deciphering the role of IS6110 in a highly transmissible *Mycobacterium tuberculosis* Beijing strain, GC1237. *Tuberculosis* 91, 117–126. doi: 10.1016/j.tube.2010.12.007

Ates, L. S., Dippenaar, A., Ummels, R., Piersma, S. R., Van Der Woude, A. D., Van Der Kuij, K., et al. (2018a). Mutations in ppe38 block PE-PGRS secretion and increase virulence of *Mycobacterium tuberculosis*. *Nat. Microbiol.* 3, 181–188. doi: 10.1038/s41564-017-0090-6

Ates, L. S., Sayes, F., Frigui, W., Ummels, R., MPM, D., Bottai, D., et al. (2018b). RD5-mediated lack of PE_PGRS and PPE-MPTR export in BCG vaccine strains results in strong reduction of antigenic repertoire but little impact on protection. *PLoS Pathog.* 14:e1007139. doi: 10.1371/journal.ppat.1007139

Beggs, M. L., Eisenach, K. D., and Cave, M. D. (2000). Mapping of IS6110 insertion sites in two epidemic strains of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* 38, 2923–2928. doi: 10.1128/JCM.38.8.2923-2928.2000

Brisson-Noel, A., Aznar, C., Chureau, C., Nguyen, S., Pierre, C., Bartoli, M., et al. (1991). Diagnosis of tuberculosis by DNA amplification in clinical practice evaluation. *Lancet* 338, 364–366. doi: 10.1016/0140-6736(91)90492-8

Brosch, R., Philipp, W. J., Stavropoulos, E., Colston, M. J., Cole, S. T., and Gordon, S. V. (1999). Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra strain. *Infect. Immun.* 67, 5768–5774. doi: 10.1128/IAI.67.11.5768-5774.1999

Comín, J., Cebollada, A., Ibarz, D., Viñuelas, J., Vitoria, M. A., Iglesias, M. J., et al. (2021). A whole-genome sequencing study of an X-family tuberculosis outbreak focus on transmission chain along 25 years. *Tuberculosis* 126:102022. doi: 10.1016/j.tube.2020.102022

Comin, J., Chaure, A., Cebollada, A., Ibarz, D., Viñuelas, J., Vitoria, M. A., et al. (2020). Investigation of a rapidly spreading tuberculosis outbreak using whole-genome sequencing. *Infect. Genet. Evol.* 81:104184. doi: 10.1016/j.meegid.2020.104184

Comín, J., Monforte, M. L., Samper, S., and Otal, I. (2021). Analysis of *Mycobacterium africanum* in the last 17 years in Aragon identifies a specific location of IS6110 in lineage 6. *Sci. Rep.* 11:10359. doi: 10.1038/s41598-021-89511-x

Dale, J. W. (1995). Mobile genetic elements in mycobacteria. *Eur. Respir. J. Suppl.* 20, 633s–648s.

Dale, J. W., Tang, T. H., Wall, S., Zainuddin, Z. F., and Plikaytis, B. (1997). Conservation of IS6110 sequence in strains of *Mycobacterium tuberculosis* with single and multiple copies. *Tuber. Lung Dis.* 78, 225–227. doi: 10.1016/S0962-8479(97)90002-2

Fang, Z., Doig, C., Morrison, N., Watt, B., and Forbes, K. J. (1999). Characterization of IS1547, a new member of the IS900 family in the *Mycobacterium tuberculosis* complex, and its association with IS6110. *J. Bacteriol.* 181, 1021–1024. doi: 10.1128/JB.181.3.1021-1024.1999

Fang, Z., and Forbes, K. J. (1997). A *Mycobacterium tuberculosis* IS6110 preferential locus (ipl) for insertion into the genome. *J. Clin. Microbiol.* 35, 479–481. doi: 10.1128/jcm.35.2.479-481.1997

Gonzalo-Asensio, J., Pérez, I., Aguiló, N., Uranga, S., Picó, A., Lampreave, C., et al. (2018). New insights into the transposition mechanisms of IS6110 and its dynamic distribution between *Mycobacterium tuberculosis* complex lineages. *PLoS Genet.* 14:e1007282. doi: 10.1371/journal.pgen.1007282

Hermans, P. W. M., Van Soolingen, D., Bik, E. M., De Haas, P. E. W., Dale, J. W., and Van Embden, J. D. A. (1991). Insertion element IS987 from Mycobacterium bovis BCG IS located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect. Immun.* 59, 2695–2705. doi: 10.1128/iai.59.8.2695-2705.1991

Isabel Millan-Lou, M., Isabel López-Calleja, A., Colmenarejo, C., Antonia Lezcano, M., Asunción Vitoria, M., Del Portillo, P., et al. (2013). Global study of is6110 in a successful *Mycobacterium tuberculosis* strain: clues for deciphering its behavior and for its rapid detection. *J. Clin. Microbiol.* 51, 3631–3637. doi: 10.1128/JCM.00970-13

McEvoy, C. R. E., Falmer, A. A., van Pittius, N. C. G., Victor, T. C., van Helden, P. D., and Warren, R. M. (2007). The role of IS6110 in the evolution of *Mycobacterium tuberculosis*. *Tuberculosis* 87, 393–404. doi: 10.1016/j.tube.2007.05.010

McEvoy, C. R., Van Helden, P. D., Warren, R. M., and Van Pittius, N. C. G. (2009). Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic *Mycobacterium tuberculosis* PPE38 gene region. *BMC Evol. Biol.* 9, 237–221. doi: 10.1186/1471-2148-9-237

Mendiola, M. V., Martin, C., Otal, I., and Gicquel, B. (1992). Analysis of the regions responsible for IS6110 RFLP in a single *Mycobacterium tuberculosis* strain. *Res. Microbiol.* 143, 767–772. doi: 10.1016/0923-2508(92)90104-V

Otal, I., Martin, C., Vincent-Levy-Frebault, V., Thierry, D., and Gicquel, B. (1991). Restriction fragment length polymorphism analysis using IS6110 as an epidemiological marker in tuberculosis. *J. Clin. Microbiol.* 29, 1252–1254. doi: 10.1128/jcm.29.6.1252-1254.1991

Reyes, A., Sandoval, A., Cubillos-Ruiz, A., Varley, K. E., Hernández-Neuta, I., Samper, S., et al. (2012). IS-seq: a novel high throughput survey of in vivo IS6110 transposition in multiple *Mycobacterium tuberculosis* genomes. *BMC Genomics* 13:249. doi: 10.1186/1471-2164-13-249

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genome viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Roychowdhury, T., Mandal, S., and Bhattacharya, A. (2015). Analysis of IS6110 insertion sites provide a glimpse into genome evolution of *Mycobacterium tuberculosis*. *Sci. Rep.* 5, 1–10. doi: 10.1038/srep12567

Safi, H., Barnes, P. F., Lakey, D. L., Shams, H., Samten, B., Vankayalapati, R., et al. (2004). IS6110 functions as a mobile, monocyte-activated promoter in *Mycobacterium tuberculosis*. *Mol. Microbiol.* 52, 999–1012. doi: 10.1111/j.1365-2958.2004.04037.x

Sagasti, S., Millán-Lou, M. I., Soledad Jiménez, M., Martín, C., and Samper, S. (2016). In-depth analysis of the genome sequence of a clinical, extensively drug-resistant *Mycobacterium bovis* strain. *Tuberculosis* 100, 46–52. doi: 10.1016/j.tube.2016.06.005

Sampson, S. L., Warren, R. M., Richardson, M., van der Spuy, G. D., and van Helden, P. D. (1999). Disruption of coding regions by IS6110 insertion in *Mycobacterium tuberculosis*. *Tuber. Lung Dis.* 79, 349–359. doi: 10.1054/tuld.1999.0218

Sampson, S. L., Warren, R. M., Richardson, M., Victor, T. C., Jordaan, A. M., Van der Spuy, G. D., et al. (2003). IS6110-mediated deletion polymorphism in the direct repeat region of clinical isolates of *Mycobacterium tuberculosis*. *J. Bacteriol.* 185, 2856–2866. doi: 10.1128/JB.185.9.2856-2866.2003

Sekine, Y., Izumi, K. I., Mizuno, T., Ohtsubo, E., and Ishihama, A. (1997). Inhibition of transpositional recombination by OrfA and OrfB proteins encoded by insertion sequence IS3. *Genes Cells* 2, 547–557. doi: 10.1046/j.1365-2443.1997.1440342.x

Soto, C. Y., Menéndez, M. C., Pérez, E., Samper, S., Gómez, A. B., García, M. J., et al. (2004). IS6110 mediates increased transcription of the phoP virulence gene in a multidrug-resistant clinical isolate responsible for tuberculosis outbreaks. *J. Clin. Microbiol.* 42, 212–219. doi: 10.1128/JCM.42.1.212-219.2004

Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., et al. (2016). *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* 48, 1535–1543. doi: 10.1038/ng.3704

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rüsch-Gerdes, S., Willery, E., et al. (2006). Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* 44, 4498–4510. doi: 10.1128/JCM.01392-06

Tanaka, M. M. (2004). Evidence for positive selection on *Mycobacterium tuberculosis* within patients. *BMC Evol. Biol.* 4, 31–38. doi: 10.1186/1471-2148-4-31

Thabet, S., Namouchi, A., and Mardassi, H. (2015). Evolutionary trends of the transposase-encoding open reading frames A and B (orfA and orfB) of the mycobacterial IS6110 insertion sequence. *PLoS One* 10:e0130161. doi: 10.1371/journal.pone.0130161

Thierry, D., Brisson-Noel, A., Vincent-Levy-Frebault, V., Nguyen, S., Guesdon, J. L., and Gicquel, B. (1990). Characterization of a *Mycobacterium tuberculosis* insertion sequence, IS6110, and its application in diagnosis. *J. Clin. Microbiol.* 28, 2668–2673. doi: 10.1128/jcm.28.12.2668-2673.1990

Vera-Cabrera, L., Hernández-Vera, M. A., Welsh, O., Johnson, W. M., and Castro-Garza, J. (2001). Phospholipase region of *Mycobacterium tuberculosis* is a preferential locus for IS6110 transposition. *J. Clin. Microbiol.* 39, 3499–3504. doi: 10.1128/JCM.39.10.3499-3504.2001

Wall, S., Ghanekar, K., McFadden, J., and Dale, J. W. (1999). Context-sensitive transposition of IS6110 in mycobacteria. *Microbiology* 145, 3169–3176. doi: 10.1099/00221287-145-11-3169

Warren, R. M., Sampson, S. L., Richardson, M., Van Der Spuy, G. D., Lombard, C. J., Victor, T. C., et al. (2000). Mapping of IS6110 flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. *Mol. Microbiol.* 37, 1405–1416. doi: 10.1046/j.1365-2958.2000.02090.x

Yesilkaya, H., Dale, J. W., Strachan, N. J. C., and Forbes, K. J. (2005). Natural transposon mutagenesis of clinical isolates of *Mycobacterium tuberculosis*: how many genes does a pathogen need? *J. Bacteriol.* 187, 6726–6732. doi: 10.1128/JB.187.19.6726-6732.2005

Conflict of Interest: 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.

Table S1. Primers used and insertion points for the IS*6110* amplification in MtZ strain. *Point in CDC1551 strain genome. **Points in *M. bovis* AF2122/97 genome.

Primer	Sequence	Point of insertion	IS*6110* direction	Reference
Rv0755-F	ATCTTGATCGCATCGGAAGC	850536	Forward	27
Rv0755-R	AACCCGCTGAGCAGCATCGC			27
Rv0795-F	CCTATCGCCCACCAGACGC	889029	Forward	27
Rv0795-R	ACACCGTGCGCCTCTACCTGC			27
Rv1668-F	ACCCGCAATCCGCTACGC	1895651	Reverse	27
Rv1668-R	AGCACGAATCATGGACTCGGC			27
Rv2286c-F	AGCTCATGACCGATCGCTGC	2559568	Reverse	27
Rv2286c-R	ACCGCCGCGTTGACCTGTTTCG			27
Rv2349-F	AATGCCGACGTCATATCGCC	2627492	Forward	27
Rv2349-R	ACGTCATCAACAACACGCTGC			27
DR43-F	ACCCGGTGCGATTCTGCG	3119660-3121879	Reverse	27
DR22-R	AGACGGCACGATTGAGAC			27
Rv2823c-F	AAGAGCTGTGCGGTCAGGC	3129520	Reverse	27
Rv2823c-R	AAGGTGATCGAGGAGAAGTACCGGC			27
Rv3229c-F	AGCATCTGCCGTAGGTACCAC	3606308	Reverse	27
Rv3229c-R	AACACCATCCTTGCGATCG			27
Cut1-F	ACGTATCCGAGAGTGTGAC	1979901-1988923**	Forward	27
RvD3-R	CAATCATGCTGTTTGGC			27
PPE38-F	AAGTGCGGATTTTCGGTGTGG	2633977	Reverse	27
PPE38-R	ATCAAACGAGGACGCCGAGG			27
MT2423-F	GTTGCCGTTGTTTCCGTTAC	2633461	Reverse	27
MT2423-R	CAGCGGTAACCAAGGTGAC			27
MT3429-F	GCAATCAGAACGTCGGTGT	3709089*	Forward	This work
RvD5-R	GTACCCGCACCACCTGCT			27
PPE71-F	CCGTTACCGAAGTTCGTGTC	3078253**	Reverse	This work
PPE71-R	CAACCTGGGCAGCGGTA			This work

Table S2. Primers used and insertion points for the IS*6110* amplification in CG1237 Beijing strain. *Point in *M. bovis* AF2122/97 genome.

Primer	Sequence	Point of insertion	IS*6110* direction	Reference
DnaAj-F	CAATCGACAAAGCGCTGGC	1594	Forward	This work
DnaAj-R	TGGGGTGTGTGTTGGGT			This work
Rv0795j-F	TGGCGCCTGGTCCTGTTG	889072	Reverse	This work
Rv0795r	GTCCTTCTCACCTAGTGCCG			28
pipF	ACGAATGGCTGAAGATGTGAA	937116	Forward	28
pipR	AATCGGAGGTCAAGTGGA			28
Rv1371d	ATCGTTGCAGTACTGGCCGTGGG	1543972	Reverse	28
Rv1371r	TATACGGGTGGTGCAGCAGGGTG			28
CtpDF	ACCCCGCCAGAATGGTTAATCC	1657016	Reverse	28
CtpDR	TTGTTCCGCAACGCTTGTGC			28
RD152d	CCGGGTTGAGCAATGCGGATATCAGTGGAC	1986638-1998625	Reverse	40
RD152r	TGGGATAGTTCAGGTGGCCATCGTGGGCAT			40

PPE34t	TGCCCGACGTTGTAGATTCCC	2163392	Reverse	28
PPE34-R	CCTGGGCATTAACAGTTC			This work
Rv2016d	ATGCATAGCCGGTGTCTGTCC	2263627	Reverse	28
Rv2016r	AATGCGAACCTCTTCCAGCC			28
Mb2047cd	CCTAAAAGGATAGCGTTGAA	2248012*	Forward	28
Mb2047cr	TGTTCTCAAGTGTCCCCAA			28
Rv2078d	ATCGAGACCACGCAGAGGGC	233468	Forward	28
Rv2078r	AGTCCTCATGATGCGCACG			28
Rv2180c-F	CGGTGTCGTAGAAGTACCGC	2442348	Reverse	This work
Rv2180c-R	GGATCAAGCTGTGGCTGCA			This work
Rv2286cj-F	TCCCACGGGTGCTTCTTG	2559506	Reverse	This work
Rv2286cj-R	GCGCAACGATAACGACGAG			This work
PPE38BG-F	GATCACCGCATCAAACGAGG	2634048-2639267	Forward	This work
PPE38BG-R	CGGGTCAATTGAGTCATCTGG			This work
RD207d	GACGAGTTCGCGCTCAAAATGT	3076124-3084489*	Forward	28
RD207r	CCCCGGCGAGGAACAGAA			28
PPE46BG-F	CGTCACAACGACCCACC	3378553	Reverse	This work
PPE46BG-R	ATCTGACGGCGAGTAATTGG			This work
Rv3326cj-F	CACGAATCGGGCCGTTAGC	3711737	Reverse	This work
3326R	ACTTGTGCGCATCGGTTCC			28
IdsBd	TTTGAGGATTGTTTATTGGAGGG	3797823	Forward	28
IdsBr	AACGCAAGACCCAACTCATGGC			28
Rv3427cd	TATCGCACCATCAAGGGC	3844681	Reverse	28
Rv3427cr	AATTCCAGATGCCCCAGG			28

Table S3. Primers used and insertion points for the IS*6110* amplification in low copy number strains.

Primer	Sequence	Point of insertion	Strains	IS*6110* direction	Reference
DR43-F	CCCGGTGCGATTCTGCG	3120523-3121879	HMS 2382 HMS 2485 HMS 2445	Reverse	27
DR22-R	AGACGGCACGATTGAGAC		HMS 2382 HMS 2407 HCU 3445 HMS 2485 HMS 2445 *M. bovis* B		27
DR-ISA3-F	CCTGTATTTCGCTGGTTTCCGTC		HMS 2407 HCU3445		This work
DR-F	ACAACCTCGCCCTGCAAG		*M. bovis* B		27
Rv0963c-F	ACCGGTGTTGACCGACAG	1075948	HMS 2382	Reverse	This work
Rv0963c-R	CAGTGACCCGCGAAAGGTG				This work
MoaX-F	ATCGGGTCATTACCGGCGGC	3709622	HMS 2382 HMS 2407	Reverse	26
MoaX-R	CCAGTCGACGCGGTTGGGG				26
LipX-F	GCTCAGGCTCTCATCGTCG	1300195	HMS 2407	Reverse	This work
LipX-R	GCCGTTTCCCCAATCGAATC				This work
PhoP-F	GCCGTCCATCCCGGGCATC	851536	*M. bovis* B	Forward	30

PhoP-R	CCATGTTCAAACCGGTGTC				30
Rv0403c-F	ATTGCTGAAAGTCCTTGCCGATG	483296	HMS 2485 HMS 2445	Reverse	25
Rv0403c-R	GCGGCTGCACTCGGTGTT				25

Table S4. Primers used for the expansion of the study in LAM, Haarlem and other families of *M. tuberculosis* and insertion points. *Point in CDC1551 genome.

Primer	**Sequence**	**Point of insertion**	**Strains**	**IS*6110* direction**	**Reference**
Rv0835d	TTGCTCCACTGCTGCCAAGTCGG	932202	LAM	Reverse	28
Rv0836d	AACAATTGGGACCACTTCGAGG				28
Rv3113d	CAATTCATCGCGCCGCTGT	3480371	LAM	Forward	28
Rv3113r	AAAATAGGTGTTGCCACCCG				28
Rv1754c-Up	CGGGTCTCCTGGGTGATTC	1986623	LAM and Haarlem	Reverse/ Forward	25
Rv1754c-Lo	GTATGCCTCCGTGACCGTGT				25
Rv3324Ar	TAAACCGTGAGCGCTGTCACC	3705513*	LAM	Reverse	28
MT3427-R	CAATGACGTTGTGCAGAT				This work
RvD5-R	GTACCCGCACCACCTGCT				27
Rv0403c-F	ATTGCTGAAAGTCCTTGCCGATG	483296	Haarlem	Reverse	25
Rv0403c-R	GCGGCTGCACTCGGTGTT				25
Rv0963c-F	AGCTTCCTTGACCATGTCCC	1075948	Haarlem	Reverse	This work
Rv0963c-R	TGCTGCAACGAGAACTCACC				This work
Rv2336-Up	AATGCCGTCGTGGTCAA	2610859	Haarlem	Forward	25
Rv2336-Lo	CGGTTTCTCGGGTGCTAC				25
pks9-F	AGAATTCGGCGCGGTATC	1889066	S	Reverse	This work
pks9-R	GGCTGAGCAAACTCTGTGCT				This work
DR-18	CGGGCGAGCTGCAGATG	DR region	LAM	Reverse	This work

Publicación 3

scientific reports

Check for updates

OPEN

Estimation of the mutation rate of *Mycobacterium tuberculosis* in cases with recurrent tuberculosis using whole genome sequencing

Jessica Comín[1]✉, Alberto Cebollada[2], Aragonese Working Group on Molecular Epidemiology of Tuberculosis (EPIMOLA)* & Sofía Samper[1,3,4]

The study of tuberculosis latency is problematic due to the difficulty of isolating the bacteria in the dormancy state. Despite this, several in vivo approaches have been taken to mimic the latency process. Our group has studied the evolution of the bacteria in 18 cases of recurrent tuberculosis. We found that HIV positive patients develop recurrent tuberculosis earlier, generally in the first two years (*p* value = 0.041). The genome of the 36 *Mycobacterium tuberculosis* paired isolates (first and relapsed isolates) showed that none of the SNPs found within each pair was observed more than once, indicating that they were not directly related to the recurrence process. Moreover, some IS*6110* movements were found in the paired isolates, indicating the presence of different clones within the patient. Finally, our results suggest that the mutation rate remains constant during all the period as no correlation was found between the number of SNPs and the time to relapse.

Mycobacterium tuberculosis has afflicted and co-evolved with man over thousands of years. Its success is due to its ability to infect a host and persist in a dormancy state for years[1,2]. During this period, the host is asymptomatic and not infectious, making the study of this state unmanageable. The bacteria stay in the granuloma, a barrier made by the immune systems cells, until not-well characterised signals or a weakening of the immune system allows the bacteria to escape and develop an active disease[3]. In recent years, some in vitro approaches have been designed for trying to increase the knowledge of latency[4] and a debate about the physiological state of *M. tuberculosis* during this period has emerged. The number of mutations during a time period can be used as a molecular clock to study the evolution of the pathogen[6–9]. There were two studies with apparently contradictory results. Ford et al.[10] used bacteria isolated from macaque lesions that mimic those of tuberculosis (TB) latent infection and concluded that generation time for latent TB would be similar to active TB, so the bacteria is physiologically active. On the other hand, Colangeli et al.[11], studying an outbreak in New Zealand, concluded that generation times during latency are longer than during active TB. Recently Colangeli et al.[12], pairing index cases to their TB contacts as a latency approach, concluded that both studies were correct: during the first two years of latency, the generation time is similar to that of the active disease, while later it starts to increase for long periods of time and a reduced mutation rate is observed. Looking for a novel in vivo approach, we carried out the analysis of isolates from individuals known to have developed several episodes of active TB for studying the evolution of the bacteria during the period between these episodes. Since 2004, all strains of *M. tuberculosis* have been genotyped in Aragon, Spain, which allowed us to identify the cases of TB relapses. The DNA previously used for genotyping remained in storage and could be used for whole genome sequencing (WGS) to analyse the variability of the isolates in the different episodes of the disease.

Results

Patient selection and risk factors. The search of cases with recurrent TB among the total of cases in Aragon revealed 127 patients from 2004 until 2019 (4.97%). The genotype of the isolates revealed that 114 patients were infected by the same or very similar RFLP pattern strain, which would imply a potential relapse,

[1]Instituto Aragonés de Ciencias de la Salud, C/de San Juan Bosco, 13, 50009 Zaragoza, Spain. [2]Unidad de Biocomputación, Instituto Aragonés de Ciencias de la Salud, C/de San Juan Bosco, 13, 50009 Zaragoza, Spain. [3]Fundación IIS Aragón, C/de San Juan Bosco, 13, 50009 Zaragoza, Spain. [4]CIBER de Enfermedades Respiratorias, Av. Monforte de Lemos, 3-5. Pabellón 11, Planta 0, 28029 Madrid, Spain. *A list of authors and their affiliations appears at the end of the paper. ✉email: jcomin.iacs@aragon.es

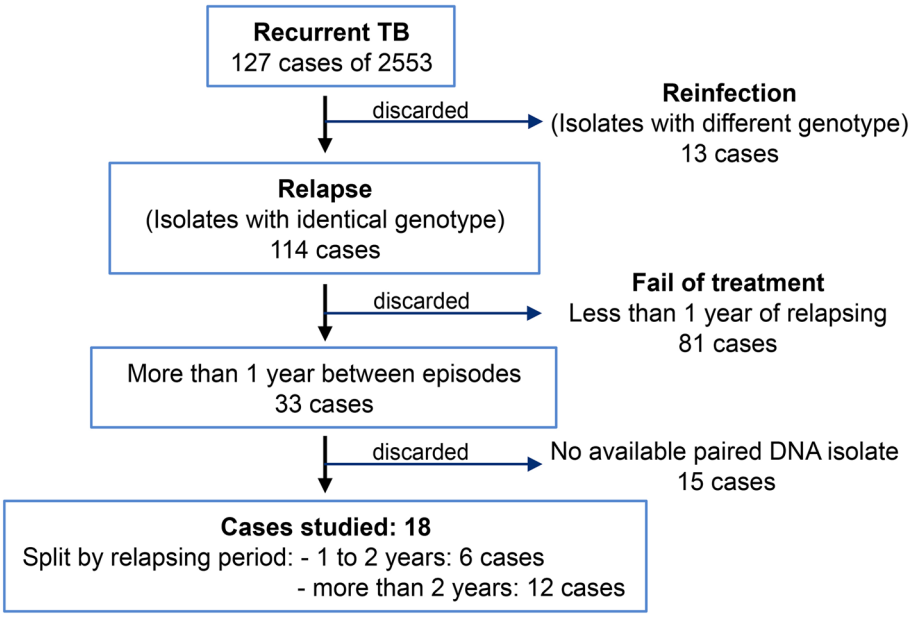


Figure 1. Diagram with the discarded and selected cases for the recurrent cases study.

while 13 patients were infected with non-related strains, i.e., re-infections, so these were not considered for this study. Among the potential relapses, we selected the cases with at least one year between episodes. Eighty-one patients had isolates with less than one year between them, therefore this group was discarded. Based on the time distance between the isolates, the cases were split into two groups: cases with ≥ 1 year but ≤ 2 years between isolates (12 patients), and cases with more than two years between the diagnosis of their isolates (21 patients). Eighteen pairs of *M. tuberculosis* isolates with available DNA of at least two different episodes were studied: twelve patients in the > 2 years group and six patients in the ≥ 1 but ≤ 2 years (Fig. 1). The lineages of the selected isolates are shown in Fig. 2.

Several selected patients had risk factors to develop TB. At least 38.9% were HIV+, 22.2% declared a high alcohol consumption, 38.9% were smokers and 16.7% were intravenous (IV) drugs users. The treatment for all patients was the standard for susceptible TB. Despite some of them did not follow it correctly, no drug resistance was developed. We split the patients into two groups: [1–2 years until relapse] and (2–14 years until relapse] in order to investigate if some of these risk factors were related to a shorter or longer relapsing period (time between the first and the second episode). The intervals were fixed according to the results obtain by Colangeli et al.[12], being 160 months the maximum time between episodes observed in our study. Results are shown in Table 1.

HIV status was significant (p value = 0.041) between the two groups, showing that HIV positive patients suffered relapse in the first two years more frequently than HIV negative patients. The gender was marginally significant for males, suggesting a trend of males suffering relapses in the first two years more frequently than females (p value = 0.066).

Analysis of the genomes. *SNPs versus relapsing period.* The number of existent SNPs between the first and its correspondent relapsed isolate ranged from 0 to 8. These SNPs were usually in the relapsed isolates, but we could also find some of them in the earliest isolates that then disappeared, showing different clones co-existing in the patient. The mutation effect of the SNPs and also the functional categories of the affected genes were analyzed; 26.3% belonged to cell wall and cell processes category and 21.1% to intermediary metabolism and respiration category. None of them were present in more than one case, therefore they do not seem to be directly implicated in relapsing. The detailed SNPs can be found in Table 2.

In order to represent the number of SNPs developed between the first and the relapsed isolates of the same patient versus the time between the diagnosis of both isolates (in months), resembling in that way the latency period, we reproduced the study of Colangeli et al.[12] using the Poisson regression model. Equally to Colageli et al. 2020, we found this correlation not significant (p value = 0.34), meaning that those isolates with a longer relapsing period were not necessarily those with more SNPs. The results are shown in Fig. 3. Otherwise, we observed that pairs of L4.1 had a higher mutation rate per genome per year (0.93 SNPs) than other sublineages (0.58 SNPs in L4.8 and 0.32 in L4.3), and it is above the average found by this study (0.64 SNPs).

Mutation rate versus generation time. In order to analyse the correlation between the mutation rate and the relapsing period, we used the Poisson model, as described by Colangeli et al. [12]. The generation time was fixed at 18 h as seen in *M. tuberculosis* actively replicating in vitro. Results are shown in Fig. 4. The mutation rate tends to diminish in longer relapsing periods, being marginally significant (p value = 0.061).

When we considered the data in (1–2) and (2–14) years of relapsing periods, the mutation rate is slightly lower for the second, estimated at 2.728×10^{-10} [95% CI: 1.433×10^{-10}, 5.193×10^{-10}] mutations per (bp × generation),

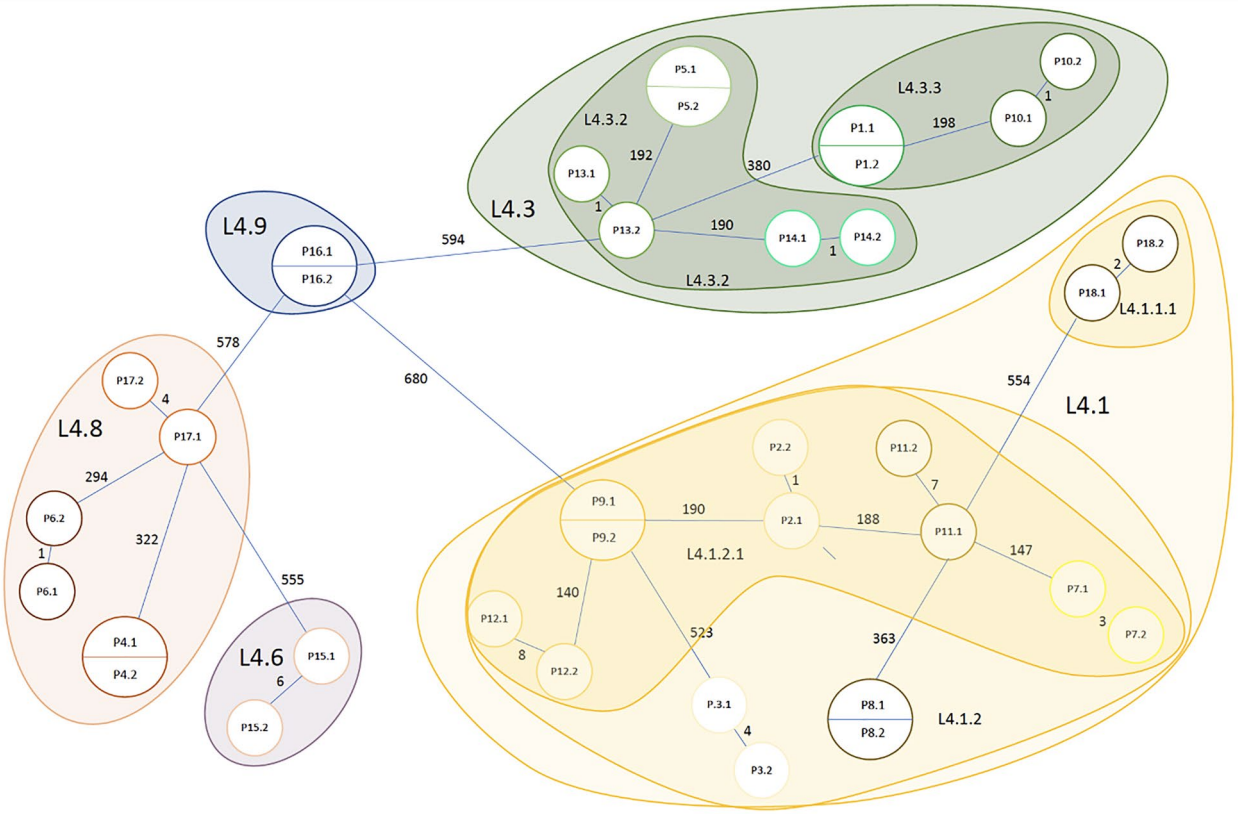


Figure 2. Drawing of a minimal spanning tree (not scaled) with all the studied isolates. The number of SNPs and the TB lineages are indicated.

	[1–2 years]	(2–14 years]	
	N = 6	*N* = 12	*p* value
Gender			0.066
Male	7 (87.5%)	4 (40.0%)	
Female	1 (12.5%)	6 (60.0%)	
HIV status			0.041
Negative	2 (25.0%)	7 (87.5%)	
Positive	6 (75.0%)	1 (12.5%)	
Alcohol consumption			1.000
High	1 (25.0%)	3 (42.9%)	
Low/No	3 (75.0%)	4 (57.1%)	
Immunosuppression			0.072
No	1 (20.0%)	6 (85.7%)	
Yes	4 (80.0%)	1 (14.3%)	
Smoker			0.576
No	2 (50.0%)	2 (28.6%)	
Yes	2 (50.0%)	5 (71.4%)	
Users of IV drugs			0.491
No	2 (50.0%)	6 (85.7%)	
Yes	2 (50.0%)	1 (14.3%)	

Table 1. Risk factors of the 18 selected cases.

Pair	Relapsing period	N	Point	Gene	Functional category	Mutation effect
P1	13 m/8736 h	0				
P2	31 m/20,832 h	1	1,042,251	*pstS1*	Cell wall and cell processes	Non-synonymous
P3	92 m/61,842 h	4	2,793,634	*lipQ*	Intermediary metabolism and respiration	Non-synonymous
			3,031,623*	*Rv2719c:lexA*	Intergenic region	
			3,869,344	*Rv3448*	Cell wall and cell processes	Synonymous
			4,134,177*	*moxR2*	Regulatory proteins	Non-synonymous
P4	12 m/8064 h	0				
P5	51 m/34,272 h	0				
P6	13 m/8736 h	1	156,912	*fbpC*	Lipid metabolism	Non-synonymous
P7	64 m/43,008 h	3	1,480,972*	*Rv1319c*	Intermediary metabolism and respiration	Synonymous
			1,845,545	*uvrA*	Information pathways	Synonymous
			3,179,330	*Rv2867c:ispG*	Intergenic region	Synonymous
P8	160 m/107,520 h	0				
P9	31 m/20,832 h	0				
P10	31 m/20,832 h	1	245,322	*mmpL3*	Cell wall and cell processes	Non-synonymous
P11	42 m/28,224 h	7	144,506*	*fadD7*	Lipid metabolism	Non-synonymous
			157,655	*fbpC:htdZ*	Intergenic region	
			1,952,766*	*Rv1726*	Intermediary metabolism and respiration	Synonymous
			2,278,535	*hspX*	Virulence, detoxification, adaptation	Non-synonymous
			3,912,404	*mce4F*	Virulence, detoxification, adaptation	Non-synonymous
			456,028	*secE2*	Cell wall and cell processes	Non-synonymous
			1,837,204	*uvrB*	Information pathways	Non-synonymous
P12	39 m/26,208 h	8	1,049,482*	*Rv0939*	Intermediary metabolism and respiration	Synonymous
			2,124,862*	*Rv1875:Rv1876*	Intergenic Region	
			2,145,578	*lppD*	Cell wall and cell processes	Non-synonymous
			2,555,780*	*Rv2282c*	Regulatory proteins	Non-synonymous
			3,111,151*	*Rv2802c*	Conserved hypotheticals	Non-synonymous
			4,364,271*	*mycP1*	Intermediary metabolism and respiration	Non-synonymous
			4,394,960*	*Rv3909*	Conserved hypotheticals	Non-synonymous
P13	32 m/21,504 h	1	3,574,424	*Rv3201c*	Information pathways	Synonymous
P14	14 m/9408 h	1	1,543,706	*Rv1371*	Cell wall and cell processes	Synonymous
P15	104 m/69,888 h	6	1,505,180	*murI*	Cell wall and cell processes	Non-synonymous
			2,477,975	*Rv2212*	Intermediary metabolism and respiration	Synonymous
			2,675,882	*mbtB:mbtA*	Intergenic Region	
			2,701,688	*lepA*	Intermediary metabolism and respiration	Non-synonymous
			3,160,108	*Rv2851c*	Intermediary metabolism and respiration	Synonymous
			3,470,698*	*ftsE*	Cell wall and cell processes	Synonymous
P16	14 m/9408 h	0				
P17	59 m/39,648 h	4	1,840,457*	*Rv1634*	Cell wall and cell processes	Synonymous
			1,923,131	*Rv1698*	Cell wall and cell processes	Synonymous
			4,338,635	*whiB6:Rv3863*	Intergenic region	
			4,400,765	*sigM*	Information pathways	Non-synonymous
P18	16 m/10,752 h	2	397,372	*Rv0331:Rv0332*	Intergenic region	
			1,247,257	*ephC*	Virulence, detoxification, adaptation	Non-synonymous

Table 2. SNPs found between the first and the relapsed paired isolates. Relapsing period (time between the first episode and the second) in months (m) and hours (h), Number of SNPs found between the paired isolates (N), Point referred to H37Rv genome, Gene name, Functional Category and the Effect of the SNP are detailed. *SNPs detected in the first isolate and not in the relapsed isolate.

than in the first period, estimated at 2.798×10^{-10} [95% confidence interval (CI): 1.209×10^{-10}, 6.477×10^{-10}]. However, this difference was not statistically significant (p value = 0.96). Results are shown in Fig. 5.

IS6110 copies variation between the isolates. All the IS*6110* copies of the isolates were analyzed using the WGS data. We found several IS*6110* movements between the first and the relapsed isolates as a result of different clones. The relapsed strain gained extra IS*6110* copies in P7 (two copies gained) and P12 (one copy gained). On the other hand, we also observed four cases in which the relapsed isolate had lost some IS*6110* copies, present

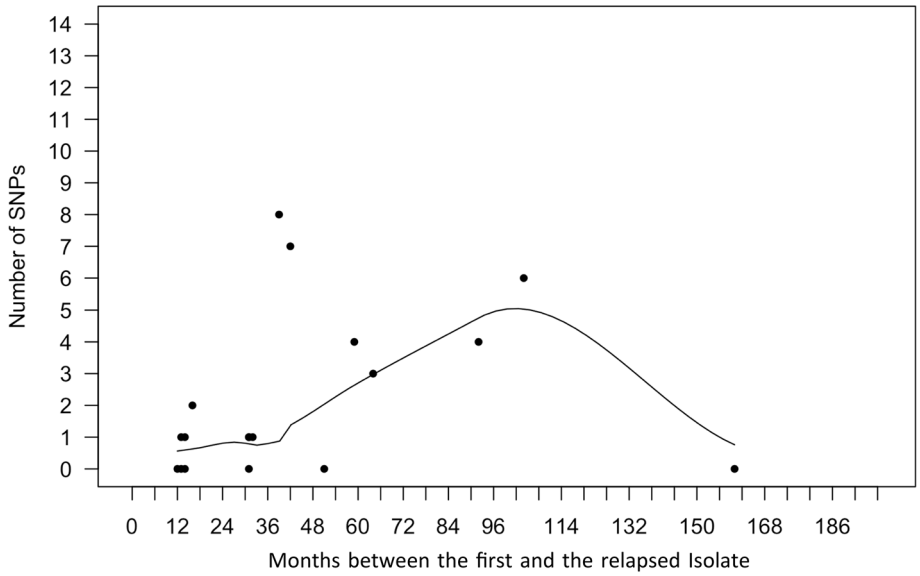


Figure 3. Scatter plot showing the number of SNPs developed between the first and the relapsed isolate of the patient versus time between the diagnosis of both isolates (in months). A Poisson regression model was used. A trend of increase in the number of SNPs is observed as the months of relapsing period increase, however, it was not significant (p value = 0.34).

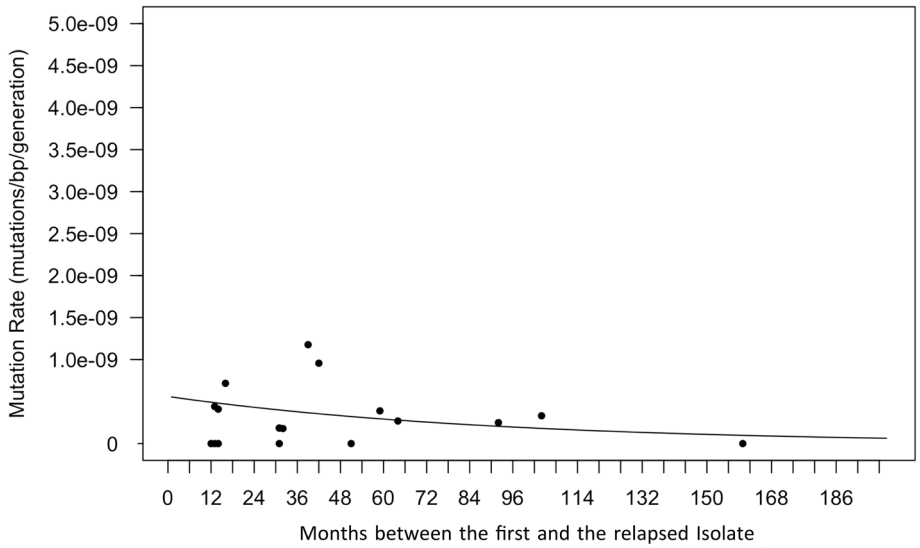


Figure 4. Mutation rate versus time until relapse. Scatter plot showing the number of SNPs that differed between the 18 patients' paired isolates (y-axis) as a function of the time between episodes (x-axis). The generation time is held constant at 18 h as seen in actively replicating *M. tuberculosis* in vitro. The relation between the mutation rate and the time between episodes was marginally significantly different from 0.0 (p value = 0.0613), indicating a trend of a higher mutation rate in the first months of relapsing period.

in the first isolate: P10 (one copy lost), P13 (three copies lost), P14 (one copy lost) and P15 (one copy lost). All these extra and absent copies usually had a lower number of reads than the fixed copies, indicating that they were not completely extended in the bacteria population. All these movements were observed in strains with more than 2 years between the isolates, except P14 pair (14 months). The IS*6110* locations can be found in Table S1.

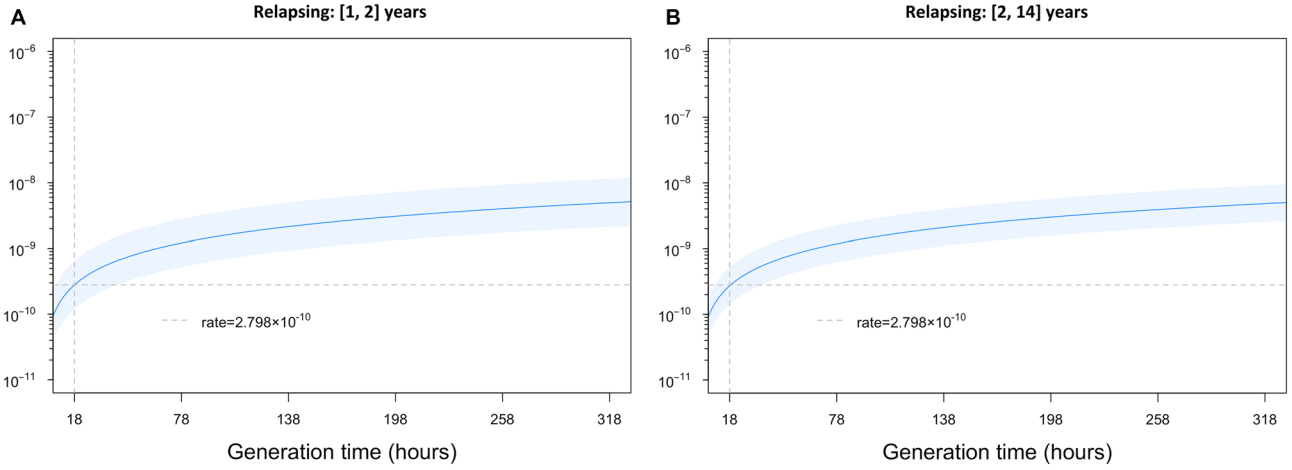


Figure 5. Changes in mutation rate during the relapsing period with varying generation times. Mutation rate (mutations per (bp × generation)) is shown for generation times ranging from 18 to 320 h for each pair grouped by the number of years between the first and relapsed isolates. The dark blue line is obtained from the regressions and shows the estimated mutation rate for a given generation time (x-axis) and light blue regions show 95% confidence intervals. The first panel shows the relationship between mutation rate and generation time during early relapsing (in ≤ 2 years) based on n = 6 pairs. The second panel shows the relation between mutation rate and generation time for relapse in 2–14 years based on n = 12 pairs. In both panels the grey dashed vertical line is fixed at 18 h, and the horizontal line indicates the mutation rate of 2.798×10^{-10} mutations per (bp × generation) as seen in early relapsing with generation times held constant at 18 h, fixed in both graphs for comparison availability.

Discussion

Studying *M. tuberculosis* latency in humans is harsh due to the difficulty of isolating the dormant bacteria, which is not possible until the active disease. Much has been published regarding latent TB and the percentages of reactivation and disease, but the latency data in patients who have already passed the disease have not been studied. Different approaches were used to mimic this process[10–12]. This work shows, for the first time, results obtained using isolates of patients with recurrent TB. Aragon, a region in the North of Spain, has a low incidence of TB. Thanks to the surveillance protocol carried out in this region since 2004, all *M. tuberculosis* isolates are genotyped and registered, allowing to trace the clinical TB history of the patients. Around 5% of the TB cases in our population correspond to recurrent TB. Of them, 89.8% were TB cases with isolates showing identical IS*6110*-RFLP patterns, indicating a potential relapse. Most of them (71%) were later considered as fail of treatment. In contrast, 10.2% of the patients had isolates with different genotypes, considered as reinfections. Among the total of TB cases in our community, reinfection occurs in 0.5% of the TB cases, reflecting that reinfection is uncommon among our population. These data are in agreement with a previous study in Madrid population, which showed an 87.5% of relapses and 12.5% of reinfections among the cases with recurrent TB[13]. However, in a study in the Canary Islands, the results showed a higher reinfection percentage (44%) versus the 55% of relapses[14]. A more extreme result was obtained in a study in London, in which 72.6% of the repeated patients were classified as reinfections against a 27.4% of relapses[15]. The large variation of the results among the different studies suggests that they largely depend on the population sample studied. It would be very interesting to analyse the reinfection cases in each of the studies to understand the reasons for these differences. Regarding endemic TB regions, a higher percentage of recurrent TB was found. Around 9.5% of TB patients had recurrent TB in Malawi (39.6% had relapse and 14.4% reinfection, the rest was undetermined)[16] and a study carried out in India demonstrated that the majority of relapses they had were among HIV negative people (95% of TB recurrences) while the majority of reinfections were among HIV positive people (75% of TB recurrences)[17].

Regarding the epidemiological and risk factors of the relapsed TB cases studied, we found that relapse was significantly earlier in HIV positive patients (in the first two years since the first episode) when compared to HIV negative patients (p value = 0.041), what would be in accordance with a compromised immune system. We also found a trend that males suffered relapse earlier than females, which could be linked to other risk factors such as the use of IV drugs, smoking and the HIV status, which were more frequent in males in our study population. Any risk factor was found as significant for causing an earlier reactivation by Colangeli et al. [12], however they recognized that the clinical cases studied did not have in general any comorbidity.

The number of SNPs between the pairs ranged from 0 to 8. Remarkably, three among the 18 pairs had more than 5 SNPs between the first and the relapsed isolate, interpreted as not recent transmission[18], even though the bacteria were isolated from the same patient. This could be related to clinical characteristics of the patients, as immunosuppression, HIV status or the treatment adherence. Surprisingly, several SNPs were found in the first isolates that were absent in the relapsed isolates, as if they had reverted. This phenomenon was extreme in P12, in which six out of the seven SNPs found were absent in the relapsed isolate. The explanation could be the presence of different clones in the patient[19,20]. In this way, in the different disease episodes a different clone was isolated, resulting from different bottlenecks and selective pressures of the original strain[21,22]. The reinfection

with an identical strain has been described as a limitation of these kind of studies, but in our case, it can be discarded as only one of the pairs belonged to a large endemic cluster (P4, with 0 SNPs). The rest of the pairs were infected with orphan or small-outbreak strains of up to four cases, differently from other studies with large endemic clusters and high TB prevalence[22].

Same as Colangeli et at.[12], we did not find a significant correlation between the number of SNPs and the time between episodes. However, it is possible that P8 (160 months between episodes and 0 SNPs) is altering the trend of SNP accumulation when the time between episodes increases. This is one limitation when working with small sample size, that a single point could have a great impact in the results. None of the SNPs found seemed related with recurrence as all were unique and therefore not common to more than one pair of isolates. It has been described that 0.5 SNPs per genome, per year is the standard mutation rate for *M. tuberculosis*[10]. Some studies, where multiple MDR/XDR isolates coming from the same patients were sequenced, have reported that selective pressure and antibiotic resistance can increase this mutation rate as high as more than 3 SNPs[17,21]. Despite all strains had been under the selective pressure of treatment, they did not achieve such a higher rate, maybe because they were drug susceptible. The mean mutation rate found in our study was 0.64 SNPs, slightly above the standard, due to the high mutation rate found in L4.1, almost double than the standard.

The correlation between the mutation rate and the relapsing period was found just marginally significant (p value = 0.0613), differently to Colangeli et al.[12], who found it significant. It is important to remark that the approaches were completely different: they used transmission events to mimic the latency period as the time between the diagnosis of the two cases, while we used isolates from the same patient who had a previous TB episode. We eliminated all patients with less than one year between the diagnosis of the episodes, as this was considered as a treatment failure, while Colangeli et al. 2020 had latency periods from one month, which was not possible in our clinical cases as a minimum of 6 months of treatment was required. We did not find a significant correlation between the mutation rate along the variable generation times analysed when we split the data into [1–2 years] and (2–14 years), we observed just a small difference. This difference was much smaller than that found by Colangeli et al. 2020 (as high as 8×10^{-10} for early latency), suggesting that mutation rate was constant during the relapsing period in recurrent TB cases. The mutation rate found in our study, 2.7×10^{-10}, was similar to that found by Ford et al. 2011 (2×10^{-10})[10], therefore both more distant from the one found by Colangeli et al. 2020. The reason why our results are similar to those of Ford et al. 2011 could be due to the similarity of the approaches applied, as they used lesions of the same macaques for studying latency and we used relapsed isolates from the same patients.

The analysis of the IS*6110* element showed differences in the number of IS*6110* copies in six of the pairs studied, affecting more than one IS copy in several pairs. It has been observed that IS*6110* transposed more in great starvation conditions[23], which could be similar to the conditions the mycobacteria found in the granuloma[4]. It was surprising that in four of the pairs studied, the relapsed isolates had lost 1 to 3 copies that were present in the first isolates. Noteworthy, the number of reads obtained in the fastQ files for these copies was considerably lower than for the rest of the IS copies. This suggests that those lost copies were not still fixed in the complete bacteria population, therefore a selection among the different clones present in the same patient had taken place[24]. It could be that the lost copies in the relapsed isolates had some deleterious effect for the mycobacteria as the relapsed bacteria were the ones without that IS copies. The fact that five out of the six pairs with IS*6110* movements had more than 2 years of relapsing period supports the idea of IS transposing more during the asymptomatic state of the patient[23].

The main limitation to analyse the evolution of the bacteria during the dormancy period is the approach used for resembling this state. There is not a perfect approach, as it is impossible to reproduce what is happening inside the granuloma of a concrete patient, but we think that using isolates of the same patient is the closest way to do it. The difficulty to obtain the complete epidemiological information of the patients is another limitation because it does not allow to determine the accurate development of the disease's episodes. Another limitation is that some of the SNPs could be the result of a sequencing error or due to laboratory management, what would have a huge impact on the mutation rate. In addition, although there were more cases of potential relapses in our records, DNA of the isolates was not available. We decided not to re-cultivate these stored isolates to avoid more manipulation that could introduce errors such as additional SNPs that were not present in the original strains.

As a conclusion, the patients with HIV seemed to suffer reactivation in the first two years after the initial episode of TB more frequently than HIV negative patients. Besides, IS*6110* movements occurred more frequently in patients with more than two years between episodes and it seems that different clones of the original strain could be responsible for the first and the following episodes. No correlation was found between the number of SNPs and the time between episodes, neither between the mutation rate and the relapsing period, just a trend of diminishing in longer time periods. Finally, the mutation rate seemed to be constant along all the period between episodes.

Material and methods

Selection of samples and patients. Of around 2553 cases of TB in Aragon since 2004, we first looked for those with more than one isolate more than 1 year apart and of a similar genotype. We used Bionumerics v6.7 software (v7.6, Applied Maths, Kortrijk, Belgium) to confirm that both isolates coming from the same patient shared an identical IS*6110*-RFLP pattern. Eighteen pairs of *M. tuberculosis* isolates with available DNA were included in this study. When there were more than two isolates from the same patient, the two more distant in isolation dates were considered for the evolution study during relapse. All data remained anonymous during the epidemiological search. Our regional ethical committee (Comité de Ética de la Investigación de la Comunidad Autónoma de Aragón, Record No. 20/2018) approved the methodology used in this work, detailed in 18/0336 project.

DNA of the bacterial isolates was obtained using the cetrimonium bromide method, as previously described[25]. No human DNA was sequenced. All DNA extractions were stored at − 20 °C until sequencing. All the isolates were genotyped by IS*6110*-RFLP and spoligotyping as previously described[26,27]. The genetic patterns obtained were stored and analysed in Bionumerics database software.

SNP annotation and lineage identification. Thirty-six isolates corresponding to 18 different patients were sequenced using Ion Torrent technology according to manufacturer's instructions. The fastQ files obtained were mapped against the reference *M. tuberculosis* strain H37Rv (NC_000962.3) in order to obtain the Binary Aligned Map (BAM) and Variant Call Format (VCF) files, used for the SNP study. The fastQ files were uploaded in Bionumerics software for the study and comparison of the genomes. The SNP annotation was carried out using Snippy software (default parameters) and Integrative Genomics Viewer (IGV), from the Broad Institute[28]. The effect of the mutation (synonymous or non-synonymous) was observed using Genewise platform (https://www.ebi.ac.uk/Tools/psa/genewise/). All the mutation points are referred to the H37Rv reference strain. For lineage identification, the SNP-based classification stablished by Coll et al. 2014[29] was used. This classification assigns specific SNPs to each TB lineages and sublineages.

IS*6110* location. All the reads containing the first and the last 30 base pairs of the IS*6110* sequence were extracted. These reads are formed by the beginning or the ending of the IS*6110* along with part of the gene in which the IS is inserted. After the extraction of the sequences, BLAST was made in Tuberculist and Bovilist to know the insertion point. BLAST was also made automatically with the script, but manual BLAST was required for some ambiguous points. The script used in R is in the Supplementary Materials.

Mutation rate calculation. The mutation rate per (bp × generation) was calculated as previously described[10], adjusting the parameters to our own data. Briefly, the mutation rate per bp × generation is defined as

$$\mu = \frac{\sum_{i=1}^{n} m_i}{N \sum_{i=1}^{n} \left(t_i / g\right)}$$

where μ is the mutation rate, m is the number of SNPs between the first and the relapsed isolate, N is the genome size (since we had, on average, reads covering 97.4% of the *M. tuberculosis* genome, N = 0.974 × L where L is reference genome size), t is time since infection (in hours, Table 1), and g is generation time (in hours).

Statistical methods. Poisson regression was used to model the variation of mutation rate over a range of generation times. To control the deviations from distributional assumptions a robust variance of Robust Sandwich Estimator was used. Poisson models were used to obtain mutation rates per (bp × generation) by using bp × generation as an offset. Two Poisson models were fit according to the relapsing period (one for 1–2 years including n = 6 pairs and another model for 2–14 years including n = 12 pairs). We also fitted a Poisson model using the relapsing period as a continuous independent variable. The hypothesis that we tested was if the parameter associated to relapsing period was significantly different from 0. To test the Poisson model parameters, a two-sided chi square test using the robust variance was used. Software R version 4.0.5 (2021-03-31) was used to all statistics analysis. Regression Poisson of all models was implemented in R using a generalized linear model function and robust variance control with sandwich package[30].

Data availability

The genomes of the studied isolates are loaded in GenBank with the accession numbers SAMN26037035-SAMN26037070 and the BioProject ID PRJNA808219, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA808219.

Received: 23 June 2022; Accepted: 22 September 2022
Published online: 06 October 2022

References

1. Esmail, H., Barry, C. E. 3rd., Young, D. B. & Wilkinson, R. J. The ongoing challenge of latent tuberculosis. *Philos. Trans. R. Soc. Lond. Ser. B, Biol. Sci.* **369**(1645), 20130437 (2014).
2. Getahun, H., Matteelli, A., Chaisson, R. E. & Raviglione, M. Latent *Mycobacterium tuberculosis* infection. *N Engl J Med.* **372**(22), 2127–2135 (2015).
3. Veatch, A. V. & Kaushal, D. Opening Pandora's Box: Mechanisms of *Mycobacterium tuberculosis* Resuscitation. *Trends Microbiol.* **26**(2), 145–157 (2018).
4. Gibson, S. E. R., Harrison, J. & Cox, J. A. G. Modelling a silent epidemic: a review of the in vitro models of latent tuberculosis. *Pathog (Basel, Switzerland).* **7**(4), 88 (2018).
5. Behr, M. A., Edelstein, P. H. & Ramakrishnan, L. Revisiting the timetable of tuberculosis. *BMJ* **362**, 1–10. https://doi.org/10.1136/bmj.k2738 (2018).
6. Weller, C. & Wu, M. A generation-time effect on the rate of molecular evolution in bacteria. *Evolution* **69**(3), 643–652 (2015).
7. Hershkovitz, I. *et al.* Detection and molecular characterization of 9,000-year-old *Mycobacterium tuberculosis* from a Neolithic settlement in the Eastern Mediterranean. *PLoS ONE* **3**(10), 3426 (2008).
8. Wirth, T. *et al.* Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog.* **4**(9), e1000160 (2008).
9. Arnold, C. Molecular evolution of *Mycobacterium tuberculosis*. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect Dis.* **13**(2), 120–128 (2007).
10. Ford, C. B. *et al.* Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat. Genet.* **43**(5), 482–486 (2011).

11. Colangeli, R. *et al.* Whole genome sequencing of *Mycobacterium tuberculosis* reveals slow growth and low mutation rates during latent infections in humans. *PLoS ONE* **9**(3), e91024 (2014).
12. Colangeli, R. *et al. Mycobacterium tuberculosis* progresses through two phases of latent infection in humans. *Nat. Commun.* **11**(1), 4870 (2020).
13. Cacho, J. *et al.* Recurrent tuberculosis from 1992 to 2004 in a metropolitan area. *Eur. Respir. J.* **30**(2), 333–337 (2007).
14. Caminero, J. A. *et al.* Exogenous reinfection with tuberculosis on a European island with a moderate incidence of disease. *Am. J. Respir. Crit. Care Med.* **163**(3 Pt 1), 717–720 (2001).
15. Afshar, B., Carless, J., Roche, A., Balasegaram, S. & Anderson, C. Surveillance of tuberculosis (TB) cases attributable to relapse or reinfection in London, 2002–2015. *PLoS ONE* **14**(2), e0211972 (2019).
16. Guerra-Assunção, J. A. *et al.* Recurrence due to relapse or reinfection with *Mycobacterium tuberculosis*: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. *J. Infect. Dis.* **211**(7), 1154–1163 (2015).
17. Shanmugam, S. *et al.* Whole genome sequencing based differentiation between re-infection and relapse in Indian patients with tuberculosis recurrence, with and without HIV co-infection. *Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis.* **113**(Suppl), S43–S47 (2021).
18. Lalor, M. K. *et al.* The use of whole-genome sequencing in cluster investigation of a multidrug-resistant tuberculosis outbreak. *Eur. Respir. J.* **51**(6), 1702313. https://doi.org/10.1183/13993003.02313-2017 (2018).
19. Gagneux, S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat. Rev. Microbiol.* **16**(4), 202–213 (2018).
20. Moreno-Molina, M. *et al.* Genomic analyses of *Mycobacterium tuberculosis* from human lung resections reveal a high frequency of polyclonal infections. *Nat. Commun.* **12**(1), 2716 (2021).
21. Xu, Y. *et al.* In vivo evolution of drug-resistant *Mycobacterium tuberculosis* in patients during long-term treatment. *BMC Genomics* **19**(1), 640. https://doi.org/10.1186/s12864-018-5010-5 (2018).
22. Pérez-Lago, L. *et al.* Recurrences of multidrug-resistant tuberculosis: Strains involved, within-host diversity, and fine-tuned allocation of reinfections. *Transbound. Emerg. Dis.* **69**(2), 327–336. https://doi.org/10.1111/tbed.13982 (2022).
23. Gonzalo-Asensio, J. *et al.* New insights into the transposition mechanisms of IS6110 and its dynamic distribution between *Mycobacterium tuberculosis* Complex lineages. *PLoS Genet.* **14**, e1007282 (2018).
24. Tanaka, M. M. Evidence for positive selection on *Mycobacterium tuberculosis* within patients. *BMC Evol Biol.* **4**, 1–8 (2004).
25. van Soolingen, D., de Haas, P. E., Hermans, P. W. & van Embden, J. D. DNA fingerprinting of *Mycobacterium tuberculosis*. *Methods Enzymol.* **235**, 196–205 (1994).
26. Van Embden, J. D. A. *et al.* Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: Recommendations for a standardized methodology. *J. Clin. Microbiol.* **31**, 406–409 (1993).
27. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907–914 (1997).
28. Robinson, J. T. *et al.* Integrative genome viewer. *Nat. Biotechnol.* **29**(1), 24–26 (2011).
29. Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* **5**, 4–8 (2014).
30. Zeileis, A., Köll, S., & Graham, N. Various versatile variances: an object-oriented implementation of clustered covariances in R. *J. Stat. Softw.* **95**, (1 SE-Articles), 1–36 (2020).

Acknowledgements

Authors would like thank Ainhoa Telletxea and Montserrat Gutierrez for proofreading the manuscript. We would like to thank the EPIMOLA group for supplying the genotyped bacterial DNA used in this work and to acknowledge the use of Servicio General de Apoyo a la Investigación-SAI, Universidad de Zaragoza (Servicio de Análisis Microbiológico), and Servicios Científico Técnicos, IACS (Servicio de Secuenciación y Genómica Funcional and Servicio de Biocomputación). This work was supported by the Carlos III Health Institute in the context of a Grant (FIS18/0336) co-funded by European Regional Development Fund/European Social Fund "A way to make Europe"/"Investing in your future" and J.C. was awarded a scholarship by the Government of Aragon/European Social Fund, "Building Europe from Aragon".

Author contributions

S.S. "conceptualization, funding acquisition, writing the manuscript". J.C. "laboratory work, analysis the data, writing the manuscript". A.C. "statistical analysis, biocomputational work, writing the manuscript". EPIMOLA "genotyping surveillance, epidemiological support".

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-21144-0.

Correspondence and requests for materials should be addressed to J.C.

Reprints and permissions information is available at www.nature.com/reprints.

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

Aragonese Working Group on Molecular Epidemiology of Tuberculosis (EPIMOLA)

María José Iglesias[3,4,5], Daniel Ibarz[5], Jesús Viñuelas[6], Luis Torres[7], Juan Sahagún[8], María Carmen Lafoz[5], Felipe Esteban de Juanas[9] & María Carmen Malo[9]

[4]CIBER de Enfermedades Respiratorias, Av. Monforte de Lemos, 3-5. Pabellón 11, Planta 0, 28029 Madrid, Spain. [5]Universidad de Zaragoza, Zaragoza, Spain. [6]Hospital Universitario Miguel Servet, Zaragoza, Spain. [7]Hospital General Universitario San Jorge, Huesca, Spain. [8]Hospital Universitario Lozano Blesa, Zaragoza, Spain. [9]Salud Pública, Gobierno de Aragón, Zaragoza, Spain.

En la publicación de este artículo hubo un error en la **Tabla 1**. Se muestra a continuación la **Tabla 1** correcta. Se ha enviado a la revista un comunicado para que hagan la corrección de errores. El p valor es diferente, pero sigue siendo significativo para el campo «HIV status». Sin embargo, el género ya no es marginalmente significativo.

Table 1. Risk factors of the 18 selected cases.

	[1–2 years]	**(2–14 years]**	**p value**
	N=6	*N=12*	
Gender:			0.316
Male	5 (83.3%)	6 (50.0%)	
Female	1 (16.7%)	6 (50.0%)	
HIV status:			0.035
No	1 (16.7%)	8 (80.0%)	
Yes	5 (83.3%)	2 (20.0%)	
Alcohol consumption:			1.000
High	1 (33.3%)	3 (37.5%)	
Low/No	2 (66.7%)	5 (62.5%)	
Immunosuppression:			0.010
No	0 (0.00%)	7 (87.5%)	
Yes	4 (100%)	1 (12.5%)	
Smoker:			0.491
No	2 (66.7%)	2 (25.0%)	
Yes	1 (33.3%)	6 (75.0%)	
Users of IV drugs:			0.152
No	1 (33.3%)	7 (87.5%)	
Yes	2 (66.7%)	1 (12.5%)	

Table S1. IS*6110* locations of all the isolates studied. For each location point, the number of reads found, the direct repeats, the gene and the direction of the IS*6110* is shown. In blue are the extra IS*6110* found among isolates of the same patient.

P1						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
1989080/1979921 (*M. bovis*)	23	cgc	1979923 (*M. bovis*)	26	*cut1*	Reverse
2366894	29	tag	2366896	30	*Rv2106c:PE22*	Reverse
3480373	24	cag	3480371	26	*Rv3113*	Forward
3083207	41	cagc	3083204	27	*Rv2775*	Forward
3665157 (*M. bovis*)	30	caa	3665159 (*M. bovis*)	30	*MT3426:MT3427*	Reverse
483310	40	tga	483308	35	*mmpS1*	Forward
2555929	29	tca	2555931	25	*Rv2282c:Rv2283*	Reverse
4077859	31	atc	4077861	43	*Rv3638:Rv3639c*	Reverse
3743341	23	gcgg	3743344	37	*Rv3346c*	Reverse
932202	31	aac	932204	31	*lpqQ:Rv0836c*	Reverse
1987457	31	-	1986625	22	*plcD/Rv1754c*	Reverse
2807873	32	aga	2807871	37	*Rv2492*	Forward
3493190	46	agg	3493192	44	*Rv3128c*	Reverse
3078617 (*M. bovis*)	32	-	3115687/3072239 (*M. bovis*)	44	*Rv2815c:Rv2816c/Rv2809*	Reverse
2047368	27	att	2047370	35	*Rv1804c:Rv1805c*	Reverse
1995894	36	tgc	1995892	23	*Rv1762c:Rv1763*	Forward

P2						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
1986626	28	tgttc	1986622	37	*Rv1754c*	Forward
3121879	18	ccc	3120523	15	*Rv2815c:Rv2816c*	Forward
3551130	48	aag	3551132	45	*Rv3183:Rv3184*	Reverse
31200025/3076507 (*M. bovis*)	21	gtc	3076507 (*M. bovis*)	15	*Rv2813:Rv2814c*	Reverse

2610863	26	gcc	2610861	35	*Rv2336*	Forward
1715972	26	acc	1715974	30	*mmpL12*	Reverse
3668575 (*M. bovis*)	20	-	3668756 (*M. bovis*)	19	*MT3429*	Forward
483296	31	agg	483298	30	*mmpS1*	Reverse
1075948	31	acc	1075950	26	*Rv0963c*	Reverse

P3						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
1986626	24	tgttc	1986622	29	*Rv1754c*	Forward
2365413	23	-	2366769	27	*Rv2104c:Rv2105/Rv2106:PE22*	Forward
3196168	31	acg	3196166	33	*Rv2885c:Rv2887*	Forward
3721237	30	cta	3721235	35	*Rv3334*	Forward
3668657 (*M. bovis*)	11	tt	3668656 (*M. bovis*)	17	*MT3429*	Forward
2247886 (*M. bovis*)	29	gcg	2247888 (*M. bovis*)	34	*MT2080*	Reverse
1998241	21	-	1998849	13	*Rv1765c/Rv1765c:Rv1765A*	Forward
3709622	19	gtc	3709624	31	*moaX*	Reverse
483296	29	agg	483298	24	*mmpS1*	Reverse
3121879	10	ccc	3120523	16	*Rv2815c/Rv2813:Rv2814c*	Reverse
2041817	16	tggcg	2041820	33	*PPE28:PPE29*	Reverse

P4						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
3606308	98	acg	3606310	127	*desA3*	Reverse
850539	171	tgtc	850536	106	*Rv0755A:thrV*	Forward
2627494	69	tca	2627492	106	*plcC*	Forward
2559568	135	gat	2559570	108	*Rv2286c*	Reverse
3129520	116	caa	3129522	131	*Rv2823c*	Reverse
1979901 (*M. bovis*)	84	-	1998483/1988923 (*M. bovis*)	107	*cut1/Rv1765c*	Forward

2633977	61	gcg	2633979	54	*PEE38*	Reverse
2604207 (*M. bovis*)	133	gaaa	2604210 (*M. bovis*)	87	*PPE71*	Reverse
1895651	90	ccta	1895654	88	*Rv1668c:Rv1669*	Reverse
889020	84	gagg	890376	131	*Rv0795-Rv0796*	Forward
3668725 (*M. bovis*)	42	gcc	3668723 (*M. bovis*)	53	*MT3429*	Forward
3121879	89	-	3119660	72	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse

P5						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
1986623	43	aac	1986625	37	*Rv1754c*	Reverse
4075916	52	gtt	4075914	37	*Rv3636*	Forward
3668572 (*M. bovis*)	27	-	3665159 (*M. bovis*)	34	*MT3429/MT3426:MT3427*	Reverse
2163404	33	ccag	2163401	25	*PPE34*	Forward
1997628/2262185	36	cgt	1997630/2262187	29	*Rv1765c/Rv2015c*	Reverse
481422	37	gtc	481420	34	*mmpL1*	Forward
1992739	29	cgg	1992737	32	*wag22:Rv1760*	Forward
2608664	27	gcg	2608662	44	*Rv2333c:cysk1*	Forward
2010922	38	gag	2010924	49	*cyp144*	Reverse
1987018	40	ccag	1987015	33	*plcD*	Forward
2633843	129	ctc	2633841	84	*PPE38*	Forward
3480373	42	cag	3480371	42	*Rv3113*	Forward
932202	48	aac	932204	38	*lpqQ:Rv0836c*	Reverse
1481531	35	ggc	1481529	48	*Rv1319c*	Forward
2630024 (CDC1551)	60	atc	2630022 (CDC1551)	107	*MT2419:MT2420*	Forward
3078617 (*M. bovis*)	26	-	3120523/3077003 (*M. bovis*)	28	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
466508	44	gtc	466506	32	*pks6*	Forward
3709458	33	agt	3709456	36	*moaX*	Forward
2633765	35	gtcac	2633761	49	*PPE38*	Forward

1989080/1979921 (*M. bovis*)	21	cgc	1979923 (*M. bovis*)	37	*cut1*	Reverse

P6						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
3494370	31	ctt	3494372	43	*Rv3128c:Rv3129c*	Reverse
1695262	50	caa	1695260	68	*Rv1504c:Rv1505c*	Forward
987098	46	tccc	987095	35	*Rv0887:Rv0888*	Forward
2604207 (*M. bovis*)	52	gaaa	2604210	56	*PPE71*	Reverse
2682339	32	gca	2682341	28	*hemN*	Reverse
2039009	43	-	-	-	*Rv1798:lppT*	-
1895651	44	ccta	1895654	33	*Rv1668c:Rv1669*	Reverse
2198331	20	ggac	2198328	49	*Rv1947*	Forward
889020	35	gagg	890376	57	*Rv0795-Rv0796*	Reverse
1987546	30	-	1986937 (*M. bovis*)	47	*plcD/Rv1762c:Rv1765c*	Forward
1527045	24	tgg	1527043	52	*Rv1358*	Forward
3668725 (*M. bovis*)	11	gcc	3668723 (*M. bovis*)	26	*MT3429*	Forward
3121879	42	ccc	3120523	50	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse

P7						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
3491909	35	cca	3491907	38	*PPE49:Rv3126c*	Forward
47041	55	ccg	47039	46	*Rv0042c*	Forward
1998793	104	gtg	1998795	41	*Rv1765c:Rv1765A*	Reverse
1075948	60	acc	1075950	38	*Rv0963c*	Reverse
483296	51	agg	483298	48	*mmpS1*	Reverse
891081/3712442	49	cag	891083/3712444	53	*IS1547*	Reverse
2610863	44	gcc	2610861	70	*Rv2336*	Forward
3668575 (*M. bovis*)	26	-	3668756 (*M. bovis*)	19	*MT3429*	Forward

2634049	37	gat	2634051	10	*PPE38*	Reverse
3126538	60	-	3076507	62	*Rv2819c/Rv2813:Rv2814c*	Reverse
1986626	77	tgttc	1986622	56	*Rv1754c*	Forward
3709336	40	aag	3709338	57	*moaX*	Reverse
2039130	2	att	2039132	2	*Rv1798:lppT*	Reverse
3554124	5	gat	3554126	6	*Rv3187:Rv3188*	Reverse

P8						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
2610863	76	gcc	2610861	78	*Rv2336*	Forward
2633680	44	-	2627272	79	*PPE38/plcC*	Reverse
1986626	91	tgttc	1986622	104	*Rv1754c*	Forward
2626655	89	cct	2626653	72	*Rv2348c*	Forward
2164719	82	caag	2164716	81	*PPE34*	Forward
483296	93	agg	483298	99	*mmpS1*	Reverse
1987085	92	ctg	1987083	82	*plcD*	Forward
3555250	111	gcgc	3555247	78	*Rv3189*	Forward
3709624/3664146 (*M. bovis*)	113	-	3668756 (*M. bovis*)	81	*moaX/MT3429*	Forward
1900789	98	ttg	1900791	89	*Rv1675c*	Reverse
444616	85	agt	444614	53	*Rv0366c*	Forward
1075948	88	acc	1075950	81	*Rv0963c*	Reverse
2547732	78	cat	2541730	92	*Rv2267c*	Forward

P9						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
1587	12	cac	1585	30	*dnaA:dnaN*	Forward
1075948	20	acc	1075950	22	*Rv0963c*	Reverse
3121879/3078341 (*M. bovis*)	25	-	3076507 (*M. bovis*)	18	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse

3848022	20	cga	3848020	25	*Rv3430c*	Forward
1986626	27	tgttc	1986622	24	*Rv1754c*	Forward
1915216	17	gcg	1915214	28	*tyrS:lprJ*	Forward
890376	26	tca	889014	23	*Rv0795-Rv0796*	Forward
3021702	25	acgt	3021699	15	*Rv2708c*	Forward
483296	20	agg	483298	29	*mmpS1*	Reverse
2610863	17	gcc	2610861	25	*Rv2336*	Forward
3668575 (*M. bovis*)	16	-	3668756 (*M. bovis*)	25	*MT3429*	Forward

P10						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
2807873	50	aga	2807871	57	*Rv2492*	Forward
3480373	65	cag	3480371	64	*Rv3113*	Forward
2366894	61	tag	2366896	74	*Rv2106:PE22*	Reverse
2047368	98	att	2047370	64	*Rv1804c:Rv1805c*	Reverse
3083207	104	cagc	3083204	57	*Rv2775*	Forward
1998688	62	gt	1998687	36	*Rv1765c:Rv1765A*	Forward
1987457	75	-	1986625	75	*plcD/Rv1754c*	Reverse
4077859	79	atc	4077861	59	*Rv3638:Rv3639c*	Reverse
1989080/1979921 (*M. bovis*)	49	cgc	1979923 (*M. bovis*)	64	*cut1*	Reverse
3167691	69	gtc	3167689	62	*nicT*	Forward
3665157 (*M. bovis*)	56	caa	3665159 (*M. bovis*)	60	*MT3426:MT3427*	Reverse
1733206	65	cgg	1733204	60	*Rv1532c*	Forward
2555929	53	tca	2555931	53	*Rv2282c:Rv2283*	Reverse
932202	60	aac	932204	62	*lpqQ:Rv0836c*	Reverse
2245246	63	gtg	2245244	52	*Rv2000*	Forward
889020	56	tga	890376	63	*Rv0795-Rv0796*	Reverse
3078617 (*M. bovis*)	57	-	3120523/3077003 (*M. bovis*)	33	*Rv2815c*	Reverse

1989192	16	tctg	1989195	12	*cut1*	Reverse

P11						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
2610863	65	gcc	2610861	53	*Rv2336*	Forward
483296	64	agg	483298	59	*mmpS1*	Reverse
1986626	74	tgttc	1986622	64	*Rv1754c*	Forward
891081/3712442	51	cag	891083/3712444	61	*IS1547*	Reverse
3115221	62	cag	3115223	54	*Rv2808*	Reverse
1998793	54	gtg	1998795	40	*Rv1765c:Rv1765A*	Reverse
1075948	52	acc	1075950	51	*Rv0963c*	Reverse
3126529	67	cgtc	3126532	43	*Rv2818c:Rv2819c*	Reverse
2163390	17	tta	2163461	35	*PPE34*	Reverse
3668756 (*M. bovis*)	65	-	3709338	63	*MT3429/moaX*	Reverse
1987459	61	aac	1987457	46	*plcD*	Forward
4222477	68	cca	4222475	37	*serU*	Forward
1999527	30	cca	1999529	40	*Rv1765A:Rv1766*	Reverse
3121879/3078359 (*M. bovis*)	38	-	3076507 (*M. bovis*)	34	*Rv2815c:Rv1816c/Rv2813:Rv2814c*	Reverse
2629973 (CDC1551)	77	-	2634051/2629896 (CDC1551)	18	*MT2419:MT2420/PPE38 (MT2419)*	Reverse
1998474/2263031	56	cggcc	1998470/2263027	60	*Rv1765c/Rv2015c*	Forward

P12						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
483296	62	agg	483298	64	*mmpS1*	Reverse
1986626	45	tgttc	1986622	48	*Rv1754c*	Forward
2246937	47	ttg	2246939	33	*Rv2001*	Reverse
1075948	62	acc	1075950	39	*Rv0963c*	Reverse
2610863	53	gcc	2610861	46	*Rv2336*	Forward

3668575 (*M. bovis*)	41	-	3668756 (*M. bovis*)	41	*MT3429*	Forward
1715972	40	acc	1715974	43	*mmpL12*	Reverse
3121879/3078359 (*M. bovis*)	46	-	3076507 (*M. bovis*)	35	*Rv2815c:Rv1816c/Rv2813:Rv2814c*	Reverse
1543460	3	cgc	1543462	1	*Rv1371*	Reverse

P13						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
3480373	70	cag	3480371	83	*Rv3113*	Forward
3665054 (*M. bovis*)	66	cax	3665052	13	*MT3426*	Forward
2633843	65	-	2604185 (*M. bovis*)	57	*PPE38/PPE71*	Forward
3668572 (*M. bovis*)	39	-	3665159 (*M. bovis*)	18	*MT3429/MT3426:MT3427*	Reverse
1997628/2262185	68	cgt	1997630/2262187	69	*Rv1765c/Rv2015c*	Reverse
1975227 (*M. bovis*)	59	-	1995739/1986578 (*M. bovis*)	63	*MT1800:MT1801/Rv1762c*	Forward
1481531	41	ggc	1481529	68	*Rv1319c*	Forward
3551220	66	gac	3552585	72	*Rv3183:Rv3184/Rv3185:Rv3186*	Reverse
3071516	78	ctgg	3071513	59	*Rv2760c*	Forward
2010922	74	gag	2010924	79	*cyp144*	Reverse
1986623	56	aac	1986625	86	*Rv1754c*	Reverse
3668756 (*M. bovis*)	82	cgg	3668758 (*M. bovis*)	141	*MT3429*	Reverse
932202	56	aac	932204	58	*lpqQ:Rv0836c*	Reverse
3123115	47	gag	3123117	49	*Rv2815c:Rv2816c*	Reverse
1357287	30	ttgg	1357284	67	*glgC:PE14*	Forward
2614559	26	gcc	2614561	11	*moeW:mmpL9*	Forward
3275792	12	ggt	3275794	16	*papA5*	Forward
3114264	1	tga	3114262	3	*Rv2807*	Reverse

P14						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**

3302967	57	gtg	3302969	44	*fadD29:Rv2951c*	Reverse
4323409	64	tgc	4323407	78	*Rv3847:Rv3848*	Forward
2581014	12	cca	2581017	34	*Rv2308*	Reverse
2010922	47	gag	2010924	31	*cyp144*	Reverse
3379505	36	gtt	3379507	34	*PPE47*	Reverse
2633843	48	ctc	2633841	28	*PPE38*	Forward
3595154	45	gac	3595152	57	*Rv3218*	Forward
1947016	41	acag	1947013	48	*proT:Rv1720c*	Forward
1481531	58	ggc	1481529	48	*Rv1319c*	Forward
1997628/2262185	64	cgt	1997630/2262187	53	*Rv1765c/Rv2015c*	Reverse
1989080/1979921 (*M. bovis*)	51	cgc	1979923 (*M. bovis*)	60	*cut1*	Reverse
1986623	45	aac	1986625	55	*Rv1754c*	Reverse
3480373	75	cag	3480371	51	*Rv3113*	Forward
932202	48	aac	932204	31	*lpqQ:Rv0836c*	Reverse
3668836 (*M. bovis*)	63	-	3668756 (*M. bovis*)	48	*MT3429:IS1547/MT3429*	Reverse
3668836 (*M. bovis*)	63	-	3665159 (*M. bovis*)	30	*MT3329:IS1547/MT3426:MT3427*	Reverse
2263133	79	ccg	2263131	41	*Rv2015c:Rv2016*	Forward
3078617 (*M. bovis*)	46	-	3120523/3077003 (*M. bovis*)	20	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
3621245	1	gacc	3621242	1	*Rv3241c*	Forward

P15						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
3491532	57	cac	3491530	68	*PPE49*	Forward
3378894	76	gcag	3378897	56	*esxR*	Reverse
3668756 (*M. bovis*)	82	-	3710381/3664903 (*M. bovis*)	69	*MT3429 /Rv3324A*	Reverse
2039176	68	aag	2039174	54	*lppT*	Forward
2203402	63	tgtc	2203405	34	*Rv1958c*	Reverse
2265111	71	tcg	2265109	55	*Rv2017c:Rv2018*	Forward

1987563	73	-	1987291 (*M. bovis*)	57	*plcD/Rv1762c:Rv1763*	Forward
889020	70	gagg	890376	75	*Rv0795-Rv0796*	Forward
3115758	67	cagc	3115761	100	*Rv2809:Rv2811*	Reverse
1703102	87	cac	1703100	53	*gmdA*	Forward
1998278/2262835	49	cca	1998280/2262837	57	*Rv1765c/Rv2015c*	Reverse
1357080	23	gat	1357082	59	*glgC:PE14*	Reverse
3121879	40	ccc	3120523	38	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
2559602	62	aca	2559600	46	*Rv2286c:yjcE*	Forward

P16						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
1075023	50	ctta	1075026	61	*lprP*	Reverse
1901096	55	gcc	1901094	41	*Rv1676*	Forward
2555849	52	gac	2555851	46	*Rv2282c*	Reverse
2248086 (*M. bovis*)	46	gtc	2248084 (*M. bovis*)	55	*MT2080:MT2081*	Forward
888900	46	atac	888903	47	*Rv0794c:Rv0795*	Reverse
3121879	42	ccc	3120523	32	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
1982069	18	ccc	1982071	2	*PPE24*	Reverse

P17						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
2604207 (*M. bovis*)	115	gaaa	2604210 (*M. bovis*)	76	*PPE71*	Reverse
3548920	127	atc	3548918	118	*Rv3179:Rv3180c*	Forward
1987569/1972961 (*M. bovis*)	121	-	1973342 (*M. bovis*)	114	*plcD*	Forward
3668725 (*M. bovis*)	42	gcc	3668723 (*M. bovis*)	34	*MT3429*	Forward
2270392	137	tga	2270394	128	*Rv2024c:Rv2025c*	Reverse
889020	114	gagg	890376	109	*Rv0795-Rv0796*	Reverse
1998838	107	-	1979901 (*M. bovis*)	111	*Rv1765c:Rv1765A/cut1*	Reverse

1895651	100	ccta	1895654	47	*Rv1668c:Rv1669*	Reverse
3121879	78	ccc	3120523	64	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse

P18						
pre-IS point	**Number of reads**	**Direct repeat**	**post-IS point**	**Number of reads**	**Gene**	**Direction**
483296	65	agg	483298	46	*mmpS1*	Reverse
3121879	34	ccc	3120523	27	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse

Publicación 4

Infection, Genetics and Evolution 81 (2020) 104184

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Research paper

Investigation of a rapidly spreading tuberculosis outbreak using whole-genome sequencing

Jessica Comin[a], Armando Chaure[b], Alberto Cebollada[a], Daniel Ibarz[c], Jesús Viñuelas[d,h], María Asunción Vitoria[e,h], María José Iglesias[c,f,g], Sofía Samper[a,f,g,*]

[a] Instituto Aragonés de Ciencias de la Salud, Zaragoza, Spain
[b] Salud Pública, Aragón, Spain
[c] Universidad de Zaragoza, Zaragoza, Spain
[d] Hospital Universitario Miguel Servet, Zaragoza, Spain
[e] Hospital Clínico Universitario Lozano Blesa, Zaragoza, Spain
[f] CIBER de enfermedades respiratorias, Madrid, Spain
[g] Fundación IIS Aragón, Zaragoza, Spain
[h] Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica, Madrid, Spain.

ARTICLE INFO

Keywords:
Outbreak
Tuberculosis epidemiology
Genetic polymorphism
Molecular epidemiology

ABSTRACT

This paper describes the application of whole-genome sequencing (WGS) to investigate an outbreak of *Mycobacterium tuberculosis* occurring in Aragon, Spain, where strains have been submitted to genotyping since 2004. The responsible outbreak strain appeared in our region first in 2014 and it spread to 14 patients in the following three years.

WGS found low variability between the isolates with none of the SNPs differences detected more than once, all of which were attributed to a recent transmission. Although two ambiguous bases linked two cases with those who presented the SNP in the same position, the establishment of a definitive transmission route was not possible. The epidemiological data supported the existence of a super-spreader, probably responsible for the majority of the cases involved since there was a two-year delay in diagnoses among cases. This fact would also help explaining the low variability found. The index case was not identified, possibly because it was not diagnosed in Aragon. In addition WGS characterised the strain as a Linage 4.3.3/LAM family and corroborated the susceptibility to anti-tuberculosis drugs observed by the clinical laboratories.

This work shows the need to have epidemiological data to support the genomic data in order to clarify the evolution of tuberculosis outbreaks.

1. Introduction

Tuberculosis (TB) is still an important cause of death worldwide (1.5 million people died and 10 million new cases in 2018 (WHO, 2019)). Aragon applies genotyping to all *Mycobacterium tuberculosis* (MTB) complex isolates for surveillance purposes routinely since 2004. Great efforts have been made to develop molecular techniques to assist with epidemiological investigations. Genotyping based in Restriction Fragment Length Polymorphism (RFLP)-IS*6110* has demonstrated its value to determine tuberculosis transmission dynamics in previous studies (Borgdorff et al., 2000; Gonzalo-Asensio et al., 2018). In recent years whole-genome sequencing (WGS) has become an affordable technology capable of replacing previous typing methodologies (Cirillo et al., 2016), and has proven invaluable to fully elucidate transmission routes when applied retrospectively (Walker et al., 2014). Although it remains to be fully agreed upon, a strain is considered to belong to a cluster when it differs in twelve or less SNPs from the major clone; likewise a transmission is considered to be recent if there are five or less SNPs between isolates (Lalor et al., 2018). A previous study showed that clusters where the notifications of the first two cases were less than 90 days apart were likely to develop into a large cluster (Hamblion et al., 2016).

We identified a rapidly spreading outbreak of TB in Aragon, that affected fourteen cases in four years. We decided to apply WGS in order to elucidate the transmission routes of the outbreak and to describe the molecular characteristics of the causal strain.

* Corresponding author at: Laboratorio de Investigación Molecular- UIT, IACS-Hospital Universitario Miguel Servet, P Isabel la Católica 1-3, CP 50009 Zaragoza, Aragón, Spain.
E-mail address: ssamper.iacs@aragon.es (S. Samper).

https://doi.org/10.1016/j.meegid.2020.104184
Received 9 September 2019; Received in revised form 17 December 2019; Accepted 6 January 2020
Available online 10 January 2020

Table 1
Primers used for the amplification of RvD5 and RD152 regions.

Primer name	Sequence	Reference
Rv1754cr	GAACCATGAGTCCAATAGCGGC	(Alonso et al., 2011)
C2	AAACACTGCGGCCTGCTCG	(Ho et al., 2000)
MT3429-F	GCAATCAGAACGTCGGTGT	(Millan-Lou et al., 2013)
RvDR5-R	GTACCCGCACCACCTGCT	(Millan-Lou et al., 2013)

2. Materials and methods

Patients and clinical samples: Since 2004, the positive TB cultures in the microbiological services of the hospitals in our region are collected and genotyped. As a result the patterns of more than two thousand MTB strains are in our Bionumerics Database for analysis and comparison of new genotypes. In 2017, we observed an identical genotype for thirteen patients, which appeared for the first time in 2014. As the IS*6110*-RFLP, Spoligotyping and Mycobacterial Interspersed Repetitive Units-Variable Number of Tandem Repeats (MIRU-VNTR) patterns were the same for the 13 isolates, we assumed they were infected by the same strain. We had one isolate from 2014, three from 2015, seven from 2016 and three more from 2017. One case, Case 14, was diagnosed by a commercial test that resulted positive for sputum but its culture was not available.

Epidemiological contact tracing: Exhaustive contact studies for Case 4 and Case 5 were made. By using the tuberculin skin test the possible contacts in the workplace, at the school and in the family environments were investigated in order to identify those who had been infected and try to control the outbreak.

Standard molecular typing: DNA of the isolates was obtained using the cetrimonium bromide method previously described (van Soolingen et al., 1994). The standardised techniques IS*6110*-RFLP, Spoligotyping and MIRU-VNTR were performed as previously described (Van Embden et al., 1993; Kamerbeek et al., 1997; Supply et al., 2006). The genetic patterns obtained were analysed by Bionumerics software v7.6 (Applied Maths, Kortrijk, Belgium) and stored in the Bionumerics Database. Spoligotypes were systematically revised against the SITVIT database (www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE).

IonTorrent sequencing: The library was prepared using Ion Xpress™ Plus Fragment Library Kit following the manufacturer instructions. The library and the sequences were mixed in equimolar quantities, the template was prepared and the Ion Chip 530 was loaded in the Ion S5™ XL System, where the sequencing and the analysis took place. Once the sequences were obtained, they were mapped against the reference strain H37Rv (NC_000962.3) or CDC1551 (NC_002755.2) using the Torrent Mapper TMAP version 5.12.27. Variants were called, excluding high variable regions as *ppe* and *pe* genes, by the Torrent Suite plugin variantCaller. This allowed us to obtain a bam file to use for the Integrative Genome Viewer (IGV) program.

Lineage and family identification: We used the Coll et al. SNP classification (Coll et al., 2014), in which each MTB linage and family is associated with specific SNPs in different genomic regions. It allows a fast and easy classification. The SNPs that classified the studied strain as linage 4.3.3 and LAM family are described in the Results chapter.

Bioinformatic tools for the study of the genomes: The fastQ files of the sequences obtained were introduced and analysed in Bionumerics software v7.6 (Applied Maths, Kortrijk, Belgium). The programme mapped the sequences against the reference strain H37Rv and obtained the different SNPs among the genomes analysed. The SNPs analysis allowed the construction of the dendrogram using the UPGMA method. For a greater accuracy, a strict-SNP filtering which removed positions with at least one ambiguous or unreliable base, gaps, non-discriminatory positions and the *ppe* and *pgrs* genes was applied. It was also considered that the retained SNP positions had a minimum 5× coverage and that the minimum distance between SNPs was at least 12 base pairs (bp). To study the phylogeny of the isolates, their drug-resistance genomic profiles and the differential regions (RD) (Coll et al., 2014; Rindi et al., 2014), IGV program and Tuberculist website (http://genolist.pasteur.fr/TubercuList/) were used.

PCR and re-sequencing: Standard PCR (MyTaq DNA polymerase and 5× MyTaq Reaction Buffer, Bioline) and Sanger sequencing were used to study RvD5 and RD152 regions. Initial denaturation at 95 °C for 1′, followed by 35 cycles: denaturation at 95 °C for 15″, annealing at primer temperature for 15″ and extension at 72 °C for 30″. The primers used are detailed in Table 1.

3. Results

The genotype of the studied outbreak strain was first identified in 2014 as its Spoligotype (777674077560771) had no match in the SITVIT database. Thirteen more isolates from different cases with identical IS*6110*-RFLP and Spoligotype were detected in the following three years using our Bionumerics Database. Two of the isolates were also typed by 24-MIRU-VNTR and found to be identical (Fig. 1). To confirm transmission among the cases we decided to investigate the patients' epidemiological characteristics and to carry out WGS in all isolates.

3.1. Epidemiological study of the outbreak

Epidemiological data and their corresponding date of onset of symptoms and date of diagnosis of the fourteen cases of the studied outbreak is summarised in Table 2. Cases 2, 4, 5, 11 and 13 were considered infectious as they presented a positive (+) sputum baciloscopy (BK+). Furthermore, Case 5 was undiagnosed for two years and therefore remained untreated. He is considered to be the super-spreader, and most of the cases were related to this case (Fig. 2). A thorough epidemiological study was initiated after the diagnosis of this case as a result. No epidemiological relationship was found among Case 1, 2, 10 and 11 and other cases.

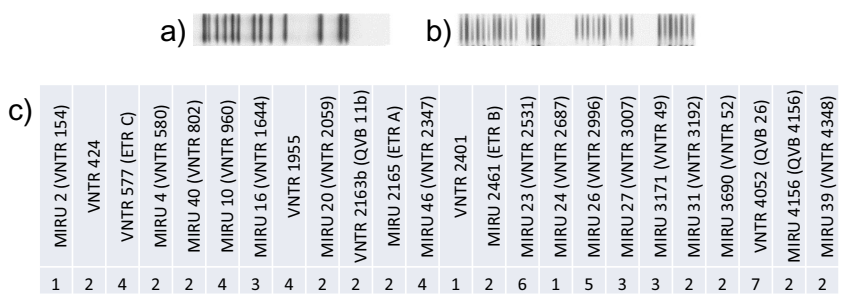


Fig. 1. a) IS*6110*-RFLP, b) Spoligotype and c) 24-MIRU-VNTR patterns of the studied strain. The RFLP and Spolygotype patterns were identical for the thirteen isolates. The 24-MIRU-VNTR was also identical for the two isolates typed by this method.

Table 2
Summary of the epidemiological data of the fourteen cases of the studied outbreak.

Case	Diagnosis Date	Symptoms	Baciloscopy	Age	Sex	Birth place	Epidemiological links	Comments
1	January 2014	Weight loss since 2013	–	9	Female	Spain	Unknown	Gambian father. Digestive TB
2	January 2015	January 2014	+	27	Male	Spain	Unknown	
3	July 2015	October 2015	–	21	Male	Spain	Football trainer of case 5	
4	December 2015	October 2015	+	41	Female	Nicaragua	Aunt of cases 5, 7 and 14	Resident of Spain for previous 9years. Travelled to Nicaragua in 2013
5	March 2016	Beginning of 2014	+	21	Male	Nicaragua	Son of case 6. Brother of case 7 and 14	
6	April 2016	Asymptomatic	–	43	Male	Nicaragua	Father of cases 5, 7 and 14	
7	April 2016	Asymptomatic	–	16	Male	Nicaragua	Son of case 6. Brother of case 5 and 14	
8	May 2016	January 2016?	–	21	Female	Spain	Classmate of case 5	
9	May 2016	January 2016?	–	25	Male	Poland	Classmate of case 5	
10	June 2016	May 2016	–	48	Female	Spain	Unknown	
11	January 2017	December 2016	+	20	Male	Spain	Unknown	
12	January 2017	December 2016	–	19	Female	Nicaragua	Probable contact with case 5 and his family	Attended Nicaraguan social meetings
13	December 2017	November 2017	+	52	Male	Portugal	Friend of case 4	In 2015 appeared in the contact study of Case 4 and was found skin test positive but rejected the prophylactic treatment
14	April 2016	Asymptomatic	–	21	Male	Nicaragua	Son of case 6. Brother of case 5 and 7	No culture grown. DNA + in the sputum.

Fig. 2. Epidemiological links: Case 5 was a super-spreader and infected the majority of the cases included two, Cases 3 and 4, diagnosed before Case 5, and which were his football trainer and his aunt. Case 5 also infected his father (Case 6), his two brothers (Case 7 and 14) and two classmates (Cases 8 and 9). The epidemiological link between Case 5 and Case 12 was not strong, which is indicated by a dashed line. Epidemiological links were not discovered for Cases 1, 2, 10 and 11 (R: relative, RC: relative and co-habitant).

3.2. SNPs analysis by WGS

A total of 13 SNPs were observed in the characterised isolates (Table 3). A low variability between the isolates, between 0 and 7 SNPs, was found (Table 4). Any of the SNPs was present in more than one isolate, which prevented us to gather data to establish the transmission chain. In order to find more differences between isolates, a direct review of the SNPs without the strict-SNP filtering was performed using the Bionumerics software. This revealed two positions with an ambiguous base, an indication of an incomplete SNP acquisition. There was one isolate with the ambiguous base, one with the SNP completely displayed while the rest of isolates did not differ from the reference. The two SNPs were at positions 1378, within *dnaA* gene (C/A/M, non-synonymous mutation), and 2,085,555, within *glcB* gene (G/A/R, non-synonymous mutation).

The phylogeny tree based in the different SNPs showed six identical isolates (Fig. 3), five of them linked epidemiologically. Thus, genomic and epidemiological data agreed in these cases. Two further connections were possible by analysing the SNPs with the ambiguous base, between Case 10, for which epidemiological links to any other cases of the outbreak could not be established, and Case 4 (2,085,555, R/A), and between Case 5 and Case 12 (1378, M/A), which were epidemiologically linked. Contrary to expected, the isolates from family related cases (Cases 5, 6 and 7), did not group together.

3.3. Genomic strain characterisation

3.3.1. The linage and resistance genotype

WGS permitted the genomic characterisation of the outbreak strain. The SNP variants determined that this strain belongs to Linage 4.3.3 (LAM family) since it has the specific mutations in *ipqQ* (codon 57; T/C), *rpoC* (codon 542; C/G) and *Rv0338c* (codon 826; G/A) genes (Coll et al., 2014). We also checked the specific RD regions in LAM family, and we determined that this strain has the RD115 deleted, but has the RD761 and the RD174 present, which corroborated it is part of the LAM family.

Regarding the genes attributed to be related to drug resistance, they presented mutations as follows: *rpoC* gene in position 764,995 (G/A, synonymous mutation); *Rv0338c* gene in position 403,364 (G/A, synonymous mutation); *kasA* gene in position 2,518,919 (G/A, non-synonymous mutation) and *gyrA* gene in position 8040 (A/C, non-synonymous mutation) (Coll et al., 2014). None of these SNPs had a high-confident association to resistance what agreed with the results of the antibiogram of the clinical laboratories, therefore we conclude that the strain was susceptible to first line antibiotics.

Table 3
Description of the thirteen SNPs found in the outbreak isolates.

Position in reference genome H37Rv	Affected Gene	Nucleotide change	Mutation effect
4943	*Rv0004*	G/T	non-synonymous mutation
682,452	*Rv0585c*	T/C	non-synonymous mutation
853,659	*phoR*	G/A	non-synonymous mutation
1,009,148	*Rv0906*	G/T	non-synonymous mutation
1,085,321	*accA2*	C/T	non-synonymous mutation
1,598,753	*whiA*	C/T	non-synonymous mutation
1,924,619	*pyrG*	A/G	non-synonymous mutation
2,068,108	*secA2*	A/C	non-synonymous mutation
3,229,502	*glnD*	C/G	non-synonymous mutation
4,031,024	*mutY*	G/A	non-synonymous mutation
4,076,216	Intergenic region	C/T	
4,338,459	*whiB6*	C/T	non-synonymous mutation
4,338,599	Intergenic region	T/G	

Table 4
Number of SNPs in the isolates and epidemiological links among the thirteen patients. The upper half of the table shows if there is a known epidemiological link between cases. The lower half of the table shows the number of SNPs between the isolates. All of them are between 0 and 7 SNPs, which means a recent contact. Case 5, the super-spreader, is coloured in blue.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12	Case 13
Case 1		no	no	no	no	no	no	no	no	no	no	no	no
Case 2	3		no	no	no	no	no	no	no	no	no	no	no
Case 3	5	4		no	yes	no	no	no	no	no	no	no	no
Case 4	2	1	3		yes	yes	yes	no	no	no	no	no	yes
Case 5	2	1	3	0		yes	yes	yes	yes	no	no	probable	no
Case 6	3	2	4	1	1		yes	no	no	no	no	probable	no
Case 7	6	5	7	4	4	5		no	no	no	no	probable	no
Case 8	3	2	4	1	1	2	5		no	no	no	no	no
Case 9	2	1	3	0	0	1	4	1		no	no	no	no
Case 10	2	1	3	0	0	1	4	1	0		no	no	no
Case 11	3	2	4	1	1	2	5	2	1	1		no	no
Case 12	3	1	3	0	0	1	4	1	0	0	1		no
Case 13	2	1	3	0	0	1	4	1	0	0	1	0	

3.3.2. Study of the variable regions RvD5 and RD152

In order to find further discriminatory characteristics of this strain, we studied the complete genome in the IGV program using CDC1551 strain (NC_002755.2) and H37Rv (NC_000962.3) as references. To study the location and number of repeats of IS*6110* we could not rely on WGS as it cannot ascertain the location of insertion sequences (ISs) in the genome because repeated sequences are not well assembled with the short lectures used by WGS. We therefore searched for the presence of large deletions that could result from recombination events caused by IS*6110,* as described for H37Rv (Fang et al., 1999 and Ho et al., 2000).

The comparison of the genomes obtained with CDC1551 in the IGV program using the bam files detected two large deleted regions. It was suspected that a copy of IS*6110* could be inserted at those points, so we amplified and re-sequenced the regions in order to check their arrangements. The first of the found deleted regions showed the loss of genes from *MT3427 to MT3429* in our strain (Fig. 4a). This region showed 100% homology with an IS*6110* and showed an entire *MT3426* gene. It was also checked in the reference H37Rv strain, which lacks these genes (RvD5) and has a truncated *Rv3324A* (*MT3426*) by an IS*6110*. These differences explain the dissimilar size of the bands in the PCR between our strain and H37Rv (Fig. 4b). The second of the found deleted region matched with RD152. In comparison to CDC1551, it showed a deletion between genes *MT1797 (Rv1754c)* and *MT1805 (cut1)*. This deletion was larger than the one in H37Rv, mainly due to the entire *plcD* (*MT1799*) deleted in the outbreak strain. The sequence of the amplicon obtained revealed that the strain has an IS*6110* inserted at gene *Rv1754c* (between the points 1,986,625–1,989,080 of the H37Rv genome). This is possibly due to a recombination, because the direct repeats created in the transposition event were absent, this strain lost all the genes between a truncated *Rv1754c* and a truncated *cut1*. The *cut1* was also found truncated in both CDC1551 and H37Rv strains, carrying also a copy of IS*6110*. However, our strain has twenty-two more nucleotides deleted (Fig. 4c).

4. Discussion

The molecular surveillance in our Autonomous Community, carried out since 2004, allowed us to identify this outbreak. Its rapid spread among the population in only four years, the short period of evolution to disease and the fact that the genotype of this strain was not present before in our region, made it of relevance for further characterisation and investigation.

4.1. Epidemiological and Molecular analysis of the outbreak

As the classical genotyping techniques used routinely were not able to establish the direction of transmission, we decided to use WGS as it had proven successful in other similar outbreaks (Lalor et al., 2018). WGS was however unable to provide a definitive answer about the route transmission as only few unique SNPs were found. This finding suggests that the SNPs appeared after the strain was established in the patient, although the SNPs generation during the in vitro process in the laboratory may also occur. The low number of found SNPs may indicate

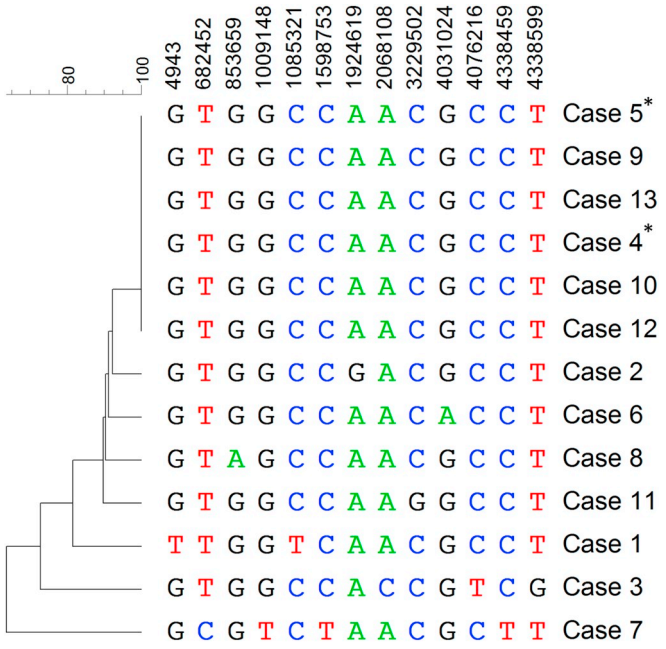


Fig. 3. Dendrogram showing the different SNPs between isolates. The numbers above indicate the position of the SNP within H37Rv genome. Isolates from Cases 4, 5, 9, 10, 12 and 13 were identical. The most different one is the isolate from Case 7. Isolates with an ambiguous base are marked with an asterisk.

that this strain has a low mutation ability or that the spread was fast. The latter can be supported by the fact that Case 5 acted as a super-spreader what played an important role in the low genetic variability found. Besides, the number of SNPs between the different isolates could be an indication of the time span between the infection and the diagnosis. There were four SNPs between Case 5 and Case 7, which is the largest genetic distance between the cases disseminated by Case 5. We suspect that Case 7 was the first of the brothers infected at the beginning of the disease of Case 5. As the diagnose of Case 7 was not until 2016, the strain could have evolved in the lungs for more time, up to two years, compared to the other isolates.

Based on WGS, we only could assess the direct transmission of two cases, in which an ambiguous SNP was detected. Remarkably, these ambiguous SNPs were eliminated of the systematic analysis due to the filter used, and searched manually after a time-consuming intervention that proved to be worth it, as it provided the only two genetic links achieved in this investigation. For Case 12, the epidemiological suspected link to Case 5 was confirmed. Besides, the ambiguous base established the only link between Case 10 and Case 4, since no epidemiological relation had otherwise been established. As reported before (Casali et al., 2016), WGS technique is not always able to establish the transmission network without enough epidemiological data and, even though, it is necessary that the isolates have a substantial number of different transmitted SNPs.

The searching of the index case did not yield results. The first diagnosed case was a young female whose father lived between Spain and Gambia (Case 1). There was not known contact of TB for her, she was BK negative and she developed a peritoneal TB, therefore it was highly unlikely that she were the spreader of the disease. The hypothesis is that a relative of Case 1, who had not been diagnosed in our region, imported the strain from Gambia.

Another possibility is that either Case 2 or 5, which presented symptoms at the same time as Case 1, were the index case. Both were contagious (BK+). Moreover, both started with symptoms in the first term of 2014, perhaps even in 2013. Case 2 liked travelling, which would support the hypothesis that he imported the strain and somehow he infected Case 1 and Case 5. As Case 5 did not refer any travel abroad, the acquisition of the strain seems to be unlikely. The possibility of someone else of foreign origin and diagnosed abroad being the index case, who infected Case 1, 2 and/or 5 is also possible.

In a previous study concerning clustering in London in a three year period (Hamblion et al., 2016), the authors suggested that if the two first cases are diagnosed in less than 90 days, the outbreak will become a large cluster of more than five cases. Our outbreak confirms the conclusion drawn from this study. The first three cases appeared at the same time (January 2014), even though the last two were diagnosed later, and it became a large cluster of more than ten cases. The study

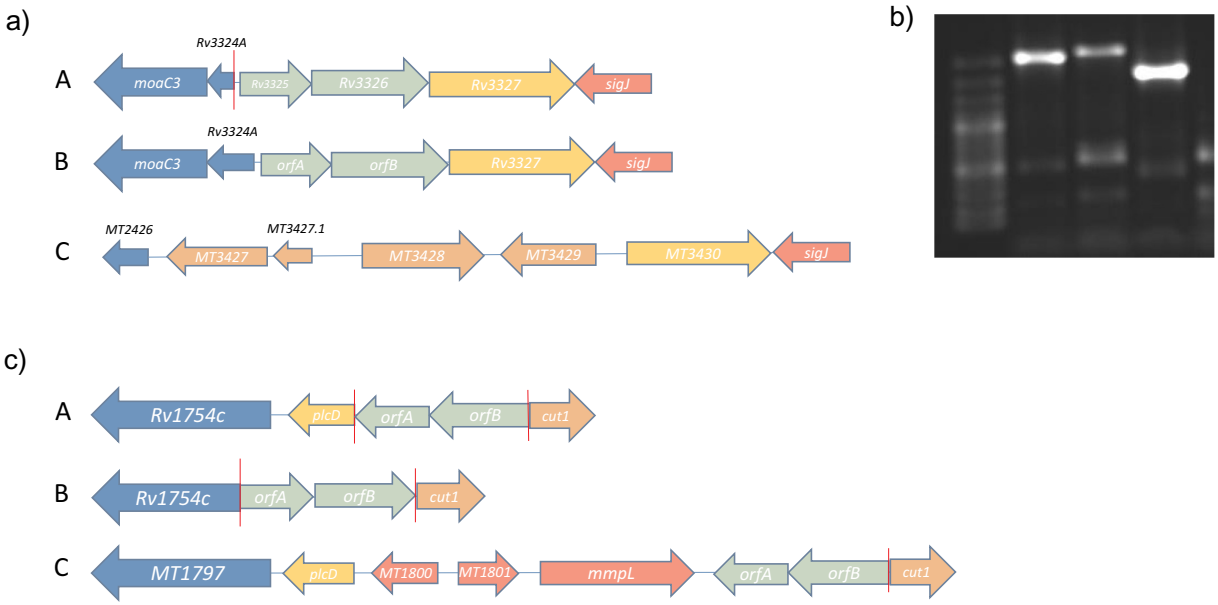


Fig. 4. Study of RvD5 and RD152. a) RvD5 region in H37Rv (A), outbreak strain (B) and CDC1551 (C). b) PCR of RvD5 region, Line 1: Case 1, Line 2: Case 4, Line 3: H37Rv DNA as control. For the amplification, we used primers from the genes *Rv3324A* and *Rv3327*, both present in our strain and in H37Rv (Table 1). It can be observed that H37Rv band is smaller than our strain band. The difference is *Rv3324A* gene: our strain has a complete copy while H37Rv has a truncated one. c) RD152 region in H37Rv (A), outbreak strain (B) and CDC1551 (C). The red line indicates a truncation of the gene. The green arrows are the IS*6110*.

also reported that in large clusters, regional clusters are more probable than local clusters. At least two cases in our outbreak were from a village, which supports this idea.

There are two aspects of the outbreak worth highlighting. The first is the fact that Case 5 remained two years or more being infectious, but not diagnosed. He was a 21 year old male, therefore he should have been healthy. The doctors never thought he could have TB in spite of his symptoms. The diagnosis of tuberculosis should have been suspected many months before and some measures should have been taken in order to avoid any similar situation again. The second is the fact that Case 13 rejected the prophylaxis when his tuberculin skin test was positive. Two years later, he developed an active TB and he could have infected other people, as he was BK+. In our opinion, the prophylaxis should be compulsory unless it is contraindicated by an exceptional situation of the patient.

4.2. Genomic characterisation of the strain

WGS resulted very useful to determine the molecular characteristics of the outbreak. The strain lineage, its family and the possible antibiotic resistance were identified in a straightforward and efficient way.

M. tuberculosis strains of Lineage 4.3, to which our strain belongs, are globally distributed. They are present in Europe, South Africa, America, Australia, Philippines and Indonesia (Stucki et al., 2016). L4 and L2 are the most successful lineages, and the majority of TB cases diagnosed in Spain are from L4. Lineage 4.3 is also present in Gambia (Stucki et al., 2016), so it is not irrational to assume that a relative of Case 1 could have imported the strain from his country of origin, thus it would not have been present before in our population.

The drug susceptibility tests performed in the clinical laboratory where the isolates were cultured, showed that all the isolates were susceptible to first line antibiotics. Despite that some SNPs were found in genes related to drug resistance, none of them were identified with certainty, so the genotype was in concordance with the phenotypical susceptibility demonstrated in the laboratory.

WGS enhances the possibility of examining other genomic regions that classical genotyping methods cannot study. The two large deletions found were probably caused by recombination events caused by the insertion of an IS*6110,* although the arrangement of the regions differed from the ones in CDC1551 and H37Rv. Both regions, RvD5 and RD152, are known for being hot spots for insertion sequences (Fang and Forbes, 1997; Vera-Cabrera et al., 2001; Talarico et al., 2005).

5. Conclusions

WGS was useful in order to characterise the strain and to study it from a molecular point of view. However, it did not seem to be of great use to determine the transmission route, perhaps due to the characteristics of the outbreak influenced by the presence of a super-spreader. Furthermore, the epidemiological data was indispensable in explaining the links between the different cases.

Author contributions

Project design: JC and SS; Experimental activity: JC and AC; Epidemiological activity: AC, JC and DI. Data and material supplier: JV, MAV and MJI. Manuscript preparation: JC, AC, SS. Final revision: JC and SS.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors would like thank Miren Telletxea and Montserrat Gutierrez for proofreading the manuscript. We also thank the use of Servicio General de Apoyo a la Investigación-SAI, Universidad de Zaragoza and Servicios Científico Técnicos de CIBA (IACS-Universidad de Zaragoza).

Formatting of funding sources

This work was supported by the Instituto de Salud Carlos III (FIS 15/0317 and 18/0336), and Gobierno de Aragón/Fondo Social Europeo, “Construyendo Europa desde Aragón”. Jessica Comín was recipient of a Government of Aragon grant “European Union-Fondo Social Europeo”.

References

Alonso, H., Aguilo, J.I., Samper, S., Caminero, J.A., Campos-Herrero, M.I., Gicquel, B., et al., 2011. Deciphering the role of IS6110 in a highly transmissible mycobacterium tuberculosis Beijing strain, GC1237. Tuberculosis. https://doi.org/10.1016/j.tube.2010.12.007.

Borgdorff, M.W., Behr, M.A., NJD, Nagelkerke, Hopewell, P.C., Small, P.M., 2000. Transmission of tuberculosis in San Francisco and its association with immigration and ethnicity. Int. J. Tuberc. Lung Dis. 4, 287–294.

Casali, N., Broda, A., Harris, S.R., Parkhill, J., Brown, T., Drobniewski, F., 2016. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: a retrospective observational study. PLoS Med. 13, 1–18. https://doi.org/10.1371/journal.pmed.1002137.

Cirillo, D.M., Cabibbe, A.M., De Filippo, M.R., Trovato, A., Simonetti, T., Rossolini, G.M., et al., 2016. Use of WGS in *Mycobacterium tuberculosis* routine diagnosis. Int. J. Mycobacteriol. https://doi.org/10.1016/J.IJMYCO.2016.09.053.

Coll, F., McNerney, R., Guerra-Assunção, J.A., Glynn, J.R., Perdigão, J., Viveiros, M., et al., 2014. A robust SNP barcode for typing mycobacterium tuberculosis complex strains. Nat. Commun. 5, 4–8. https://doi.org/10.1038/ncomms5812.

Fang, Z., Forbes, K.J., 1997. A mycobacterium tuberculosis IS6110 preferential locus (ipl) for insertion into the genome. J. Clin. Microbiol. 35, 479–481.

Fang, Z., Doig, C., Kenna, D.T., Smittipat, N., Palittapongarnpim, P., Watt, B., et al., 1999. IS6110-mediated deletions of wild-type chromosomes of mycobacterium tuberculosis. J. Bacteriol. 181, 1014–1020.

Gonzalo-Asensio, J., Pérez, I., Aguiló, N., Uranga, S., Picó, A., Lampreave, C., et al., 2018. New insights into the transposition mechanisms of IS6110 and its dynamic distribution between mycobacterium tuberculosis complex lineages. PLoS Genet. https://doi.org/10.1371/journal.pgen.1007282.

Hamblion, E.L., Le Menach, A., Anderson, L.F., Lalor, M.K., Brown, T., Abubakar, I., et al., 2016. Recent TB transmission, clustering and predictors of large clusters in London, 2010-2012: results from first 3-years of universal MIRU-VNTR strain typing. Thorax 71, 749–756. https://doi.org/10.1016/j.matdes.2016.12.002.

Ho, T.B.L., Robertson, B.D., Taylor, G.M., Shaw, R.J., Young, D.B., 2000. Comparison of mycobacterium tuberculosis genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. Yeast 1, 272–282. https://doi.org/10.1002/1097-0061(200012)17:4<272::AID-YEA48>3.0.CO;2-2.

Kamerbeek, J., Schouls, L., Kolk, A., Van Agterveld, M., Van Soolingen, D., Kuijper, S., et al., 1997. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.

Lalor, M.K., Casali, N., Walker, T.M., Anderson, L.F., Davidson, J.A., Ratna, N., et al., 2018. The use of whole-genome sequencing in cluster investigation of a multidrug-resistant tuberculosis outbreak. Eur. Respir. J. 51, 1702313. https://doi.org/10.1183/13993003.02313-2017.

Millan-Lou, M.I., López-Calleja, M.I., Colmenarejo, C., Lezcano, M.A., Vitoria, M.A., Del Portillo, P., et al., 2013. Global study of is6110 in a successful mycobacterium tuberculosis strain: clues for deciphering its behavior and for its rapid detection. J. Clin. Microbiol. https://doi.org/10.1128/JCM.00970-13.

Rindi, L., Medici, C., Bimbi, N., Buzzigoli, A., Lari, N., Garzelli, C., 2014. Genomic variability of Mycobacterium tuberculosis strains of the euro-american lineage based on large sequence deletions and 15-locus MIRU-VNTR polymorphism. PLoS One 9. https://doi.org/10.1371/journal.pone.0107150.

van Soolingen, D., de Haas, P.E., Hermans, P.W., van Embden, J.D., 1994. DNA fingerprinting of *Mycobacterium tuberculosis*. Methods Enzymol. 235, 196–205.

Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., et al., 2016. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat. Genet. 48, 1535–1543. https://doi.org/10.1038/ng.3704.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rüsch-Gerdes, S., Willery, E., et al., 2006. Proposal for standardization of optimized Mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. J. Clin. Microbiol. https://doi.org/10.1128/JCM.01392-06.

Talarico, S., Durmaz, R., Yang, Z., 2005. Insertion- and deletion-associated genetic diversity of mycobacterium tuberculosis phospholipase C-encoding genes among 106 clinical isolates from Turkey. J. Clin. Microbiol. 43, 533–538. https://doi.org/10.1128/JCM.43.2.533-538.2005.

Van Embden, J.D.A., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., et al., 1993. Strain identification of Mycobacterium tuberculosis by DNA

fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. 31 (2), 406–409.
Vera-Cabrera, L., Hernández-Vera, M.A., Welsh, O., Johnson, W.M., Castro-Garza, J., 2001. Phospholipase region of mycobacterium tuberculosis is a preferential locus for IS*6110* transposition. J. Clin. Microbiol. https://doi.org/10.1128/JCM.39.10.3499-3504.2001.
Walker, T.M., Lalor, M.K., Broda, A., Ortega, L.S., Morgan, M., Parker, L., et al., 2014. Assessment of mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. Lancet Respir. Med. https://doi.org/10.1016/S2213-2600(14)70027-X.
WHO Date Last Updated: 17 October 2019 (Date last accessed: 28 November 2019). https://www.who.int/news-room/fact-sheets/detail/tuberculosis.

Publicación 5

Tuberculosis 126 (2021) 102022

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://www.elsevier.com/locate/tube

A whole-genome sequencing study of an X-family tuberculosis outbreak focus on transmission chain along 25 years

Jessica Comín [a,*], Alberto Cebollada [a], Daniel Ibarz [b], Jesús Viñuelas [c,d,e], María Asunción Vitoria [d,f], María José Iglesias [b,e,g], Sofía Samper [a,e,g]

[a] Instituto Aragonés de Ciencias de la Salud, Zaragoza, C/de San Juan Bosco, 13, 50009, Zaragoza, Spain
[b] Universidad de Zaragoza, C/Domingo Miral S/N, 50009, Zaragoza, Spain
[c] Hospital Universitario Miguel Servet, Paseo Isabel la Católica, 1-3, 50009, Zaragoza, Spain
[d] Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica, C/Agustín de Bentacourt, No 13, 28003, Madrid, Spain
[e] Fundación IIS Aragón, C/de San Juan Bosco, 13, 50009, Zaragoza, Spain
[f] Hospital Clínico Universitario Lozano Blesa, Avda. San Juan Bosco, 15, 50009, Zaragoza, Spain
[g] CIBER de Enfermedades Respiratorias, Av. Monforte de Lemos, 3-5. Pabellón 11, Planta 0, 28029, Madrid, Spain

ARTICLE INFO

Keywords:
Tuberculosis
Tuberculosis outbreak
X-family tuberculosis strain

ABSTRACT

Lineage 4/X-family of *Mycobacterium tuberculosis* is not very notorious, except for the CDC1551 strain. One strain of this family, named Ara50, caused one of the largest tuberculosis outbreaks of the Aragon region, Spain, during the 1990s and remained until 2018. These X-strains are characterised by high transmissibility and by carrying a low copy number of IS*6110* in their genomes. Epidemiological data of the 61 patients consisted of inmates, HIV seropositives, intravenous drug users and the homeless. The application of whole-genome sequencing (WGS) to 36 out of 61 isolates, selected by IS*6110*-RFLP, allowed to confirm 32 as recent transmissions. We found 10 SNPs in genes considered as virulence factors, five of them specific of this strain. WGS identified three sub-clusters (CLSs). The largest one, sub-CLS 1, included 10 cases. Seven of them shared a SNP in the *mce3C* gene, considered a virulence factor gene. Sub-CLS 2 involved familiar cases, and no link was known for sub-CLS 3. Finally, the strain showed efficacy in latency as a confirmed epidemiological link was established between two cases, with 6 years of distance in their diagnosis. This outbreak study combined epidemiological and molecular analyses in order to elucidate tuberculosis transmission.

1. Introduction

Whole-genome sequencing (WGS) is nowadays essential for studying infectious diseases from both molecular and epidemiological points of view [1]. Tuberculosis (TB) is still an important cause of death, especially in developing countries where other diseases, such as AIDS, play an important role in the fatal result [2].

Mycobacterium tuberculosis is a complex pathogen with a long incubation and latent period, which makes the study of outbreaks and contact tracing difficult [3]. There are seven different lineages (Ls) and one animal strain branch, all of them potentially causing the disease and with specific characteristics [4]. L4 is the most extended globally [5], and its success has been associated with having a high copy number of insertion sequence (IS) *6110* [6]. However, CDC1551 strain, belonging to L4, L4.1.1.3/X-family, with only four copies of IS*6110* showed an unusually high rate of transmission in a rural area of the United States (US) in 1995 [7].

Aragon (a region in northern Spain) has applied genotyping to all *M. tuberculosis* complex isolates for surveillance purposes routinely since 2004. Studies were performed in Zaragoza for the periods 1993–1995 and 2001–2004 [8]. Combining all the genotypes available, we discovered five outbreaks caused by an X-family strain, but just one, named Ara50, grouped more than 50 cases. It was the largest outbreaks in the 1990s among our population; likewise, we could find the same genotype ongoing to 2018. This strain was selected to be studied due its similarity with CDC1551, a known well-transmitted strain [7] with low copy number of IS*6110*, its long persistence, and the large number of cases it produced. We applied WGS in order to study the molecular

* Corresponding author. Unidad de Investigación Traslacional-Hospital Universitario Miguel Servet, Paseo Isabel la Católica, 1-3, 50009, Zaragoza, Spain.
E-mail addresses: jcomin.iacs@aragon.es, jessicacp_94@hotmail.com (J. Comín), alberto@unizar.es (A. Cebollada), dibarz@unizar.es (D. Ibarz), jvinuelasbayon@yahoo.es (J. Viñuelas), avitoria@salud.aragon.es (M.A. Vitoria), iglesias@unizar.es (M.J. Iglesias), ssamper.iacs@aragon.es (S. Samper).

https://doi.org/10.1016/j.tube.2020.102022
Received 25 August 2020; Received in revised form 13 November 2020; Accepted 15 November 2020
Available online 28 November 2020

characteristics of this X strain and to elucidate its evolution along 25 years and its transmission chain, as other authors have succeeded using this technique [9–11]. We were successful in finding out the molecular characteristics of the strain and partially successful in identifying the transmission chain.

2. Materials and methods

2.1. Clinical samples and cases

Systematic genotyping of all *M. tuberculosis* strains isolated in Aragon, coordinated by public health services, has been carried out since 2004 up to now. These RFLP genotypes are included in Bionumerics software v7.6 (Applied Maths, Kortrijk, Belgium), which allowed the comparison among all the analysed patterns, including those from two previous population studies from the periods 1993–1995 and 2001–2004 in Zaragoza.

The search for X-family genotypes belonging to isolates from the period 2004–2019, among more than 6000 IS*6110*-RFLP and spoligotype patterns, was performed in the Bionumerics software. Fifty-two patterns belonged to the X-family and some of them were grouped in five clusters. One of them, named Ara50, which involved 27 isolates corresponding to 26 different cases.

The search for this Ara50 genotype in the two previous studies, carried out in an Aragon population, allowed the discovery of 34 more isolates between the periods of 1993–1995 (27 cases) and 2001–2003 (seven cases), with an IS*6110*-RFLP pattern compatible with the Ara50 X-family strain. A total of 61 genotypes were selected as possible Ara50 isolates. Enough DNA stored at −80 °C was available for 36 of them, which could be analysed by WGS.

Epidemiological, clinical and microbiological data of the patients were collected and carefully studied. All the data were anonymous in order to protect the identity of the patients.

2.2. Standard molecular typing

DNA of the isolates was obtained using the cetrimonium bromide method, as previously described [12]. All DNA extractions were stored at −80 °C until sequencing. All the isolates were genotyped by standardised techniques, including IS*6110*-RFLP (all the isolates), spoligotyping (all the isolates since 2001) and 24-MIRU-VNTR (one representative isolate), as previously described [13–15]. The genetic patterns obtained were stored and analysed by Bionumerics database software. Spoligotypes were systematically revised against the SITVIT database (www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE). Gilla Kaplan kindly provided CDC1551 strain, used for IS*6110*-RFLP and Spoligotyping techniques.

2.3. IonTorrent sequencing

Thirty-six isolates with a similar RFLP pattern, considered as part of the Ara50 cluster, were recovered and sequenced using IonTorrent technology. The library was prepared using the Ion Xpress™ Plus Fragment Library Kit, following the manufacturer's instructions, and the sequencing and analysis took place in the Ion S5™ XL System. Once the sequences were obtained, they were mapped against the reference strains H37Rv (NC_000962.3) and CDC1551 (NC_002755.2) using the Torrent Mapper TMAP version 5.12.27. Variants were called, excluding high variable regions such as *ppe* and *pe_pgrs* genes, by the Torrent Variant Caller plugin. This allowed us to obtain a bam file to use for the Integrative Genomics Viewer (IGV) program.

2.4. Lineage and family identification by WGS

We used the SNP classification of Coll et al. [16], in which each MTB lineage and family is associated with specific SNPs in different genomic regions. The SNPs that classified the studied strain as L4.1.1.3 and X-family are described in the Results section.

2.5. Bioinformatic tools for the study of the genomes

The fastQ files of the sequences obtained were introduced and analysed in Bionumerics software. The program mapped the sequences against the reference strains H37Rv and CDC1551, and obtained the different SNPs among the genomes analysed. The analysis of SNPs allowed the construction of a dendrogram using the UPGMA method. For greater accuracy, strict SNP filtering that removed positions with at least one ambiguous or unreliable base, gaps (maximun frequency 1%), non-discriminatory positions and *ppe* and *pgrs* genes, was applied. It was also considered that the retained SNP positions had a minimum 5× coverage and that the minimum distance between SNPs was at least 12 base pairs (bp). To study the phylogeny of the isolates and their drug-resistance profile, the IGV program and TubercuList website (http://genolist.pasteur.fr/TubercuList/) were used. For the IS*6110* location, reads containing the first 30 and the last 30 nucleotides of the IS*6110* were selected using BLAST, and they were subsequently studied using TubercuList.

2.6. Phylogenetic tree

The SNPs of Ara50 and X-non Ara50 strains were extracted using SNIPPY tool. The consensus sequenced obtained, among CDC1551 and H37Rv sequences, were aligned and the tree was constructed using Parsnp tool. The tree was visualized using TreeView tool.

2.7. Data availability

Sequences of Case 2 (Ara50 type) (accession number SAMN15501351) and Case 36 (evolution of Ara50 strain with one more IS*6110* copy) (accession number SAMN15501352) have been uploaded to BioSample database. They are included in the BioProject PRJNA645275.

3. Results

Based on the IS*6110*-RFLP (five bands), spoligotyping (octal code 067776777760771) and 24-MIRU-VNTR (25324433234425153322743, ordered according their appearance in the genome) patterns, we suspected 61 cases were affected by this X-family strain. Fifty-six of these isolates had an identical RFLP pattern. The other five had a slightly different pattern with five bands (but not exactly at the same position) or with six bands in case the strain could have evolved. The different patterns selected can be observed in Fig. 1A.

Thirty-six isolates of the 61 suspected cases had the required amount of DNA to apply the WGS technique. Once the WGS analysis was done, we discerned that two of the isolates were Haarlem strains, as they had the specific SNPs for this family [16], while another was an X-strain, but not from the Ara50 outbreak, as it had more than 150 SNPs in comparison to the rest of the isolates. One more isolate was excluded, as it was a cross-contamination. Finally, 32 isolates were included in the genomic study.

3.1. Genomic Ara50 strain characterisation

3.1.1. IS6110 analysis

WGS allowed the location study of the five copies of IS*6110* determined in the RFLP genotype of the Ara50 strain, the additional IS*6110* in Case 36 and in the outgroup strains (the two Haarlem and the non-Ara50 *X* strains). The results are shown in Table 1. All the points are referred to the H37Rv genome (NC_000962.3) or to the *M. bovis* genome (LT708304.1), in case the location was not present in H37Rv. Four of these IS*6110* were at the same location that the ones in CDC1551 strain.

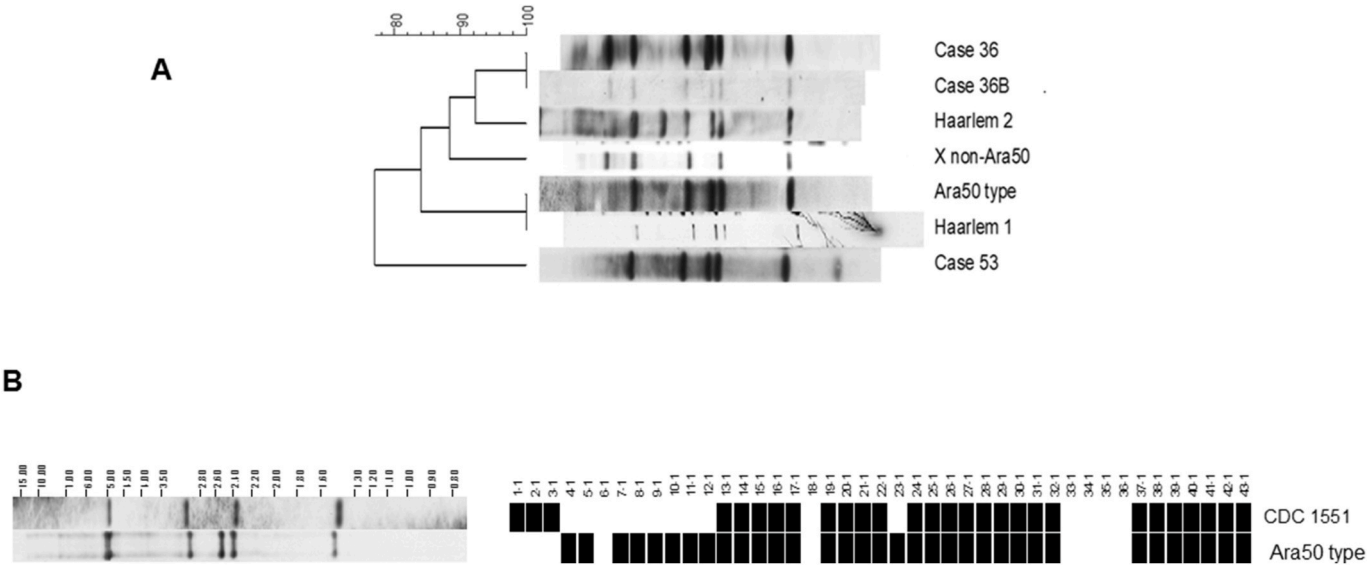


Fig. 1. A: IS*6110*-RFLP patterns of the strains selected in the Bionumerics database as possible evolution of the Ara50 strain. After WGS, we found that Case 36 and Case 36 B had an additional band corresponding to a transposition of IS*6110*. One selected isolate was an X strain but did not belong to the Ara50 outbreak, and two isolates were Haarlem strains. The genomic sequence was not available for Case 53, although it was considered as belonging to the cluster. B: IS*6110*-RFLP and Spoligotyping of Ara50 and CDC1551 strains. Four bands in the RFLP are shared by both strains.

Table 1
IS*6110* location for the Ara50-type, Case 36 (with six bands in the RFLP instead of five) and the outgroup strains (non-Ara50 *X* strain, Haarlem 1 and Haarlem 2). *Points referred to *M.bovis* genome.

Strain	Point of insertion	Gene
Ara50	483296	*mmpS1*
	1979899*	*cut1*
	3120523–3121879	DR region
	3123104	DR region
	3377327	*ppe46*
	3665328*	*MT3427* (extra IS of Case 36)
Non-Ara50 *X* strain	483296	*mmpS1*
	888920	*Rv0794c:Rv0795*
	1979899*	*cut1*
	3120523–3121879	DR region
	3377327	*ppe46*
Haarlem 1	483296	*mmpS1*
	1075948	*Rv0963c*
	1715972	*mmpL12*
	1986622	*Rv1754c*
	1987254	*plcD*
	2610861	*Rv2336*
	3120523–3121879	DR region
	3550924	*Rv3183*
	3668575-3668756*	*MT3429*
Haarlem 2	79922	*Rv0071*
	153026	*ppe19*
	483296	*mmpS1*
	1075948	*Rv0963c*
	1986622	Rv1754c
	2038634	*Rv1798:lppT*
	2610861	*Rv2336*
	2634049	*ppe38*
	3076505*	DR region
	3120523–3121879	DR region
	3491619	*ppe49*
	3668575-3668756*	*MT3429*

Ara50 had an additional IS*6110* in the DR region. The IS*6110*-RFLP and Spoligotyping patterns of Ara50 and CDC1551 strains are shown in Fig. 1B.

3.1.2. SNP analysis

According to the SNP classification established by Coll et al. [16], this strain belonged to lineage 4.1.1.3/X-family, as it has SNPs in positions 931123, 62657, 514245 and 4229087.

Regarding the genes associated with antimicrobial resistance, we found mutations in *katG* (point 2156096), *embB* (point 4249408), *gyrA* (points 7362, 7585 and 9304), *tlyA* (point 1917972) and *fbiA* (point 3,641,091). Nevertheless, none of these polymorphisms was established as conferring resistance, which is in agreement with the antibiotic test results obtained by the clinical laboratories.

This strain had 918 SNPs when compared to H37Rv genome (NC_000962.3). The comparison against CDC1551 genome revealed 408 SNPs. Fifty-seven SNPs were considered specific to the Ara50 strain, as we detected them in comparison to other sequenced strains we had from other studies (including other X strains and a variety of lineages) (Table S1). A phylogeny was constructed using CDC1551, Ara50 and the X non-Ara50 genomes using H37Rv as an outgroup (Fig. S1). X non-Ara50 genome is more closely related with CDC1551 than Ara50 strain.

The strain had non-synonymous SNPs in 10 genes considered virulence factors [17]: *mmaA4*, *mbtB*, *mmpL7*, *sigH*, *ctpC*, *pks7*, *pks5*, *katG*, *fadE29* and *lpqH*. The first five were specific SNPs, while the last four were also present in CDC1551 and in the non-Ara50 *X* strains. The mutation in *pks7* was both present in Ara50 and non-Ara50 *X* strains.

3.2. Epidemiological study of Ara50 outbreak

This strain has remained for 25 years. The first cases were detected in 1993, when our records started, maintaining the highest number of cases per year in the first three years studied (1993–1995). Since 2001, when the records started again, the number of cases per year has diminished substantially, with a punctual rise in 2006. After 3 years with no cases, 2015–2017, the strain reappeared in 2018 with two more cases. No other case has appeared since then (Fig. 2).

The epidemiological study was carried out on the 56 cases described in Fig. 2. The majority of affected patients were autochthonous with ages from 4 to 87, with those aged between 16 and 45 being the most affected. Considering just the known ones, we observed a 55.6% of HIV + patients and a 38.8% of intravenous drugs (ID) users or ex-users. Besides, 78.6% were smokers and more than 25% reported high alcohol consumption. Thirty-two cases presented a positive sputum smear result and, therefore, were potentially infectious. We also observed a high proportion of dysfunctional family structure (50%). A

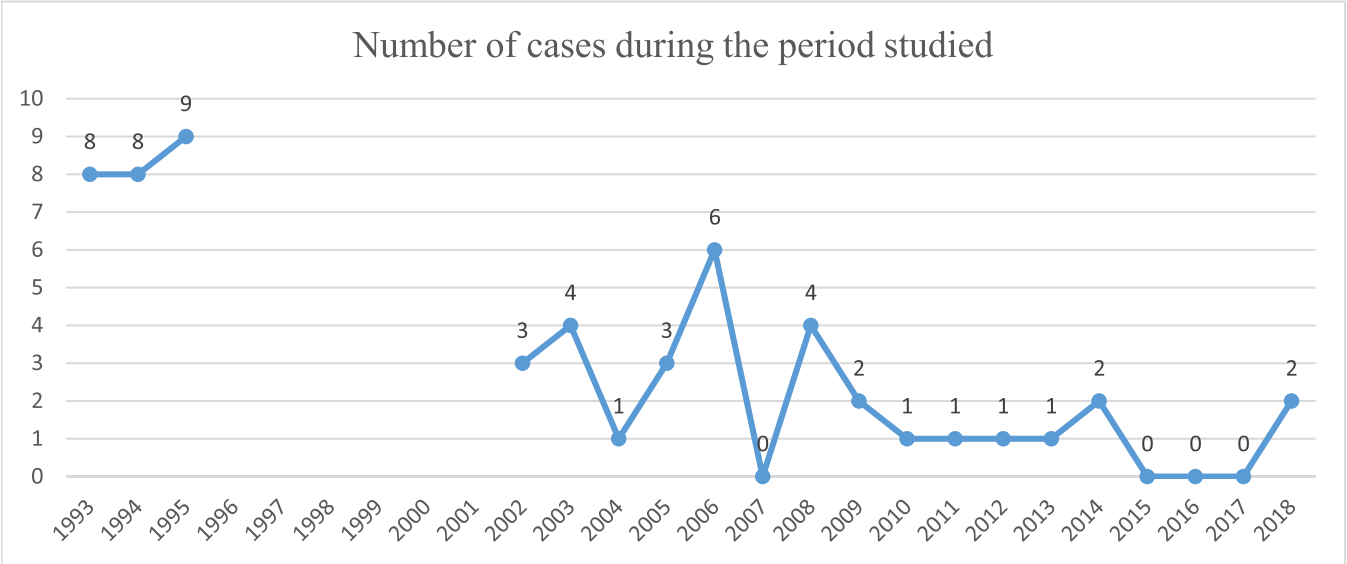


Fig. 2. Number of cases along the period studied. There was no available data during the period 1996–2000. A reduction in the number of cases during the second period of the study was observed, so the trend would seem to be the elimination of the strain. A second isolate of Case 36 and non-Ara50 cases finally excluded by WGS are omitted in the graphic.

Table 2
Age and risk factors associated with the X-family strain outbreak. *Dysfunctional family structure includes prisoners, indigents and prostitution. The unknown data are not considered in the percentage.

	N (total N = 56)
Age (years)	
0-15	2 (3.6%)
16-30	15 (26.8%)
31-45	24 (42.9%)
46-60	9 (16.1%)
>60	6 (10.7%)
HIV state	
HIV+	30 (55.6%)
HIV-	24 (44.4%)
Unknown	2
Use of Intravenous drugs (ID)	
User or ex-user of ID	19 (38.8%)
No use of ID	30 (61.2%)
Unknown	7
Smoking	
Smoker	33 (78.6%)
Non smoker	9 (21.4%)
Unknown	14
Alcohol consumption	
High	11 (27.5%)
Moderate	7 (17.5%)
Low/No consumption	22 (55%)
Unknown	16
Family structure	
Structured	14 (50%)
Dysfunctional*	14 (50%)
Unknown	28
Previous TB contact	
Yes	4 (9.8%)
No	37 (90.2%)
Unknown	15
Treatment compliance	
Correctly	8 (44.4%)
Irregularly/No	10 (55.5%)
Unknown	38
Baciloscopy (BK) result	
BK+	32 (57.1%)
BK-	24 (42.9%)

summary of the risk factors can be found in Table 2. In addition, in order to study a possible neighbourhood link, a map marking the addresses of the cases was constructed (data not shown). Six of the cases lived outside the city, while the rest were distributed in the different neighbourhoods in Zaragoza. Nine of the patients had been in prison, during or after the TB diagnosis. Moreover, four cases had been in the methadone unit closely in time. No other epidemiological links could be established using the epidemiological information alone.

3.3. *Ara50 genomic study*

The SNP analysis showed genetic links between several isolates (Fig. 3).

There was a sub-cluster (CLS), sub-CLS 1, that included 10 isolates (Cases 27, 32, 52, 3, 20, 29, 39, 38, 41 and 56) related by the SNPs in positions 2292579 and 2698983. Moreover, it was split by a third SNP in position 2212353, which was shared in seven of these isolates (Cases 3, 20, 29, 39, 38, 41 and 56). Four of them (Cases 27, 32, 29 and 39) could be epidemiologically related for being part of the methadone unit programme and two more cases for being in prison (Case 39 and 38). We know this strain circulated through one of Aragon's penitentiaries, where Case 38 was imprisoned in 2005. The epidemiological information from these 10 cases can be found in Table 3.

Another sub-CLS, sub-CLS 2, was identified using WGS. It was constituted by four cases (28, 45, 44 and 46). Case 28 was the index case of this sub-CLS. The other three were not diagnosed until 6 years later, sharing SNPs in positions 625157 and 4194065. Epidemiological information showed that Cases 28 and 44 were father and son, respectively. A new SNP appeared in position 3632058 in the isolates of the last three cases. No epidemiological link was found between Cases 44, 45 and 46.

Finally, a sub-CLS 3 consisting of two cases, Cases 22 and 37, was identified. No epidemiological link was found between them; however, both isolates had the same genomic sequence, sharing two SNPs in positions 1114864 and 1498542.

There were four genomes, belonging to the period 1993–1995, that did not show any SNPs, being Case 2 the earliest (1993). We considered them the original sequence of the strain.

In Table 4, the SNP distances among the 32 analysed genomes can be

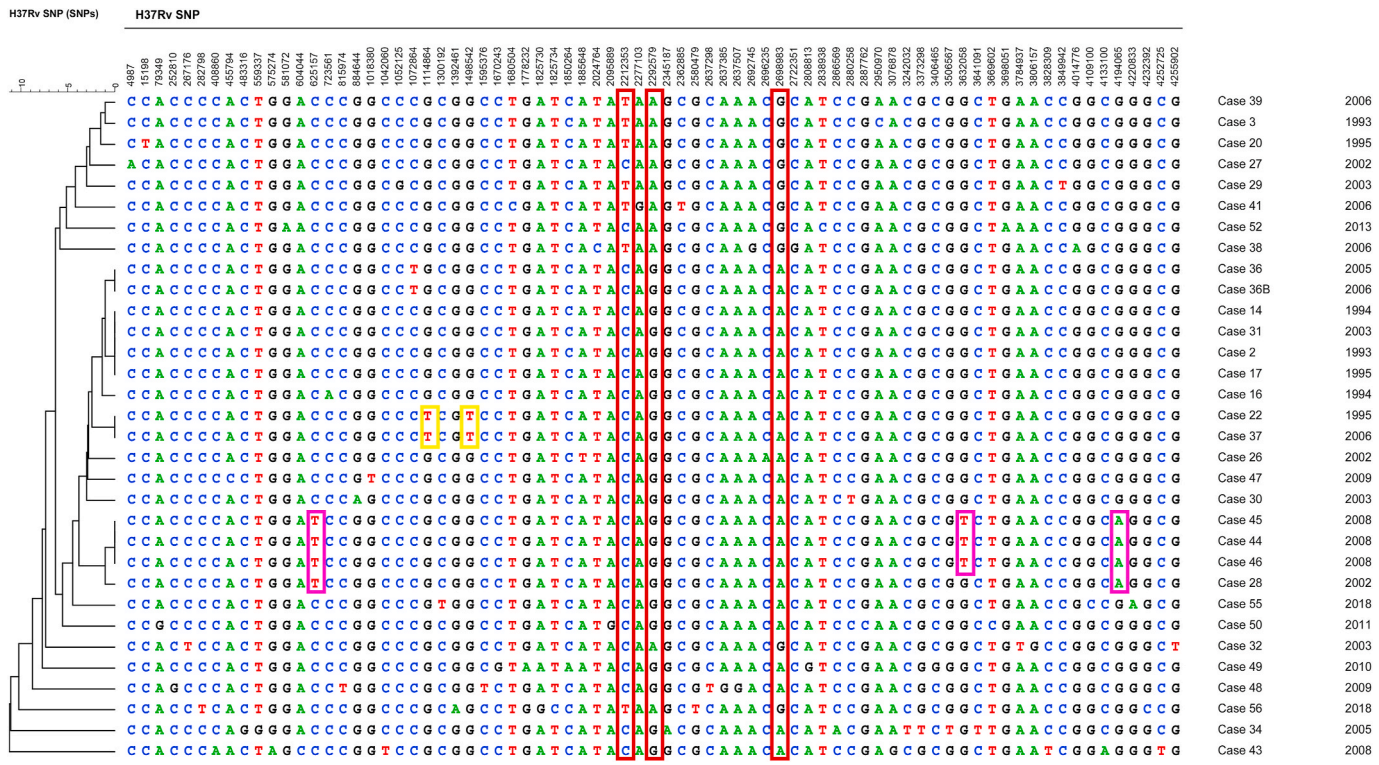


Fig. 3. Dendrogram based on SNP analysis. The numbers indicate the SNP position relative to the H37Rv genome. Specific SNPs of sub-CLS 1 are marked in dark red. Specific SNPs of sub-CLS 2 are marked in pink. Specific SNPs of sub-CLS 3 are marked in yellow.. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Epidemiological information of the isolates involved in the sub-CLS 1.

Case	Gender	Age	Culture date	BK	Country of origin	Comments
Case 3	Female	28	May 24, 1993	+	Spain	Treated of TB in 1988. Involved in prostitution.
Case 20	Female	12	June 09, 1995	–	Spain	
Case 27	Male	43	October 18, 2002	+	Spain	Previous TB episode. Methadone unit.
Case 29	Male	40	April 07, 2003	+	Spain	Methadone unit.
Case 32	Male	43	September 06, 2003	–	Spain	Methadone unit.
Case 38	Male	39	May 09, 2006	+	Spain	Penitentiary
Case 39	Male	46	May 17, 2006	+	Spain	Methadone unit. Penitentiary. Indigence.
Case 41	Male	34	July 11, 2006	–	Spain	
Case 52	Male	82	December 27, 2013	+	Spain	Possible previous TB episode.
Case 56	Female	46	December 05, 2018	–	Spain	

found. It is assumed that, if there are ≤12 SNPs between isolates, both are the same strain. If there are ≤5 SNPs between isolates, there was a recent contact [11]. For 24 isolates, the number of SNPs was ≤5 with at least one isolate. For seven isolates, the number of SNPs were ≥5 in all the cases but ≤12 SNPs with at least one isolate.

Using the link classification of Lalor et al. [11], the only "confirmed epidemiological link" was the one between Cases 28 and 44, known as being father and son. The cases in prison and attending the methadone unit were considered a "probable epidemiological link", as they spent time in the same location but the timing was unknown. Finally, the rest of the cases had a "possible epidemiological link" for sharing the same neighbourhood and social practises (those who were users of ID). The genetic and possible epidemiological links are shown in Fig. 4.

The study of the SNPs that were detected, which are detailed in the dendrogram, was also carried out. The genes involved and the effect of the SNP in terms of synonymous and non-synonymous mutations can be found in Table 5.

One non-synonymous SNP in position 2212353 was found within the *mce3C* gene, an ATP-binding cassette transporter related to virulence. This SNP was detected in seven isolates of sub-CLS 1, but a different clinical behaviour could not be determined. The other two non-synonymous SNPs shared by sub-CLS 1 were detected within *Rv2047c* and *Rv2402* genes, both considered hypothetical proteins.

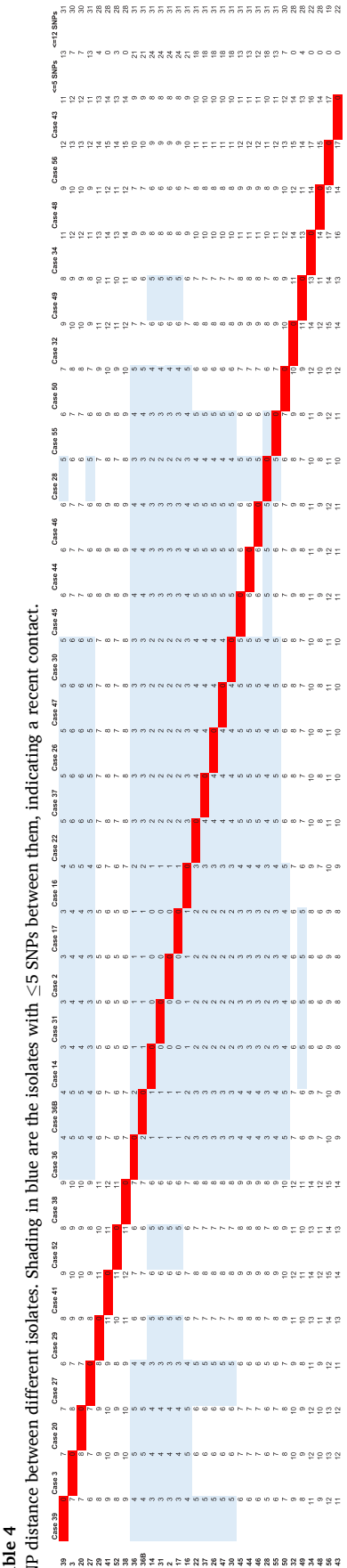
The SNPs transmitted from Case 28 to Case 44 in sub-CLS 2 were within *fabH* and *ctpJ* genes, which are involved in fatty acids biosynthesis and cation transport, respectively. The other SNP probably generated in Case 44 and shared by the other two isolates of sub-CLS 2 was located in the intergenic region *alkB:Rv3252c*.

The SNPs shared by isolates of sub-CLS 3 were in *Rv0998* gene, considered a hypothetical protein, and *dinG* gene, an ATP-dependent helicase involved in DNA repair and replication.

4. Discussion

This study focused on an X-family strain, Ara50 strain. X-family became better known when strain CDC1551 produced an outbreak in 1995. Previous studies have described the great ability of L2 and L4 for spreading globally and related that with the high copy number of IS*6110* they presented [6]. What is interesting about the X-family is that it has a low copy number of IS*6110*, specifically, our X strain carried five copies,

Table 4
SNP distance between different isolates. Shading in blue are the isolates with ≤5 SNPs between them, indicating a recent contact.



despite being L4, and demonstrated a high transmission capacity during the 1990s. Simultaneously, it has had a great ability to persist, as it has circulated for at least 25 years. Ara50 strain produced the largest outbreak in Aragon for 1993–1995 period, and continued being one of the largest during the rest of the study period, with almost 60 cases in total. For 1993–1995, 46.6% of strains were in cluster [8], for 2001–2003, 52.5% of strains [8] and for 2004–2018, 56.1% (not published), being the majority of clusters formed by two or three cases.

The discrimination power of WGS has been largely demonstrated [18]. According to this, WGS has been useful to distinguish between evolved and not related strains, which shared a non-identical but similar genotype. In this work, we could confirm the evolution of one isolate as having ≤12 SNPs when compared to the Ara50 type strain. In addition, it shared the five locations of the IS*6110* with Ara50. Modifications in the RFLP pattern were observed before for strains belonging to the same cluster [19,20]. In contrast, three of these isolates were not related to the Ara50 outbreak. The study of the IS*6110* confirmed it, as were the IS number and location, was different to the ones for Ara50. It was surprising that the number of IS*6110* turned out to be higher than expected for the observed RFLP pattern, although this was previously observed [21].

The SNP study of Ara50 revealed this strain was more closely related to CDC1551 (408 SNPs) than to H37Rv (918 SNPs), something expected as both were X family strains. In addition, the phylogenetic tree (Fig. S1) showed that X non-Ara50 strain is more related with CDC1551 than Ara50 strain, however a close common ancestor among these three strains seems likely. Ara50 had 10 virulence factor genes affected. At least five seem to be specific for the strain; therefore, it is possible that the success of the strain was due, in part, to some of these SNPs.

Tracking the transmission chain in this outbreak was difficult, as multiple infectious foci co-existed at the same time, and it has been demonstrated that identical sequences could derive from different transmission events [10]. Different from another outbreak studied in our laboratory [22], which was contextualised in household and school environments, this outbreak had a high proportion of individuals with risk factors, such as ID users, alcoholics, prisoners, homelessness and prostitution. These individuals have many possibilities to concur in more than one context, creating a suitable ground for TB transmission. Moreover, in these environments, the treatment of TB is not usually followed correctly, and some patients abandon it and suffer relapses. A remarkable fact is that we had some cases with an earlier TB episode. It could be that some of them were relapses or maybe re-infections with the same strain, as they continued frequenting the same circles and behaviours. It is noteworthy that the strain did not evolve any resistance to anti-tuberculosis drugs despite this.

WGS allowed the identification of several sub-CLSs. Sub-CLS 1 was special because a SNP in position 2212353, shared by seven out of the 10 isolates, subdivided the CLS. If epidemiological data had been considered alone, we would have assumed that the four isolates in the methadone unit transmitted the strain among them. Nevertheless, two of the isolates had this SNP, but the other two did not. As this SNP has circulated since 1993 and the methadone unit coincidence was not until 2002–2003, it suggests a different origin, despite being in the same place and close in time. However, the ability of WGS to infer the direction of transmission was not very high. We could only determine the transmission route in sub-CLS 2 by combining both epidemiological and genetic information, as the transmission is towards the accumulation of SNPs [23]. This is consistent with the findings of other authors [3,10,11]. Inferring the direction of transmission for sub-CLS 3 was not possible using the genomic information, since both sequences were identical. However, the fact that one isolate was diagnosed in 1995 and the other in 2006 suggested that Case 22 became infected first.

The study of the SNP distance between isolates revealed that most of them had 0–5 different SNPs, indicating recent contact [11]. For several isolates, unique SNPs were identified, so the SNP distance increased for more than 12 SNPs in some cases, despite belonging to the Ara50 cluster.

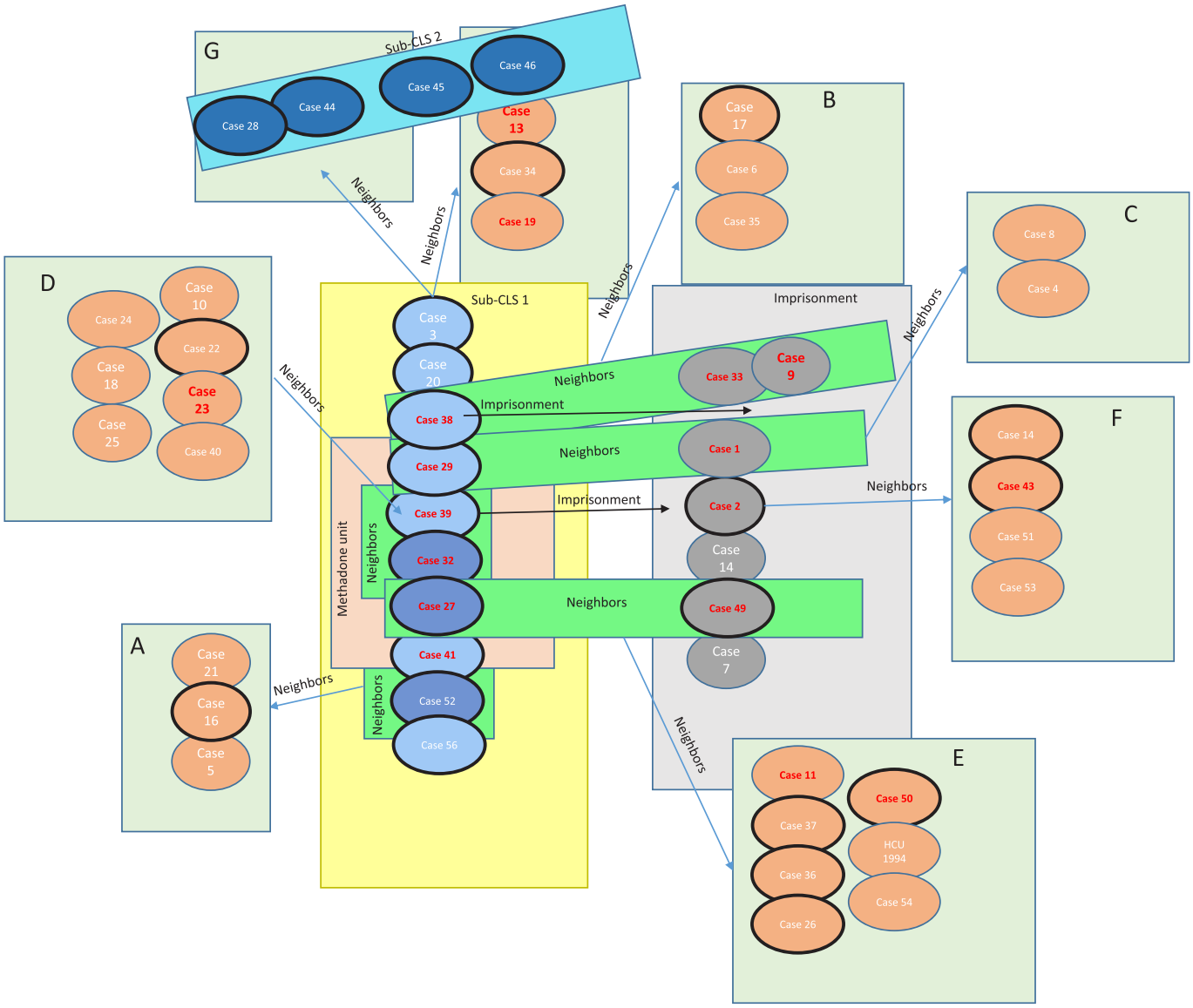


Fig. 4. Confirmed, probable and possible links traced using both genomic and epidemiological information. Letters corresponded to different neighbourhoods of Zaragoza. Within sub-CLS 1, the ones with the SNP in position 2212353 are coloured in light blue, and the ones without the SNP are coloured in dark blue. The ones in red are users or ex-users of ID. The ones with the black circles are the ones sequenced by WGS. Cases without epidemiological or genomic links do not appear in the diagram. The two cases in sub-CLS3 are not grouped together.. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

This could mean that the SNPs occurred in each isolate individually after the transmission, as can be observed in the isolates from sub-CLS 1, or were produced during the *in vitro* steps, as we observed when analysing the contaminated isolate in the laboratory, which was far from being identical to its contaminating isolate, as they differed in seven SNPs (data not shown). It is important to take into account that this strain has been circulating for more than 25 years, so the evolution of the strain has been possible. The SNP dendrogram showed four identical sequences, with no SNPs in relation to the cluster, so we considered them the genome type. Just one of these isolates, Case 2, was from 1993, when our study began. On one hand, we thought that this case could be a possible spreader case due to its epidemiological characteristics. He was a young man who lived near some of the cases that appeared later, and he was an ID user, HIV+ and imprisoned after the TB diagnosis. Besides, it can be observed that Case 2 had recent contact (≤5 SNPs) with 24 of the patients, including those in the methadone unit and those in prison. On the other hand, the fact that the three characteristic SNPs present in sub-CLS 1 were already observed in 1993 suggests that the strain was circulating some years before.

Among the transmitted SNPs, the one in *mce3C* gene in sub-CLS 1 caught our attention for being in a virulence factor gene, so it could be related with the success of this variant. This gene is part of the *mce3* operon [24], which encodes for homologous ATP-binding cassette transporters related to virulence as knock-out strains for this operon were attenuated [25,26]. Virulent *M. bovis* strains lack *mce3* operon, nevertheless this fact has been related with evolutionary differences between *M. tuberculosis* and *M. bovis*, as host range [27]. Recently, Yong Zhang et al. [28] have demonstrated that MCE3C protein is located on the surface of the cell, promoting mycobacterial adherence to macrophages and allowing their entry into them, becoming an important gene in the first steps of infection. We know that seven of the sequenced isolates had this SNP, and there could be more among those unsequenced isolates. This SNP could be conferring an advantage in the ability of the strain to infect macrophages.

A limitation of the study is that there is a period without data (1996–2000); therefore, several cases were missed, as well as the cases

Table 5
Location points, gene names and effect of the SNPs used to construct the dendrogram. In the fourth column, the specific SNPs of each sub-CLS are indicated.

Point	Gene	Effect	CLS SNP
4987	*Rv0004*	Non-synonymous (T185N)	
15198	*trpG*	Synonymous	
79349	Intergenic region	–	
252810	*pckA*	Non-synonymous (F343L)	
267176	*Rv0223c*	Non-synonymous (A197T)	
282798	*Rv0236c*	Non-synonymous (A1352T)	
408860	*Rv0340*	Non-synonymous (A76E)	
455794	*Rv0378*	Non-synonymous (N53T)	
483316	*mmpS1*	Non-synonymous (C114S)	
559337	*fadB2*	Non-synonymous (V148G)	
575274	*Rv0485*	Non-synonymous (D431N)	
581072	*regX3*	Synonymous	
604044	*hemD*	Non-synonymous (Q409P)	
625157	*fabH*	Synonymous	Sub-CLS 2
723561	*recB*	Non-synonymous (A485S)	
815974	*sppA*	Synonymous	
884644	*Rv0790c*	Non-synonymous (R53C)	
1018380	*Rv0913c*	Non-synonymous (P116T)	
1042060	*pstB*	Synonymous	
1052125	*Rv0941c*	Non-synonymous (V65L)	
1072864	*Rv0959*	Non-synonymous (A537V)	
1114864	*Rv0998*	Non-synonymous (L39F)	Sub-CLS 3
1300192	Intergenic region	–	
1498542	*dinG*	Synonymous	Sub-CLS 3
1595376	*uvrC*	Synonymous	
1670243	*moxR1*	Non-synonymous (L321V)	
1680504	*Rv1490*	Synonymous	
1778232	*bioD*	Non-synonymous (R125H)	
1825730	*cydA*	Non-synonymous (F53S)	
1825734	*cydA*	Non-synonymous (K52E)	
1850264	*lysX*	Non-synonymous (R591L)	
1885648	*pks8*	Non-synonymous (E1315D)	
2024764	Intergenic region	–	
2095889	*Rv1845c*	Synonymous	
2212353	*mce3C*	Non-synonymous (A243V). Virulence factor.	Sub-CLS 1
2277103	*Rv2030c*	Non-synonymous (F462L)	
2292579	*Rv2047c*	Non-synonymous (R650C)	Sub-CLS 1
2345187	*pknJ*	Synonymous	
2362885	*helZ*	Synonymous	
2580479	*Rv2308*	Non-synonymous (V21F)	
2637298	Intergenic region	–	
2637385	Intergenic region	–	
2637507	Intergenic region	–	
2692745	Intergenic region	–	
2696235	*cysT*	Non-synonymous (G138V)	
2698983	*Rv2402*	Non-synonymous (Y152C)	Sub-CLS 1
2722351	*Rv2425c*	Non-synonymous (V320L)	
2808813	*pdhC*	Non-synonymous (M376T)	
2838938	*Rv2522c*	Non-synonymous (S202G)	
2866569	*lppA*	Synonymous	
2880258	*Rv2560*	Non-synonymous (P62S)	
2887762	*Rv2566*	Non-synonymous (D464H)	
2950970	*Rv2624c*	Non-synonymous (I113S)	
3076878	Intergenic region	–	
3242032	Intergenic region	–	
3373298	*ligA*	Synonymous	
3406465	*adhC*	Non-synonymous (P61A)	
3506567	*fadE 24*	Non-synonymous (G402V)	
3632058	Intergenic region	*alkB:Rv3252c*	Sub-CLS 2
3,641,091	*fbiA*	Synonymous	
3669602	*usfY*	Non-synonymous (Q133R)	
3698051	*sapM*	Non-synonymous (G285D)	
3784937	*Rv3371*	Synonymous	
3806157	*acrA 1*	Synonymous	
3828309	*choD*	Non-synonymous (R140H)	
3849942	*Rv3431c*	Synonymous	
4014776	*fadE 34*	Synonymous	
4109100	*acs*	Non-synonymous (A437P)	
4133100	*Rv3691*	Non-synonymous (L195I)	
4194065	*ctpJ*	Non-synonymous (L437F)	Sub-CLS 2
4220833	*lipE*	Synonymous	
4232392	*Rv3785*	Non-synonymous (I402M)	
4252725	*fadE 35*	Synonymous	
4255902	*accD4*	Non-synonymous (A16E)	

prior to 1993, as the isolates were not genotyped. Besides, DNA was not available for all the isolates, so the WGS was not applied to all known cases. Some of these cases could be a link we have lost. A major difficulty in the study of TB outbreaks is the latency period. As a person can become infected but not develop the disease until years later [29], it is a challenge to determine when the transmission occurs and who infected whom. This strain proved to be successful in latency, as can be observed in sub-CLS 2. Case 28 developed the disease in 2002, then potentially infected his son, who did not develop the disease until 2008.

5. Conclusion

In conclusion, this work analysed an outbreak caused by an X-strain, which successfully persisted for at least 25 years among our population. We were able to discard four cases included by IS*6110*-RFLP in the Ara50 outbreak that finally were not caused by the Ara50 strain. Besides, it showed TB at its worst, as this strain affected vulnerable people, such as ID users, incarcerated, HIV+ and homeless. The strain did not develop antibiotic resistance, a remarkable fact taking into account the number of relapses and poor adherence to treatment. WGS was useful in the identification of three sub-CLS, one of them with the acquisition of a SNP in a gene considered as a virulence factor. This SNP, together with the 10 SNPs in genes involved in virulence, could explain the successful spread of this strain and its sub-CLS 1 variant, which has co-existed with the Ara50 strain since the beginning of the study in 1993. Finally, although the last registered case was in 2018, we do not rule out the possibility of new cases in the next years, as the strain has demonstrated efficacy in latency.

DNA sequences and GenBank accession numbers

Sequences of Case 2 (Ara50 type) (accession number SAMN15501351) and Case 36 (evolution of Ara50 strain with one more IS*6110* copy) (accession number SAMN15501352) have been uploaded to BioSample database. They are included in the BioProject PRJNA645275.

Acknowledgements

We thank the Working Group on Molecular Epidemiology of Tuberculosis in Aragon and the use of Servicio General de Apoyo a la Investigación-SAI, Universidad de Zaragoza and Servicios Científico Técnicos del CIBA (IACS-Universidad de Zaragoza). We also thank the Instituto de Salud Carlos III and Gobierno de Aragón for the funding and Aragon Health Department for its constant support in this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tube.2020.102022.

Funding

This work was supported by the Instituto de Salud Carlos III (FIS18/0336), and Gobierno de Aragón/Fondo Social Europeo, "Construyendo Europa desde Aragón. Jessica Comín was recipient of a Government of Aragon grant "European Union-Fondo Social Europeo". The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Jessica Comín: Conceptualization, Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization Alberto Cebollada: Software, Formal Analysis, Data Curation Daniel Ibarz: Investigation Jesús Viñuelas: Resources María Asunción Vitoria: Resources María José Iglesias: Resources Sofía Samper: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Funding acquisition.

References

[1] Cremers AJH, Coolen JPM, Bleeker-Rovers CP, van der Geest-Blankert ADJ, Haverkate D, Hendriks H, et al. Surveillance-embedded genomic outbreak resolution of methicillin-susceptible Staphylococcus aureus in a neonatal intensive care unit. Sci Rep 2020;10(1):2619.
[2] World Health. Organization webpage. Available from: https://www.who.int/tb/global-report-2019.
[3] Hatherell HA, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. BMC Med 2016;14:21.
[4] Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A 2002;99(6):3684–9.
[5] Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet 2016;48(12):1535–43.
[6] Gonzalo-Asensio J, Pérez I, Aguiló N, Uranga S, Picó A, Lampreave C, et al. New insights into the transposition mechanisms of IS6110 and its dynamic distribution between Mycobacterium tuberculosis Complex lineages. PLoS Genet 2018;14(4):e1007282.
[7] Manca C, Tsenova L, Barry CE, Bergtold A, Freeman S, Haslett PA, et al. Mycobacterium tuberculosis CDC1551 induces a more vigorous host response in vivo and in vitro, but is not more virulent than other clinical isolates. J Immunol 1999;162(11):6740–6.
[8] López-Calleja AI, Lezcano MA, Vitoria MA, Iglesias MJ, Cebollada A, Lafoz C, et al. Genotyping of Mycobacterium tuberculosis over two periods: a changing scenario for tuberculosis transmission. Int J Tubercul Lung Dis 2007;11(10):1080–6.
[9] Zakham F, Laurent S, Esteves Carreira AL, Corbaz A, Bertelli C, Masserey E, et al. Whole-genome sequencing for rapid, reliable and routine investigation of Mycobacterium tuberculosis transmission in local communities. New Microbes New Infect 2019;13:100582.
[10] Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in london: a retrospective observational study. PLoS Med [Internet 2016;13(10):1–18. https://doi.org/10.1371/journal.pmed.1002137. Available from:.
[11] Lalor MK, Casali N, Walker TM, Anderson LF, Davidson JA, Ratna N, et al. The use of whole-genome sequencing in cluster investigation of a multidrug-resistant tuberculosis outbreak. Eur Respir J [Internet] 2018;51(6):1702313. Available from: http://erj.ersjournals.com/lookup/doi/10.1183/13993003.02313-2017.
[12] van Soolingen D, de Haas PE, Hermans PW, van Embden JD. DNA fingerprinting of Mycobacterium tuberculosis. Methods Enzymol 1994;235:196–205.
[13] Van Embden JDA, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 1993;31(2):406–9.
[14] Kamerbeek J, Schouls L, Kolk A, Van Agterveld M, Van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol 1997;35(4):907–14.
[15] Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. J Clin Microbiol 2006;44(12):4498–510.
[16] Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun 2014;5:4–8.
[17] Forrellad MA, Klepp LI, Gioffré A, García JS, Morbidoni HR, de la Paz Santangelo M, et al. Virulence factors of the mycobacterium tuberculosis complex. Virulence. 1st4. 66: 3; 2013.
[18] Cannas A, Mazzarelli A, Caro A Di, Delogu G, Girardi E. Molecular typing of Mycobacterium tuberculosis strains: a fundamental tool for tuberculosis control and elimination. Infect Dis Rep 2016;8(2):6567.
[19] Millán-Lou MI, Otal I, Monforte ML, Vitoria MA, Revillo MJ, Martín C, et al. Vivo IS6110 profile changes in a Mycobacterium tuberculosis strain as determined by tracking over 14 years. J Clin Microbiol 2015;53(7):2359–61.
[20] Shitikov E, Guliaev A, Bespyatykh J, Malakhova M, Kolchenko S, Smirnov G, et al. The role of IS6110 in micro- and macroevolution of Mycobacterium tuberculosis lineage 2. Mol Phylogenet Evol 2019;139:106559.
[21] Reyes A, Sandoval A, Cubillos-Ruiz A, Varley KE, Hernández-Neuta I, Samper S, et al. IS-seq: a novel high throughput survey of in vivo IS6110 transposition in multiple Mycobacterium tuberculosis genomes. BMC Genom 2012;13:249.
[22] Comin J, Chaure A, Cebollada A, Ibarz D, Viñuelas J, Vitoria MA, et al. Investigation of a rapidly spreading tuberculosis outbreak using whole-genome sequencing. Infect Genet Evol 2020:81.
[23] Schürch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ, et al. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol 2010;48(9):3403–6.
[24] Ahmad S, El-Shazly S, Mustafa AS, Al-Attiyah R. The six mammalian cell entry proteins (Mce3A-F) encoded by the mce3 operon are expressed during in vitro growth of Mycobacterium tuberculosis. Scand J Immunol 2005;62(1):16–24.
[25] Gioffré A, Infante E, Aguilar D, De La Paz Santangelo M, Klepp L, Amadio A, et al. Mutation in mce operons attenuates Mycobacterium tuberculosis virulence. Microb Infect 2005;7(3):325–34.
[26] Senaratne RH, Sidders B, Sequeira P, Saunders G, Dunphy K, Marjanovic O, et al. Mycobacterium tuberculosis strains disrupted in mce3 and mce4 operons are attenuated in mice. J Med Microbiol 2008;57(Pt 2):164–70.
[27] Zumárraga M, Bigi F, Alito A, Romano MI, Cataldi A. A 12.7 kb fragment of the Mycobacterium tuberculosis genome is not present in Mycobacterium bovis. Microbiology 1999;145(Pt 4):893–7.
[28] Zhang Y, Li J, Li B, Wang J, Liu CH. Mycobacterium tuberculosis Mce3C promotes mycobacteria entry into macrophages through activation of β2 integrin-mediated signalling pathway. Cell Microbiol 2018;20(2).
[29] Salgame P, Geadas C, Collins L, Jones-López E, Ellner JJ. Latent tuberculosis infection - revisiting and revising concepts. Tuberculosis 2015;95(4):373–84.

Table S1. Location points, nucleotide changes, gene names and the mutation effect of Ara50 specific SNPs. The points are referred to H37Rv genome.

Point	Change	Gene	Mutation effect
108	C/T	*dnaA*	Synonymous
255919	C/T	*Rv0213c*	Neutral
284403	G/A	*Rv0236c*	Neutral
401678	C/A	*Rv0336*	Neutral
417516	C/T	*Rv0347*	Synonymous
459674	C/A	*clpB*	Neutral
559218	G/A	*fadB2*	Synonymous
597033	C/T	*mmpS2*	Deleterious
633535	G/C	*Rv0541c*	Synonymous
647079	C/T	*menD*	Deleterious
698024	G/A	*Rv0600c*	Synonymous
731417	G/GT	Intergenic region	-
735645	G/A	*rplA*	Synonymous
736873	G/T	*mmaA4*	Neutral
750554	G/A	*Rv0654*	Synonymous
751114	G/C	*Rv0654*	Neutral
765756	A/G	*rpoC*	Neutral
804811	C/T	*rpsC*	Deleterious
900847	A/C	*cpsY*	Neutral
916303	G/T	*Rv0822c*	Neutral
1174496	C/T	*Rv1051c*	Deleterious
1252229	C/T	*Rv1128c*	Synonymous
1348388	G/A	*Rv1204c*	Synonymous
1414145	C/G	*pknH*	Neutral
1439087	C/A	*cysN*	Synonymous
1505464	C/T	*murI*	Synonymous
1639696	C/T	*qor*	Deleterious
1682091	G/GT	Intergenic region	-
2009062	G/C	*Rv1774*	Neutral
2117784	C/G	*Rv1868*	Synonymous
2140945	A/G	*Rv1894c*	Synonymous
2238104	T/A	Intergenic region	-
2283073	G/T	*Rv2037c*	Synonymous
2419009	G/T	*murE*	Neutral
2674765	G/A	*mbtB*	Deleterious
2696712	C/T	*subI*	Neutral
2788092	G/A	*plsB2*	STOP
3011258	G/A	*Rv2693c*	Synonymous
3161596	G/A	*mqo*	Neutral
3260089	C/T	*ppsC*	Synonymous
3285146	C/T	*mmpL7*	Neutral
3418187	C/G	*Rv3057c*	Deleterious
3599381	C/A	*sigH*	Deleterious
3638593	C/T	*pmmA*	Synonymous
3652581	G/A	*ctpC*	Deleterious
3684733	C/G	*atsB*	Deleterious
3785262	T/C	*Rv3371*	Deleterious

3791079	G/A	*Rv3377c*	Synonymous
4020500	C/A	*arsB2*	Deleterious
4047400	T/C	*Rv3604c*	Deleterious
4163903	C/T	*Rv3720*	Synonymous
4264301	G/A	*Rv3802c*	Neutral
4350254	C/T	*Rv3871*	Synonymous
4356831	A/C	*Rv3878*	Neutral
4379359	A/G	*Rv3894c*	Synonymous
4401018	C/G	*Rv3912*	Deleterious
4403914	G/A	*Rv3915*	Synonymous

Figure S1. Phylogenetic tree based on the genomes including CDC1551, Ara50, X non-Ara50 and H37Rv strains. X non-Ara50 strain seemed to be more related with CDC1551 than Ara50 strain. H37Rv appeared as an outgroup.

Publicación 6

frontiers | Frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 24 May 2022
doi: 10.3389/fcimb.2022.887134

The MtZ Strain: Molecular Characteristics and Outbreak Investigation of the Most Successful *Mycobacterium tuberculosis* Strain in Aragon Using Whole-Genome Sequencing

Jessica Comín[1], Jan Madacki[2], Isabel Rabanaque[3,4,5], María Zúñiga-Antón[3,4,5], Daniel Ibarz[6], Alberto Cebollada[7], Jesús Viñuelas[8,9], Luis Torres[10], Juan Sahagún[11], Christophe Klopp[12], Jesús Gonzalo-Asensio[6], Roland Brosch[2], María-José Iglesias[5,6,13] and Sofía Samper[1,5,13]*

[1] Grupo de Genética de Micobacterias, Instituto Aragonés de Ciencias de la Salud, Zaragoza, Spain, [2] Unit for Integrated Mycobacterial Pathogenomics, Institut Pasteur, Université de Paris, CNRS UMR 3525, Paris, France, [3] Departamento de Geografía y Ordenación del Territorio, Universidad de Zaragoza, Zaragoza, Spain, [4] Instituto Universitario de Investigación en Ciencias Ambientales de Aragón, Zaragoza, Spain, [5] Fundación Instituto de Investigación Sanitaria (IIS) Aragón, Zaragoza, Spain, [6] Grupo de Genética de Micobacterias, Facultad de Medicina, Universidad de Zaragoza, Zaragoza, Spain, [7] Unidad de Biocomputación, Instituto Aragonés de Ciencias de la Salud, Zaragoza, Spain, [8] Hospital Universitario Miguel Servet, Zaragoza, Spain, [9] Grupo de Estudio de Infecciones por Micobacterias (GEIM), Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica, Madrid, Spain, [10] Hospital San Jorge, Huesca, Spain, [11] Hospital Clínico Universitario Lozano Blesa, Zaragoza, Spain, [12] MIAT INRA, Castanet-Tolosan, France, [13] Centro de Investigación Biomédica en Red (CIBER) de Enfermedades Respiratorias, Madrid, Spain*

Since 2004, a tuberculosis surveillance protocol has been carried out in Aragon, thereby managing to detect all tuberculosis outbreaks that take place in the community. The largest outbreak was caused by a strain named *Mycobacterium tuberculosis Zaragoza* (MtZ), causing 242 cases as of 2020. The main objective of this work was to analyze this outbreak and the molecular characteristics of this successful strain that could be related to its greater transmission. To do this, we first applied whole-genome sequencing to 57 of the isolates. This revealed two principal transmission clusters and six subclusters arising from them. The MtZ strain belongs to L4.8 and had eight specific single nucleotide polymorphisms (SNPs) in genes considered to be virulence factors [*ptpA*, *mc3D*, *mc3F*, *VapB41*, *pks15* (two SNPs), *virS*, and *VapC50*]. Second, a transcriptomic study was carried out to better understand the multiple IS*6110* copies present in its genome. This allowed us to observe three effects of IS*6110*: the disruption of the gene in which the IS*6110* is inserted (*desA3*), the overexpression of a gene (*ppe38*), and the absence of transcription of genes (*cut1:Rv1765c*) due to the recombination of two IS*6110* copies. Finally, because of the disruption of *ppe38* and *ppe71* genes by an IS*6110*, a study of PE_PGRS secretion was carried out, showing that MtZ secretes these factors in higher

OPEN ACCESS

Edited by:
Alexandra Aubry, Sorbonne Universités, France

Reviewed by:
Isabelle Bonnet, Hôpitaux Universitaires Pitié Salpêtrière, France
Kaixia Mi, Institute of Microbiology (CAS), China

***Correspondence:**
Jessica Comín
jcomin.iacs@aragon.es

Specialty section:
This article was submitted to Molecular Bacterial Pathogenesis, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: *01 March 2022*
Accepted: *11 April 2022*
Published: *24 May 2022*

Citation:
Comín J, Madacki J, Rabanaque I, Zúñiga-Antón M, Ibarz D, Cebollada A, Viñuelas J, Torres L, Sahagún J, Klopp C, Gonzalo-Asensio J, Brosch R, Iglesias M-J and Samper S (2022) The MtZ Strain: Molecular Characteristics and Outbreak Investigation of the Most Successful Mycobacterium tuberculosis Strain in Aragon Using Whole-Genome Sequencing. Front. Cell. Infect. Microbiol. 12:887134. doi: 10.3389/fcimb.2022.887134

amounts than the reference strain, thereby differing from the hypervirulent phenotype described for the Beijing strains. In conclusion, MtZ consists of several SNPs in genes related to virulence, pathogenesis, and survival, as well as other genomic polymorphisms, which may be implicated in its success among our population.

Keywords: tuberculosis outbreak, molecular epidemiology, tuberculosis, WGS, tuberculosis virulence

INTRODUCTION

Airborne tuberculosis (TB) has existed alongside humanity since the beginning of civilization, being first recorded among Egyptian mummies dating back to 2,400 BC (Morse et al., 1964). In 2019, an estimated 10 million people fell ill with TB, causing 1.2 million deaths (World Health Organization, 2020). While COVID-19 surpassed TB as the world's leading deadly infectious disease in 2020, TB remains second. *Mycobacterium tuberculosis* (MTB) is the causative pathogen, which usually produces respiratory disease but can also cause extrapulmonary disease (García-Rodríguez et al., 2011).

Traditionally, restriction fragment length polymorphism (RFLP), spoligotyping, and mycobacterial interspersed repetitive units-variable number of tandem repeats (MIRU-VNTR) have been used to investigate outbreaks due to its value in interpreting transmission dynamics (Borgdorff et al., 2000; Gonzalo-Asensio et al., 2018; Iglesias et al., 2020). The development of whole-genome sequencing (WGS) technologies marked a milestone in outbreak investigation, as it offers unsurpassed resolution power (Nikolayevskyy et al., 2019). Moreover, WGS has become an affordable technique, thus being proposed as a replacement for previous molecular typing techniques (Cirillo et al., 2016). However, its potential to elucidate the direction of transmission is still uncertain (Hatherell et al., 2016).

A TB molecular surveillance program has been carried out in the Autonomous Community of Aragon, in the north of Spain, since 2004. All MTB cases were genotyped using IS*6110*-RFLP and spoligotyping, and recently, WGS has been applied to relevant isolates. In a previous study for 2001–2004, a cluster of 85 cases was discovered (López-Calleja et al., 2007), the largest of the period. Subsequently, this strain was named *M. tuberculosis Zaragoza* (MtZ) and was classified as belonging to the principal genetic group 3 (López-Calleja et al., 2009). A study of the MtZ strain in terms of fitness demonstrated its high virulence in a low-dose aerosol and intravenous *in-vivo* infection model (Caceres et al., 2012). Some epidemiological links were found among several cases, but the virulence mechanisms that allowed this strain to spread affecting 242 cases until today were unknown.

The aims of this work were to complete the study of this outbreak, the largest in our region since our records began in 1993, by means of WGS and to determine the molecular characteristics that could be responsible for the diffusion of the MtZ strain.

MATERIALS AND METHODS

Clinical Samples and Cases

Aragon has three capital provinces, including Zaragoza province where almost three-quarters of the Aragonese population lives, the majority of them in Zaragoza City. The microbiologists working in the mycobacterial units were the professionals in charge of encoding the tuberculosis strains (strain code) that arrived at the microbiology laboratories of Aragon's hospitals. These microbiologists carried out the drug susceptibility tests for all the isolates in their laboratory routine. In our laboratory, a nurse determined the new TB cases and assigned the four-digit code for each case. It is necessary to have controlled the repeated cases. The strain code was the key field linked to the necessary variables for the study. No person from our research laboratory was involved in this process. All data remained anonymous. Our regional ethical committee (Comité de Ética de la Investigación de la Comunidad Autónoma de Aragón, Record No. 20/2018) approved this methodology, detailed in the 18/0336 project.

Since 2004, all *M. tuberculosis* isolates have been genotyped as part of a surveillance protocol, and the DNA remains frozen at −20°C. All MtZ cases were selected based on the identical IS*6110*-RFLP pattern (12 bands) of their isolates, analyzed by BioNumerics software (v7.6, Applied Maths, Kortrijk, Belgium). There were some exceptions, previously described by Millán-Lou et al. in 2015 (Isabel Millan-Lou et al., 2013) as evolved isolates with an extra or a lost band. Two cases were identified in Aragon from 1993 to 1995, and 240 cases were identified from January 2001 to December 2020. From all the samples with available DNA, 57 were sequenced using WGS: one from 1995 (considered case 0 in this study), 15 from 2001, two from 2002, two from 2003, three from 2004, three from 2006, three from 2007, two from 2009, four from 2010, three from 2011, three from 2012, one from 2013, two from 2015, two from 2016, three from 2017, five from 2018, and three from 2019.

DNA Extraction and Genotyping

The DNA of the MtZ isolates, stored at −20°C until sequencing, was extracted from bacterial growth in solid media cultures at the time of diagnosis of the cases using the cetrimonium bromide method previously described (van Soolingen et al., 1994). IS*6110*-RFLP and spoligotyping were performed for all the isolates as previously described (Van Embden et al., 1993; Kamerbeek et al., 1997).

WGS

Illumina sequencing (Nextera Flex, San Diego, USA) was applied using the manufacturer's instructions for 43 samples. Ion Torrent technology was used on 14 samples according to the manufacturer's instructions. After sequencing, millions of sequences contained in the fastQ files were mapped against the reference strain H37Rv (NC_000962.3) in order to obtain the Binary Aligned Map (BAM) and Variant Call Format (VCF) files, used for the single nucleotide polymorphism (SNP) study. The depth of the sequences was at least 30× and the average coverage was around 98%.

Lineage Identification Using WGS

For lineage identification, the SNP classification established by Coll et al. in 2014 (Coll et al., 2014) was applied. This classification associates one strain to one lineage based on several representative SNPs of each TB lineage.

Cluster Classification Using PCR

We designed three pairs of primers to be able to identify which main cluster the non-sequenced isolates of interest belonged to: deoA-F (CGACAAGTCACCTCCCTG) and deoA-R (GAGATGAACTGCCCGCT) for the SNP at point 3702632 (original CLS), dacB1-F (GCGCACCTGCTCGACTAC) and dacB1-R (CTAGTGCTGCGGCCG) for the SNP at point 3716874 (CLS-2), and PE12-F (TGGCCGTTTCGATATTGG) and PE12-R (GTCGAACCGTGGGTGC) for the SNP at point 1301938 (CLS-1).

Bioinformatics for the Study of the Genomes

From the fastQ files obtained after sequencing, the BAM and VCF files were produced. These files were studied using the Integrative Genomics Viewer software [IGV, from the Broad Institute (Robinson et al., 2011)] to search antibiotic resistances and to compare the MtZ strain with other L4.8 strains we had sequenced for another study. Tuberculist (http://genolist.pasteur.fr/TubercuList/), Mycobrowser (https://mycobrowser.epfl.ch/) and UniProtKB (https://www.uniprot.org/uniprot/) websites were used to find information about the genes and proteins with interesting SNPs. We used GeneWise (https://www.ebi.ac.uk/Tools/psa/genewise/) to determine if the SNP was synonymous or non-synonymous and PROVEAN (http://provean.jcvi.org/seq_submit.php) to determine the effect of the non-synonymous SNPs (neutral or deleterious).

The fastQ files were also uploaded into BioNumerics software, where the phylogenetic trees were constructed using the UPGMA method. The program mapped the sequences against H37Rv and found the SNPs between the different isolates. SNPs in genes related to resistance were also determined. For greater accuracy, strict SNP filtering that removed positions with at least one ambiguous or unreliable base, gaps (maximum frequency 1%), non-discriminatory positions, and *ppe* and *pgrs* genes was applied. The retained SNP positions had a minimum of 5× coverage and the minimum distance between SNPs was at least 12 base pairs (bp). As the program was not able to compare sequences obtained by different sequencing platforms, SNPs were eye reviewed in both different phylogenetic trees constructed for the Illumina and Ion Torrent sequences to detect the shared SNPs.

Geodatabase and Map Design

The location (street and number) is available for each MtZ case, and therefore, its assignment to a basic health zone (BHZ) is possible. This allowed the development of a geodatabase that would facilitate cross-referencing with sociodemographic data and its analysis. With this aim, we designed two maps to draw the spatial distribution of TB cases in the city of Zaragoza from 2001 to 2020 together with the two cases identified in the 1993–1995 period. Both maps used the BHZ of the city of Zaragoza as a spatial base, represented by polygons in a vector model of 133 elements. The coordinate system was ETRS 1989 UTM, Zone 30N. Cartography was implemented in a Geographic Information System (ESRI ArcGIS 10.7 software), where we prepared the base map and built the geodatabase. The first map had a multivariate design representing two variables: the number of cases and the TB rate per 100,000 population, calculated using the total number of cases due to the MtZ strain in each neighborhood taking into account the population of each BHZ in 2019. This was represented by color hue and lightness, and the number of cases was represented by size. Establishing spatial patterns in the first variable (TB rate) was based on a sequential scheme (yellow to brown) suited to ordered data that progress from high to low. The number of cases was represented using a graduated symbol system, where data were sorted into ranges to which a symbol size was assigned to represent the class mark. The second map showed the distribution of cases by CLS using a binary color scheme (orange–gray) showing nominal differences divided into two categories (presence–absence).

Transcriptomic Analysis

With the objective of analyzing the effect of the IS*6110* in the *M. tuberculosis* genome, we studied three different MtZ isolates having the same genome background but carrying one different IS copy (case 0, the original strain with 12 IS*6110* bands; case 129, with an extra IS*6110* located in *dnaA*:*dnaN*; and case 241, with one less IS*6110* due to recombination around the DR region). They were cultured and sent to the STAB-VIDA Company (Caparica, Portugal, https://www.stabvida.com) for transcriptomic analysis. The library construction of cDNA molecules was carried out using a Ribosomal Depletion Library Preparation Kit. The generated DNA fragments were sequenced in the Illumina NovaSeq platform, using 150-bp paired-end sequencing reads. The analysis of the generated raw sequence data was carried out using CLC Genomics Workbench 12.0.3. The high-quality sequencing reads were mapped against the reference genome *M. tuberculosis* H37Rv (NC_000962.3). For the transcriptomic study, we used Integrative Genome Browser (IGB) software (Freese et al., 2016).

PE_PGRS Secretion Analysis

One isolate of MtZ (MS 387, case 0) and H37Rv as a control were grown for protein extraction in 7H9-0.05% Tween-80 supplemented with dextrose, NaCl, and catalase to avoid albumin contamination in the secreted fraction until an optical density of 0.6–0.8 at 37°C was reached and then were pelleted. Supernatant fraction and whole-cell protein were extracted following the protocol described by Pérez et al. (2020). Secretion analysis was performed as described previously (Ates et al., 2018; Madacki et al., 2021). Briefly, the whole-cell protein and supernatant fractions were loaded on a NuPage 10% Bis-Tris gel (Thermo Fisher, Waltham, USA), followed by a dry Western blot transfer onto a nitrocellulose membrane (iBlot, Thermo Fisher). The membrane was incubated with anti-PGRS primary antibody 7C4.1F7 (1:2,000) (the antibody-producing clone was a gift from M. J. Brennan, Aeras, Rockville, MD, USA, and the purified antibody was a gift from W. Bitter, Amsterdam UMC, Amsterdam, the Netherlands) and a horseradish peroxidase (HRP)-conjugated IgG anti-mouse secondary antibody (Amersham) (1:5,000). As a loading control, anti-SigA (1:5,000) primary antibody (a gift from I. Rosenkrands, Statens Serum Institut, Copenhagen, Denmark) followed by HRP-conjugated IgG anti-rabbit secondary antibody (Amersham) (1:5,000) was used.

RESULTS

The MtZ strain was first observed in an epidemiological/genomic study (López-Calleja et al., 2007) carried out in the Aragon region, which produced the largest outbreak ever recorded there. The RFLP pattern showed 12 bands and a unique spoligotype, defined as unknown in the SITVIT2 database (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/description.jsp). Its IS*6110* locations were studied by Millán-Lou et al. in 2013 (Isabel Millan-Lou et al., 2013). At that time, WGS was not developed enough for routine application in laboratories; therefore, the molecular characterization was not exhaustively studied. Now, we can observe its molecular properties in much more detail in order to find possible explanations for the high incidence of this strain in the local population of Aragon.

MtZ Genomic Characterization

Using the phylogenetic SNP classification established by Coll et al. in 2014 (Coll et al., 2014), we concluded that the MtZ strain belongs to lineage L4.8 because it has the SNPs in positions 931123 (T/C, *lpq*Q) and 3836739 (G/A, *groEL1*). In total, the MtZ strain has 194 SNPs compared with the H37Rv genome. Twenty-three of them were also present in other L4.8 strains we had sequenced, so we considered that 171 SNPs were specific for the MtZ outbreak strain (**Table S1**).

According to Forrellad et al. and Ramage et al. (Ramage et al., 2009; Forrellad et al., 2013), MtZ isolates had eight non-synonymous SNPs in genes considered as potential virulence factors: seven were specific for the outbreak strain and one was also present in the other L4.8 strains studied (**Table 1**). The ones specific for the strain were in *mce3D* and *mce3F* genes, encoding proteins of the Mce3 cluster, which was shown to be important for virulence in a murine model (Senaratne et al., 2008), but which at the same time is absent from all virulent *Mycobacterium bovis* strains due to the RD7 deletion (Brosch et al., 2002); the *Rv2601A* gene (antitoxin VapB41); the *pks15* gene (two SNPs), whereby the *pks15* gene is likely already a pseudogene in L4 strains due to the frameshift mutation that derives *pks15* and *pks1*, which are corresponding to a functional gene involved in phenolphthiocerol synthesis in L2 strains (Constant et al., 2002); the *virS* gene, part of a virulence operon which controls phagosome–lysosome fusion, acid stress response inside the macrophages, and the expression of enzymes, cell wall and envelope proteins, efflux pumps, ionic transporters, and transcriptional regulators (Singh et al., 2019); and the *Rv3749c* gene (toxin VapC50). The SNP also present in other L4.8 strains was in the *ptpA* gene, involved in host–pathogen interaction and interfering with vesicular trafficking in the macrophage.

We also checked the non-synonymous SNPs in genes that Mycobrowser classified in some of these categories: "required for survival in primary murine macrophages" (Rengarajan et al., 2005), "required for growth in C57BL/6J mouse spleen" (Sassetti et al., 2003), and "disruption of this gene provides a growth advantage for *in vitro* growth" (DeJesus et al., 2017). The first two are related to infection ability, so we grouped both. MtZ has SNPs in *Rv0020c*, *Rv0101*, *Rv0187*, *Rv0193c*, *Rv1145*, *Rv1304*, *Rv1337*, *Rv1844c*, *Rv2113*, *Rv2113*, *Rv2411c*, *Rv2702*, *Rv3234c*, and *Rv3701c*. Information about these genes can be found in **Table 1**.

Any of the sequenced isolates had SNPs in genes related to first-line drug resistance; however, case 124 was the only one showing molecular resistance to quinolones (*gyrA*, D94G). Among the non-sequenced isolates, case 31 showed phenotypic resistance to rifampicin, case 174 to ethambutol, case 181 to isoniazid and ethambutol, and case 239 to isoniazid. The rest of the isolates were susceptible to all antituberculosis drugs.

Epidemiological Study

Until December 2020, 242 cases of MtZ distributed throughout the region were detected. The number of cases per year can be found in **Figure 1**. The highest number of cases corresponded to the first decade of the 2000s, with a special rise in 2004. In 2009, there was a drop in the number of cases, which has continued to date. Only two cases of MtZ were identified for 1993–1995, both presenting an extrapulmonary form of the disease.

Focusing on the epidemiological characteristics of the population affected by this strain, the more affected group of age was between 31 and 45 years (92 cases), followed by the 16–30 group (66 cases). The majority of the cases were autochthonous, with only 27 being foreign-born (coming from 17 different countries). Sixty percent of cases had a positive smear result (BK+), and 53 cases had an X-ray result as cavitated pathology and 80 as non-cavitated pathology. More detailed information can be found in **Table 2**.

Two hundred and five cases lived in Zaragoza City, the community capital. The rest of the cases lived in smaller cities or villages. The cases were distributed among 40 health zones, and the map related to the TB rate in Zaragoza showed that the

TABLE 1 | Mutations in genes related to virulence and transmission.

Nucleotide locus	Reference	Variant	Amino acid change	Gene	Functional category	Gene product
2213395	A	G	I181V	*Rv1969*	Virulence, detoxification, adaptation	Mce-family protein Mce3D
2216011–2216012	TG	T	Reading frame alteration	*Rv1971*	Virulence, detoxification, adaptation	Mce-family protein Mce3F
2930254	G	A	G62D	*Rv2601A*	Virulence, detoxification, adaptation	Antitoxin VapB41
3296843	A	G	R290L	*Rv2947c*	Lipid metabolism	Polyketide synthase Pks15
3296972	C	A	V333A	*Rv2947c*	Lipid metabolism	Polyketide synthase Pks15
3447666	T	C	E254G	*Rv3082c*	Virulence, detoxification, adaptation	Virulence-regulating transcriptional regulator VirS
4197895	C	G	K81N	*Rv3749c*	Conserved hypotheticals	Toxin VapC50
2507254*	G	A	A37T	*Rv2234*	Regulatory proteins	Phosphotyrosine protein phosphatase PtpA
24445	C	G	V334L	*Rv0020c*	Regulatory proteins	Conserved protein with FHA domain, FhaA
113371	A	G	H1124R	*Rv0101*	Lipid metabolism	Peptide synthetase Nrp
1460992	A	C	H250P	*Rv1304*	Intermediary metabolism and respiration	ATP synthase a chain AtpB
219104-219105	GA	G	Reading frame alteration	*Rv0187*	Intermediary metabolism and respiration	O-methyltransferase
3017134*	G	T	A93S	*Rv2702*	Intermediary metabolism and respiration	Polyphosphate glucokinase PpgK
4144477	G	A	T147I	*Rv3701c*	Conserved hypotheticals	Conserved hypothetical protein
226475	C	T	A33T	*Rv0193c*	Conserved hypotheticals	Hypothetical protein
1273071*	T	C	S217P	*Rv1145*	Cell wall and cell processes	Conserved transmembrane transport protein MmpL13a
1504420	G	A	R22H	*Rv1337*	Cell wall and cell processes	Integral membrane protein
2094479	T	C	D237G	*Rv1844c*	Intermediary metabolism and respiration	6-Phosphogluconate dehydrogenase Gnd1
2373539	C	T	P304S	*Rv2113*	Cell wall and cell processes	Integral membrane protein
2373696	A	ACCG	Reading frame alteration	*Rv2113*	Cell wall and cell processes	Integral membrane protein
2709030	G	T	F312L	*Rv2411c*	Conserved hypotheticals	Conserved hypothetical protein
3610610	G	A	Q194 STOP	*Rv3234c*	Lipid metabolism	Triacylglycerol synthase Tgs3

Following the functional classification used in Mycobrowser, the SNPs in genes considered to be virulence factors are colored in orange. In green are genes classified as "required for survival in primary murine macrophages" and "required for growth in C57BL/6J mouse spleen." In blue are genes classified as "disruption of this gene provides a growth advantage for in vitro growth".

**Present also in other L4.8 strains. All the points refer to the H37Rv genome.*

highest concentrations were in the city center (San Pablo, Independencia, Rebolería) and the outer edges (Utebo, Torrero) (**Figure 2**). In general, higher rates corresponded to socioeconomically vulnerable areas (San Pablo, Rebolería, Las Fuentes, Torrero), with one exception (Independencia). There was spatial contiguity between the areas with the highest values and basic spatial patterns can be seen: concentric behavioral patterns (center, intermediate pattern, peripheral pattern). Moreover, a map locating all the cases was constructed in order to study possible links related to neighborhood contacts (data not shown).

Some epidemiological links were found in the previous study of López-Calleja et al. in 2009 (López-Calleja et al., 2009): nine cases (cases 76, 78, 79, 80, 81, 82, 83, 84, and 85) were part of a

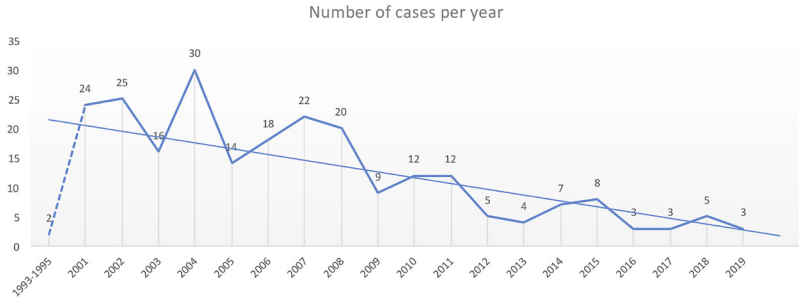


FIGURE 1 | Number of cases of the *Mycobacterium tuberculosis Zaragoza* (MtZ) strain during the study period. The two cases in 1993–1995 were identified after the discovery of the strain in the 2001–2004 study. The highest number of cases was in 2004, with 30 cases, and the number continued to be high during the first decade of the 2000s. The number of cases diminished in 2009, and in the last few years, it has been quite low. No cases were diagnosed in 2020.

TABLE 2 | Characteristics of the 242 TB cases involved in the MtZ outbreak.

	N (%)
Age (years)	
0–15	17 (7.0)
16–30	66 (27.3)
31–45	92 (38.0)
46–60	49 (20.2)
>60	18 (7.4)
Intravenous drug users	
Yes	17 (7.0)
No	124 (51.2)
Unknown	101 (41.8)
Smokers	
Yes	92 (38.0)
No	40 (16.5)
Unknown	110 (45.5)
HIV status	
Positive	24 (9.9)
Negative	152 (62.8)
Unknown	66 (27.3)
Alcohol consumption	
High	43 (17.8)
Moderate	11 (4.5)
Low/no	79 (32.7)
Unknown	109 (45.0)
X-ray result	
Normal	2 (0.8)
Cavitated pathology	53 (21.9)
Non-cavitated pathology	80 (33.1)
Unknown	107 (44.2)
Bacilloscopic (BK) result	
Positive	145 (59.9)
Negative	90 (37.2)
Unknown	7 (2.9)

nursery outbreak, involving eight babies and their carer. Case 3 was the uncle of case 10 and the cousin of case 13. Cases 5 and 57 were father and son, as were cases 40 and 46. Cases 70 and 71 were married, as were cases 16 and 20. Cases 54 and 95 lived in the same building, but no specific link could be established between them. The same occurred for cases 38 and 48. Four cases (cases 6, 32, 45, and 55) were in prison. New epidemiological investigations revealed that cases 115 and 126 were brother and sister, and case 115 was a workmate of case 129. Cases 122 and 140 lived in the same building, and cases 221 and 222 were married.

Genomic Study

The sequence analysis of the genomes was carried out in BioNumerics software. The data are summarized in **Figure 3** for simplicity. Only two isolates could be considered as the original sequence of the strain (case 0, belonging to the period of 1993–1995, and case 7 from 2001). The rest of the isolates had an SNP in position 3702632; therefore, there had to be an intermediate isolate that gained this mutation that was not sequenced (case x). From this intermediate isolate, two branches appeared. Several cases conserved the original base at position 3716874 (CLS-1), while the rest of the isolates had an SNP in that position (CLS-2) (**Figure 3**).

In addition, isolates from CLS-1 also shared an SNP at position 1301938, and two of the isolates shared one more SNP at position 2544858, indicating transmission between these two cases (CLS-1.1). There were no epidemiological links among cases of CLS-1, except that three of them lived in the area of Huesca, one of the provincial capitals. A few more cases, not sequenced, also lived in this area. We confirmed by PCR that four of these cases had the characteristic SNPs of CLS-1. One sequenced case who lived in this area was classified in CLS-2, therefore not sharing these SNPs.

Eight cases of CLS-2 were considered the original sequence of this cluster, as they only shared the SNPs in positions 3702632 and 3716874. The majority of these cases were diagnosed in 2001, at the beginning of the outbreak. Twenty-four more cases shared those SNPs, having some other unique SNPs, not transmitted to other isolates; therefore, 32 isolates could be considered as part of the CLS-2 original sequence. From CLS-2, five more subclusters appeared. These isolates conserved the two characteristic SNPs of CLS-2 together with independent SNPs, resulting in five new transmission sub-CLSs. The genetic and epidemiological links among cases of CLS-2 can be found in **Figure 4**.

Isolates of CLS-2.1 shared the SNPs at positions 862566 and 3143468. In addition, they had an extra IS*6110* inserted in *dnaA:dnaN*. Only two cases among the sequenced isolates had those SNPs. However, 10 more cases could be included in this cluster after the study of IS*6110* in *dnaA:dnaN*. Moreover, the majority of CLS-2.1 cases lived in the same neighborhood.

Isolates of CLS-2.2 shared three SNPs at positions 898986, 2770751, and 3070366. Four sequenced isolates belonged to this cluster. Nine cases belonged to CLS-2.3, sharing an SNP at position 3081980, and two isolates constituted CLS-2.4 sharing the SNPs at positions 551 and 2180630. Finally, CLS-2.5 was formed by two isolates that shared the SNP at position 46642. Additional information about these transmitted SNPs in the different subclusters can be found in **Table 3**.

A strain was considered to belong to a cluster when there were ≤12 SNPs between at least two isolates of the potential outbreak. If there were ≤5, it was considered recent contact (Lalor et al., 2018). A table with the genomic distance among the different isolates of the MtZ outbreak was constructed (**Table S2**). All the isolates had ≤12 SNPs with at least one isolate, meaning all of them belonged to the MtZ outbreak, and 50 isolates had recent contact (≤5 SNPs) with at least one isolate.

We have followed the classification established by Lalor et al. in 2018 for the different links among the outbreak cases (Lalor et al., 2018). For the MtZ outbreak, we found several "confirmed epidemiological links" as relatives, husband and wife, cohabitants, work contact, and the nursery cluster. "Probable epidemiological links" could be established among the cases in prison and cases who live in the same building, as they spent time in the same location, but the timing was unknown. Finally, possible epidemiological links could be established among cases who lived in the same area. These detailed links can be found in **Figure 4**.

The geographical distribution of the different MtZ CLSs in Zaragoza City is shown in **Figure 5**. The TB case distribution map reveals an uneven distribution by CLS. The CLS with the highest presence in the city was CLS-2 (17 areas), followed by

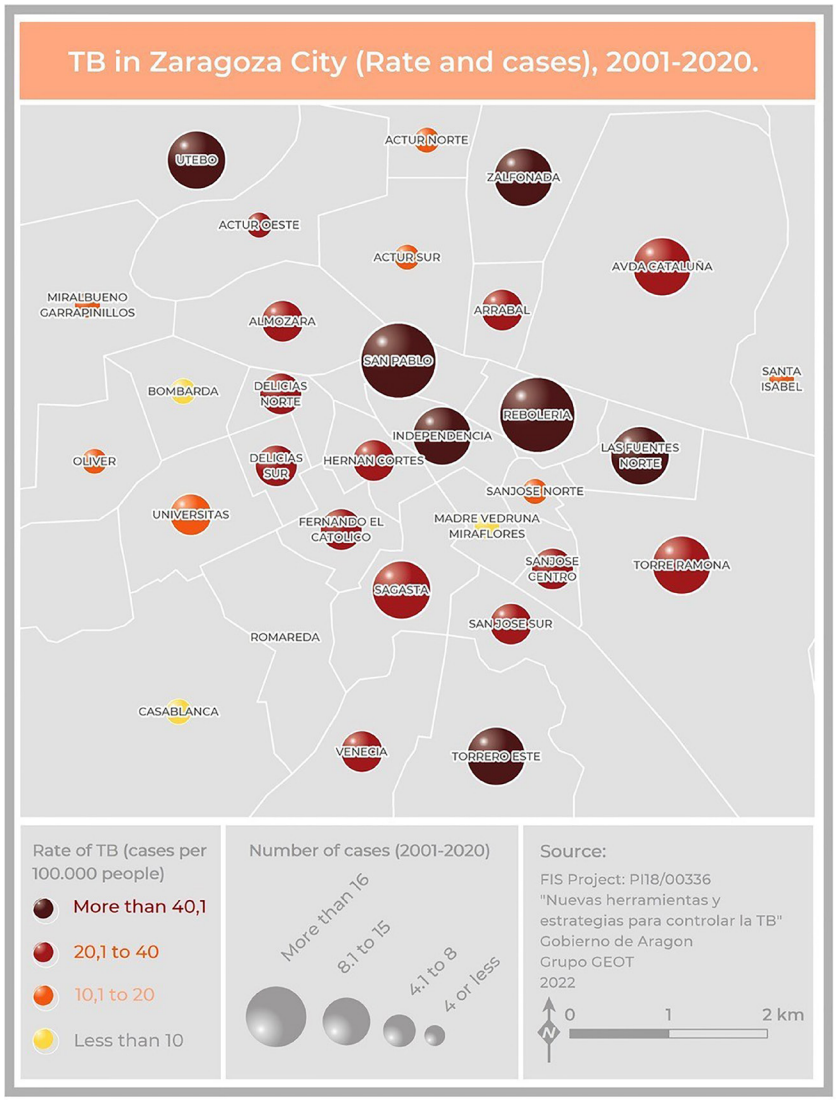


FIGURE 2 | Map of Zaragoza City showing the number of cases and the incidence of TB due to the MtZ strain in the different neighborhoods. The map represents two variables: TB incidence per 100,000 population, represented by color hue and lightness (yellow to brown), and number of cases, represented by size. San Pablo and Rebolería, in the old town, are the neighborhoods with a higher incidence and number of cases. Cases detected outside the city were not included.

CLS-2.1 (six areas) and CLS-2.3 (six areas). In some CLSs, there were no cases in the city at all (CLS-1.1).

RNAseq and PE_PGRS Secretion Studies

In order to analyze the effect of IS*6110* in the adjacent genes of the genome of the MtZ strain, we performed a transcriptomic study using three different MtZ isolates: case 0, considered the original isolate with 12 IS*6110* copies; case 129, which belonged to a CLS-2.1 isolate with an extra IS*6110* in *dnaA:dnaN*; and case 241, which lost one IS*6110* copy due to recombination between two of them around the DR region. The study was carried out in both exponential and stationary growth phases. Relevant findings can be found in **Figure 6**.

The transcriptional analysis allowed us to observe three different effects of IS*6110* for the three MtZ isolates in both growth phases. First, transcription was interrupted in the *desA3* gene upon reaching the IS*6110* insertion site (3606310) while continuing in the H37Rv reference strain (**Figure 6A**). Second, overexpression of the *ppe38* gene in the exponential growth phase, where the MtZ strain had an IS*6110* inserted at point 2633977, was observed (**Figure 6B**). Finally, the transcriptomic study confirmed the lack of transcription (zero reads) of all the genes lost in the region *cut1:Rv1765c* due to the recombination of two close IS*6110* (**Figure 6C**).

Focusing on the different IS*6110* copies among the three isolates analyzed, nothing different was observed for the *dnaA:dnaN* region in case 129 compared with the reference strain or

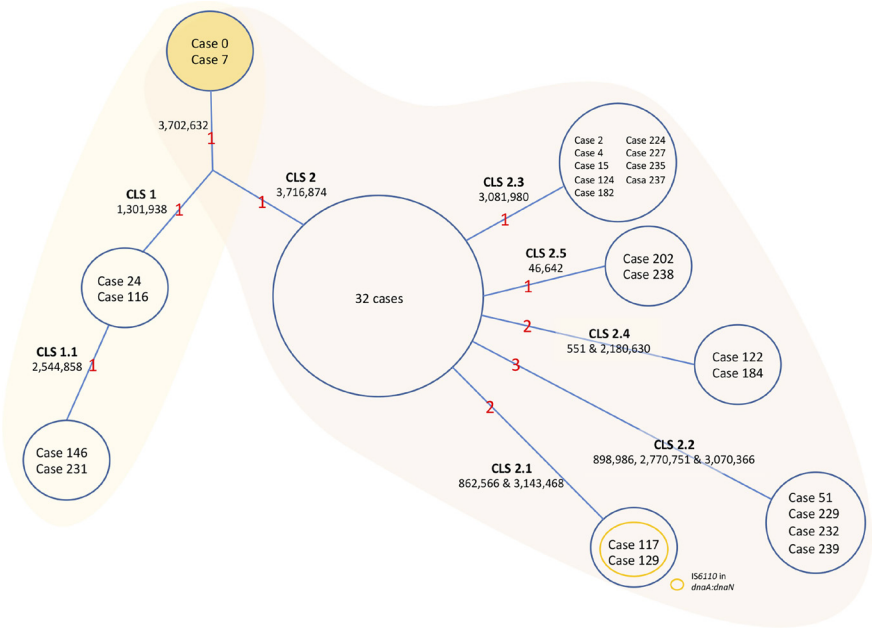


FIGURE 3 | Evolution of the MtZ strain. One acquired SNP at point 3702632 separates Cases 0 and 7, who share the original base, from the rest of the sequenced cases. A Case x must exist, though not sequenced, from which CLS-1 and CLS-2 emerged. CLS-1 emerged after the acquisition of a single nucleotide polymorphism (SNP) at position 1301938. A sub-CLS (CLS-1.1) emerged from this after the acquisition of another SNP at point 2544858. CLS-1 was mostly located in the Huesca area. We could classify cases 65, 113, 139, and 216 in this CLS-1 by PCR. Formerly, an SNP acquired from the hypothetical case x at point 3716874 led to CLS-2, in which the majority of MtZ cases are included. The red numbers are the number of acquired SNPs. The cases appearing in the figure are the ones sequenced. The unique SNPs were excluded from the figure to simplify the data.

with the other two MtZ isolates. For the case 241 isolate, the adjacent genes to the DR region had no transcription, confirming the deletion of genes *Rv2816c* to *Rv2823c*.

Analyzing the transcriptional areas of IS*6110* using H37Rv locations, we observed higher transcription peaks for the MtZ isolates compared with the H37Rv reference strain, being even higher in the stationary phase, thus meaning that IS*6110* was overexpressed in the MtZ isolates (**Figure 6C**). No major differences were found among the transcriptomes of the three MtZ isolates studied.

It has been observed (Ates et al., 2018) that MTB strains belonging to the Beijing lineage, often thought to be hypervirulent, are defective in the secretion of PE_PGRS (polymorphic GC-rich repetitive sequences) proteins and that the secretion of more than 80 proteins of this protein family is dependent on the presence of at least one copy of *ppe38/ppe71* genes. Ates et al. (2018) demonstrated a possible link between the hypervirulent phenotype of some Beijing strains and changes in this genomic region caused by IS*6110* insertion and/or recombination between *ppe38/71*. As MtZ also has both of these genes targeted by IS*6110* insertions, we wanted to investigate whether PE_PGRS secretion was affected in this strain. Interestingly, the analysis of the PE_PGRS secretion by Western blot revealed that the MtZ strain produced and secreted PE_PGRS, even in higher amounts than H37Rv, used as reference (**Figure 7**), suggesting that the specific *IS6110* insertions did not disrupt the functionality of the remaining PPE proteins.

DISCUSSION

The MtZ strain is the most successful strain ever documented in the Aragon region, circulating from the 1990s until now. Thanks to the molecular surveillance protocol of TB carried out in Aragon, all MTB clinical isolates are genotyped (except for those during 1996–2000). At the same time that MtZ arose, something similar was happening in other parts of the world: strain GC in Gran Canaria (Caminero et al., 2001), strain C in New York (Friedman et al., 1997), strain CH in Leicester (Watson and Moss, 2001), the Harlingen strain in the Netherlands (van Soolingen et al., 1999), and the Danish cluster 1 and 2 strains in Denmark (Lillebaek et al., 2004). All these strains, MtZ included, had in common high transmissibility and drug susceptibility. It is difficult to understand how a drug-susceptible strain of *M. tuberculosis* can spread so widely among the first-world population. A higher capacity of infection could explain how this strain became responsible for such a large proportion of TB cases in a city with a strong health coverage. On the other hand, the cases were correctly treated, and consequently, few drug resistances developed. The observed tendency suggests that the outbreak is

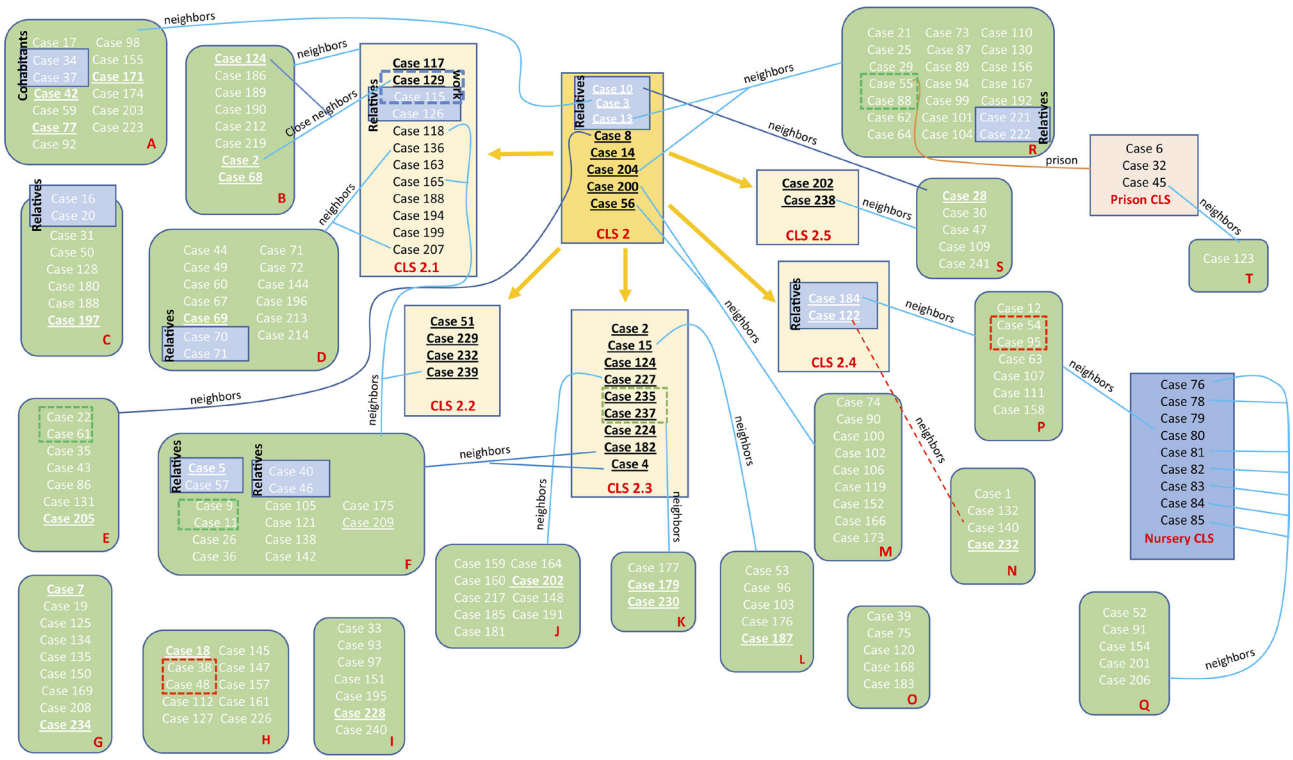


FIGURE 4 | Schematic view of the transmission chain for CLS-2 with regard to the molecular and epidemiological links. Five transmission CLSs emerged from CLS-2, the original sequence of this MtZ branch. Epidemiological links are included in the picture. Each red letter corresponds to a different neighborhood. Those cases presented in a blue box had a confirmed epidemiological link (relatives, work contact, and the nursery cluster). Those in a red dotted box or line live at the same building although no familiar link was established (probable epidemiological link). Those cases in a green dotted box are nearby neighbors (living on the same street). Those underlined are the ones sequenced. Cases 2, 124, 188, 202, and 232 appear twice (in the sub-CLS and the neighborhood they belong to) in order to simplify the diagram.

about to conclude, as no cases were diagnosed in 2020; however, at least one case was detected in 2021 (not included in the study).

The development of WGS has allowed a deeper study of the molecular characteristics of outbreak strains. MtZ belongs to L4.8, the closest phylogenetic branch to the reference strain H37Rv (L4.9), showing less than 200 SNPs between them We could confirm the molecular susceptibility to all the first-line drugs in all cases; however, molecular monoresistance to quinolones was found in case 124, a great advantage of WGS over other genotyping methods, offering a global vision of the genetic resistance pattern of the isolates. This resistance could have developed as a consequence of a previous treatment since a residual previous TB was observed in the patient's thorax X-ray.

Moreover, the MtZ strain had eight SNPs in genes that have been listed as virulence factors (Ramage et al., 2009; Forrellad et al., 2013) but whose real impact on virulence remains to be demonstrated. Additionally, mutations in 14 genes that could be somehow related to *in-vivo* growth, pathogenesis, and survival in the murine host (Sassetti et al., 2003; Rengarajan et al., 2005; DeJesus et al., 2017) could also interfere with the specific dominance of this strain. Regarding the SNPs in potential virulence factors, *mce* genes have been traditionally related to cell entry (Arruda et al., 1993). For the *mce3* operon, this function was demonstrated for the *mce3C* gene (Zhang et al., 2018); however, the complete *mce3* operon seems more related to lipid transport (Klepp et al., 2022). The MtZ strain has two SNPs in this operon, one of them a frameshift, suggesting a possible defect in this lipid transport. Nevertheless, it has been shown that the different *mce* operons can interact with each other (Klepp et al., 2022); therefore, complementation between systems could be possible. In addition, the entire *mce3* operon is absent from *M. bovis* and most other animal-adapted strains of the *M. tuberculosis complex*, suggesting that it is not required for infecting the macrophages and preventing that the frameshift SNP attenuates the MtZ strain. Concerning the two SNPs in the *pks15* gene, they might affect the lipid metabolism, but this remains unclear, as *pks15* is already a pseudogene in L4 strains (Constant et al., 2002). Some SNPs also suggest that MtZ could have a better survival ability inside the macrophages: the *virS* gene which controls the vesicular trafficking and survival inside the macrophages regulating the host immune system (Singh et al., 2019); and the *ptpA* gene, responsible for the inhibition of macrophage phagosome–lysosome fusion and phagosome acidification (Bach et al., 2008; Wong et al., 2011;

TABLE 3 | SNPs transmitted in the different CLSs of the MtZ outbreak.

Point	Gene	Amino acid change	Mutation effect	Functional category	Gene description	CLS
551	*dnaA*	R184Q	Non-synonymous (deleterious)	Information pathways	Plays an important role in the initiation and regulation of chromosomal replication.	CLS-2.4
46642	*Rv0042c*	L189R	Non-synonymous (deleterious)	Regulatory proteins	Possibly involved in transcriptional mechanism	CLS-2.5
862566	*Rv0769*	G52D	Non-synonymous (deleterious)	Intermediary metabolism and respiration	Dehydrogenase/reductase	CLS-2.1
898986	*Rv0805*	–	Synonymous	Intermediary metabolism and respiration	Hydrolyzes cyclic nucleotide monophosphate to nucleotide monophosphate	CLS-2.2
1301938	*PE12*	–	Synonymous	Pe/ppe	PE family protein PE12	CLS-1
2180630	*Rv1928c*	I196M	Non-synonymous (neutral)	Intermediary metabolism and respiration	Short-chain type dehydrogenase/reductase	CLS-2.4
2544858	*lppN*	A54V	Non-synonymous (neutral)	Cell wall and cell processes	Lipoprotein LppN	CLS-1.1
2770751	*pepN*	V589A	Non-synonymous (deleterious)	Intermediary metabolism and respiration	Aminopeptidase with broad substrate specificity to several peptides	CLS-2.2
3070366	*Rv2757c*	W74L	Non-synonymous (deleterious)	Virulence, detoxification, adaptation	VapC21 toxin	CLS-2.2
3081980	*dapB*	R121Q	Non-synonymous (neutral)	Intermediary metabolism and respiration	Involved in the biosynthesis of diaminopimelate and lysine from aspartate semialdehyde	CLS-2.3
3143468	*dinF*	L54P	Non-synonymous (deleterious)	Information pathways	Induction by DNA damage	CLS-2.1
3702632	*deoA*	L279R	Non-synonymous (deleterious)	Intermediary metabolism and respiration	Catalyzes the reversible phosphorolysis of pyrimidine nucleosides	CLS-1 and CLS-2 (including the subclusters)
3716874	*dacB1*	–	Synonymous	Cell wall and cell processes	Involved in peptidoglycan synthesis	CLS-2 (including the subclusters)

Points in the genome and the affected genes are in reference to the H37Rv strain. The amino acid change, the kind of mutation and its potential effect (according to PROVEAN), the functional category (according to Mycobrowser), and information concerning the genes are supplied.

Poirier et al., 2014; Wang et al., 2015; Wang et al., 2016). The toxin–antitoxin systems play an important role in stress response and genome stability (Ramage et al., 2009), so the two SNPs found in *VapB41* and *VapC50* genes could also be affecting the success of the infection. Considering all this, some of these mutations could be related to the high transmissibility and success of the MtZ strain.

The explosion of cases took place at the beginning of the 2000s, as only two cases were detected for 1993–1995 (with no data for 1996–2000). The six first cases affected with the MtZ strain in 2001 were diagnosed over the course of a few days, and the outbreak turned into the largest one ever experienced in the region. In a study carried out by Hamblion et al. in 2016 (Hamblion et al., 2016) regarding clustering in London for a 3-year period, they observed that if the two first cases of an outbreak are diagnosed with less than 90 days between them, it is probable that the outbreak becomes larger. Genomic information revealed the gain of two mutations (points 3702632 and either 3716874 or 1301938) for almost all MtZ cases since 2001; therefore, both facts may be related to its successful spread in the population. However, none of those SNPs is in a gene considered to be a virulence factor or relevant for pathogenesis.

It was curious that people infected with the MtZ strain did not have, in general, any risk factor, which was already observed by López-Calleja et al. (2009) for the first 85 cases. Different from an X-family outbreak which also began in the 1990s among the local population and with a high proportion of HIV^+ individuals, users of intravenous drugs, and prisoners (Comín et al., 2021), the MtZ outbreak did not show a high percentage of these individuals, just a high proportion of smokers (38%) and a moderate proportion of alcohol dependence (18%). On the other hand, in both outbreaks, young people were the most affected group, with more than 70% of the cases under 45 years old.

Although some links were established using epidemiological information, the genomic study revealed additional transmission chains, as well as the evolution of the strain. From cases 0 and 7, an SNP was gained at position 3702632 (case x, missing from our study). Subsequently, one SNP in 1301938 led to CLS-1, and another in 3716874 led to CLS-2. CLS-2 comprised the largest number of isolates studied, and five subclusters arose from it. As a result of the WGS analysis, a higher number of cases could be included in the transmission chain than just those connected by epidemiological links. However, the direction of the transmission could not be established in the majority of cases, something recurrent in TB outbreak studies (Casali et al., 2016; Hatherell et al., 2016; Lalor et al., 2018; Comin et al., 2020; Comín et al., 2021). As an example, cases 51 and 229 from CLS-2.2 had an identical genomic sequence, but case 51 was diagnosed 15 years before, so we think that case 51 would be the first case, as transmission always moves toward SNP accumulation (Schürch et al., 2010). Case 51 and/or case 229 or other related but not sequenced isolates should have infected case 232 and case 239 (also belonging to CLS-2.2), which accumulated more SNPs. Another extra difficulty to study the TB outbreaks and the reconstruction of the transmission chain is the latency period, as the infection could take place years before the diagnosis (Salgame et al., 2015). We observed this phenomenon, for

FIGURE 5 | Map showing the geographical distribution of MtZ CLSs within Zaragoza City, using a presence–absence legend (orange–gray). CLS-2 was the most widespread, involving the highest number of neighborhoods. There were no cases of CLS-1.1 because they were in the Huesca area, another province of the community.

example, for cases 5 and 57 (WGS not available), who were father and son, respectively. The father was diagnosed in 2001, probably when he infected his son, who developed the disease 2 years later, in 2003.

Regarding the SNPs transmitted among the different clusters, four of them caught our attention for being in genes related to virulence and pathogenesis. An SNP in the *Rv0805* gene, which regulates the intracellular concentration of cAMP and could have been altering the properties of the cell wall, was transmitted among cases of CLS-2.2 (Shenoy et al., 2005; Podobnik et al., 2009). In addition, an SNP in the *pepN* gene, which encodes for an M1 family zinc metallo-aminopeptidase that plays a crucial role in survival, cell maintenance, growth and development, virulence, and pathogenesis, was also transmitted among CLS-2.2 isolates (Ingmer and Brøndsted, 2009). It has been hypothesized that PepN may cleave pathogen and host proteins in host macrophages to regulate virulence levels (Sharma et al., 2019). A third SNP transmitted in CLS-2.2 was in the *Rv2757c* gene, VapC21 toxin, whose importance in pathogenesis has been described previously as part of a toxin–antitoxin system (Ramage et al., 2009). We could not find any isolate with only one or two of these SNPs, as if they occurred in a single isolate that later was transmitted. Finally, an SNP transmitted among cases of CLS-2.3 was in the *dapB* gene, an essential reductase whose inhibition blocks the production of meso-diaminopimelate, leading to inhibition of *de-novo* lysine biosynthesis and peptidoglycan assembly. Both of these pathways are crucial for the survival of the pathogen (Usha et al., 2012). Among the subclusters obtained by WGS, CLS-2.2 and CLS-2.3 were the largest, which may be related to the presence of some of these SNPs. Another important fact is the extra IS*6110* inserted in *dnaA:dnaN* observed for CLS-2.1. This variant was transmitted to 12 isolates over the entire period considered by the study, so the variant persisted in the population. We do not

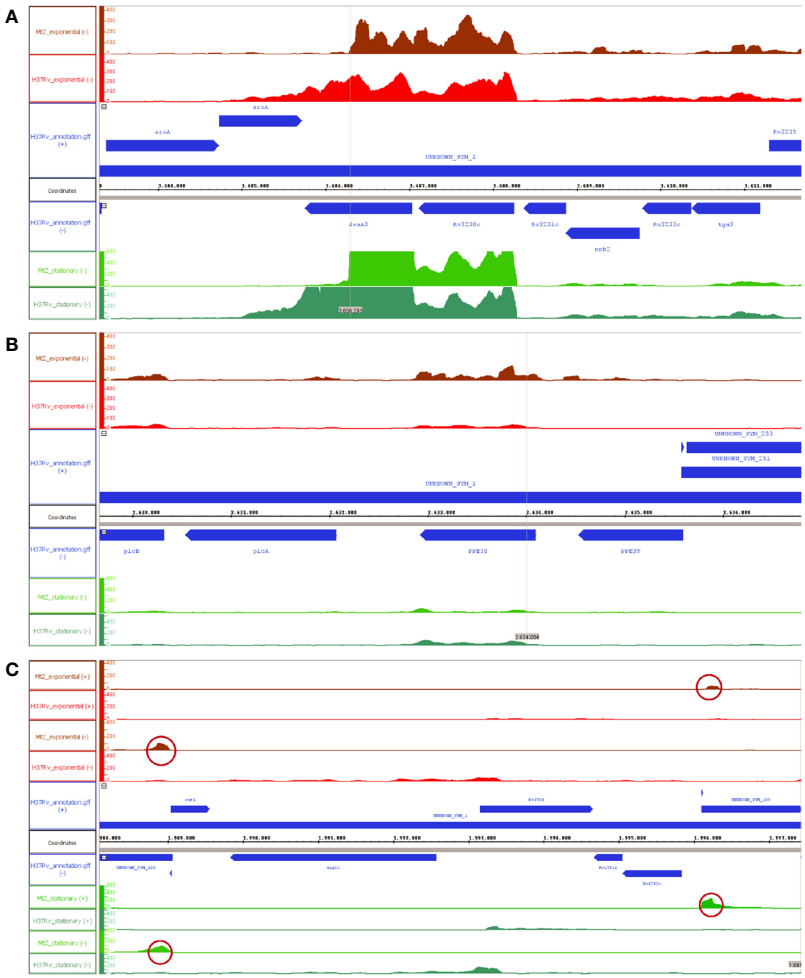


FIGURE 6 | Effects of the IS*6110* in the transcription of the MtZ strain. **(A)** Transcriptomic profile of MtZ and H37Rv in the *desA3* gene region. The transcription is interrupted in MtZ. **(B)** Transcriptomic profile of MtZ and H37Rv in the *ppe38* gene region. This gene is overexpressed in MtZ in the exponential growth phase. **(C)** Transcriptomic profile of the *cut1:Rv1765c* region. The absence of transcription confirms the loss of the intermediary genes. The peaks circled in red coincided with IS*6110* copies in the genome of the H37Rv strain showing overexpression in MtZ. For all pictures: above is the exponential growth phase (in red colors) and below is the stationary growth phase (in green colors). In the center is the annotation for the H37Rv reference strain (in blue color). "−" is the negative DNA strand; "+" is the positive DNA strand.

know if this copy conferred any biological advantage, but at least it did not seem to be detrimental to the bacteria. Nothing remarkable was found in the transcriptomic study related to this additional IS*6110* copy for this MtZ variant.

Regarding the genetic distances, the majority of isolates could be considered recent contact (≤5 SNPs) with at least one isolate. Six cases were more distant (≥5 and ≤12), but it is important to have in mind that the strain was circulating for many years, so evolutionary SNPs occurred. In addition, many unique SNPs detected in certain isolates could either have been generated in the culture process or have been introduced by the sequencing.

In order to better understand the role of the multiple IS*6110* copies in the genome, a transcriptomic study was carried out in three MtZ isolates that differed for one IS*6110* copy. As we knew the location of all the copies, we wondered about the role of this element in the transcription of the adjacent genes. This analysis allowed us to confirm different pieces of evidence. First, we could see the three general effects that the insertion of the IS*6110* could produce in the genome: the disruption of the gene in which it has been inserted (*desA3*) (Sampson et al., 1999; Beggs et al., 2000; Warren et al., 2000; Yesilkaya et al., 2005), the deletion or inversion of the DNA in between (*cut1:Rv1765c*) due to recombination (Sampson et al., 2003), and the overexpression effect (*ppe38*) (Beggs et al., 2000; Safi et al., 2004; Soto et al., 2004; Alonso et al., 2011) (**Figure 6**). No differences were detected in the transcription of *dnaA* or *dnaN* genes between case 0 and case 129. Second, we could observe a higher transcription of IS*6110* in the three MtZ isolates analyzed compared with H37Rv. This means that MtZ had some of its IS copies overexpressed. We speculate that it could be the one in *desA3*, as this gene has high

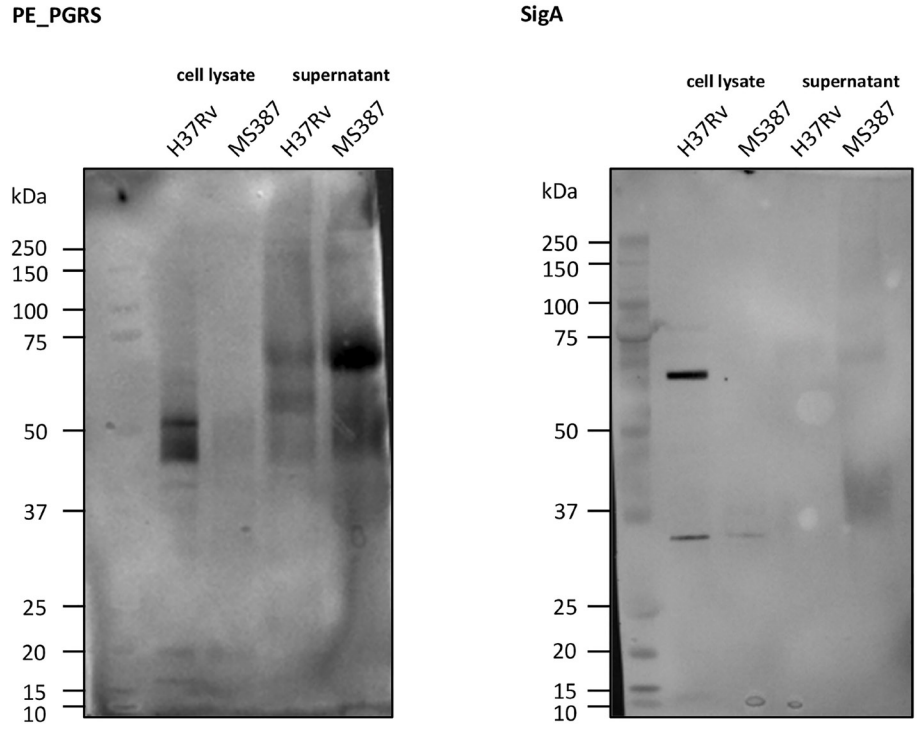


FIGURE 7 | Western blot of PE_PGRS secretion. As can be observed, MS 387 (MtZ strain, case 0) secretes PE_PGRS, even in a higher amount than the reference H37Rv strain. sigA control did not appear in MS 387 because the protein concentration in the cell lysate was too low.

levels of transcription, even higher in the stationary growth phase. Consequently, there will be a large number of IS*6110* reads mapped with all of the H37Rv ISs in the genome responsible for the high peaks we have observed.

According to other authors (Ates et al., 2018), PE_PGRS secretion is related to virulence. These authors have observed that deletion of the *ppe38/71* locus prevents this secretion from making Beijing strains hypervirulent. We demonstrated that MtZ, despite having both *ppe38/71* genes affected by an IS*6110* insertion, still secretes PE_PGRS proteins. The cause of having overexpressed the PE_PGRS proteins could be precisely the IS*6110* inserted in *ppe38*. Although IS*6110* is inserted within the gene, an ORF predictor showed the existence of an ORF in which the resulted protein was PPE38 lacking the first 40 amino acids. We hypothesize that this protein could conserve part of its functional activity. The possibility that IS*6110* could be acting as a promoter would be supported by the fact that the transcriptomic profile also showed overexpression of the *ppe38* gene in the exponential growth phase. It should be mentioned that in a previous work, we showed that an IS*6110* insertion upstream of the two-component regulator PhoP strongly increased the expression of this gene (Soto et al., 2004) and also had phenotypic consequences. The specific insertion of the IS*6110* in the herein observed case might have a similar function.

The study had some limitations. First, just 23.7% of the isolates belonging to the cluster were sequenced; therefore, several links and new transmission clusters have been lost. Because of this, we cannot know the true extent of CLS-2.2 and CLS-2.3, for which transmitted SNPs in genes related to pathogenesis were involved. Moreover, data for 1996–2000, when the strain probably started to spread with more force, are not available. The COVID-19 pandemic required that we perform the sequences with different platforms, making the analysis in BioNumerics software more difficult. This was solved by analyzing the SNPs detected in the isolates singularly.

In conclusion, the MtZ strain produced the largest outbreak ever reported in the Aragon region, and this outbreak remains active even today despite a general lack of risk factors for developing TB among the population. Fortunately, it is a drug-susceptible strain and patients can be cured with treatment. WGS allowed us to clarify the transmission chain more than epidemiological information alone although the direction of transmission was not solved for all cases. MtZ includes several SNPs in genes considered to be virulence factors as well as genes involved in pathogenesis and survival, which could be the cause of its success. More research is needed to know how the higher IS*6110* transcription and the oversecretion of PE_PGRS observed in the MtZ strain affect its virulence and pathogenesis.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and

accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, SAMN23235155–SAMN23235211.

AUTHOR CONTRIBUTIONS

JC wrote the manuscript, contributed to the conception and design of the study, analyzed the data, and performed the laboratory work. JM and RB performed some laboratory work, wrote sections of the manuscript, and revised the submitted manuscript. IR and MZ-A performed the geographical part of the study and wrote sections of the manuscript. DI, JV, LT, JS, and M-JI provided epidemiological support. AC, JG-A and CK provided bioinformatic support. SS wrote the manuscript, contributed to the conception and design of the study, and was responsible for funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Carlos III Health Institute in the context of a grant (FIS18/0336) and JC was awarded a scholarship by the Government of Aragon/European Social Fund, "Building Europe from Aragon". JM and RB acknowledge the support by the Agence Nationale de la Recherche (grants ANR-10-LABX-62-IBEID and ANR-20-CE15-0013-03).

ACKNOWLEDGMENTS

The authors would like to acknowledge Servicio General de Apoyo a la Investigación-SAI, Universidad de Zaragoza (Servicio de Análisis Microbiológico) and Servicios Científico Técnicos, IACS (Servicio de Secuenciación y Genómica Funcional and Servicio de Biocomputación). We would like to thank the EPIMOLA group for supplying the genotyped bacterial DNA used in this work. This work was supported by the Carlos III Health Institute in the context of a grant (FIS18/0336) and JC was awarded a scholarship by the Government of Aragon/European Social Fund, "Building Europe from Aragon". JM and RB acknowledge the support by the Agence Nationale de la Recherche (grants ANR-10-LABX-62-IBEID and ANR-20-CE15-0013-03).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2022.887134/full#supplementary-material

REFERENCES

Alonso, H., Aguilo, J. I., Samper, S., Caminero, J. A., Campos-Herrero, M. I., Gicquel, B., et al. (2011). Deciphering the Role of IS6110 in a Highly Transmissible Mycobacterium Tuberculosis Beijing Strain, GC1237. *Tuberculosis* 91 (2), 117–126. doi: 10.1016/j.tube.2010.12.007

Arruda, S., Bomfim, G., Knights, R., Huima-Byron, T., and Riley, L. W. (1993). Cloning of an M. Tuberculosis DNA Fragment Associated With Entry and Survival Inside Cells. *Science* 261 (5127), 1454–1457. doi: 10.1126/science.8367727

Ates, L. S., Dippenaar, A., Ummels, R., Piersma, S. R., van der Woude, A. D., van der Kuij, K., et al. (2018). Mutations in Ppe38 Block PE-PGRS Secretion and Increase Virulence of Mycobacterium Tuberculosis. *Nat. Microbiol.* 3 (2), 181–188. doi: 10.1038/s41564-017-0090-6

Bach, H., Papavinasasundaram, K. G., Wong, D., Hmama, Z., and Av-Gay, Y. (2008). Mycobacterium Tuberculosis Virulence Is Mediated by PtpA Dephosphorylation of Human Vacuolar Protein Sorting 33B. *Cell Host Microbe* 3 (5), 316–322. doi: 10.1016/j.chom.2008.03.008

Beggs, M. L., Eisenach, K. D., and Cave, M. D. (2000). Mapping of IS6110 Insertion Sites in Two Epidemic Strains of Mycobacterium Tuberculosis. *J. Clin. Microbiol.* 38 (8), 2923–2928. doi: 10.1128/JCM.38.8.2923-2928.2000

Borgdorff, M. W., Behr, M. A., Nagelkerke, N. J., Hopewell, P. C., and Small, P. M. (2000). Transmission of Tuberculosis in San Francisco and Its Association With Immigration and Ethnicity. *Int. J. Tuberc. Lung Dis. Off. J. Int. Union. Against. Tuberc. Lung Dis.* 4 (4), 287–294.

Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., et al. (2002). A New Evolutionary Scenario for the Mycobacterium Tuberculosis Complex. *Proc. Natl. Acad. Sci. U. S. A.* 99 (6), 3684–3689. doi: 10.1073/pnas.052548299

Caceres, N., Llopis, I., Marzo, E., Prats, C., Vilaplana, C., García de Viedma, D., et al. (2012). Low Dose Aerosol Fitness at the Innate Phase of Murine Infection Better Predicts Virulence Amongst Clinical Strains of Mycobacterium Tuberculosis. *PloS One* 7 (1), e29010. doi: 10.1371/journal.pone.0029010

Caminero, J. A., Pena, M. J., Campos-Herrero, M. I., Rodríguez, J. C., García, I., Cabrera, P., et al. (2001). Epidemiological Evidence of the Spread of a Mycobacterium Tuberculosis Strain of the Beijing Genotype on Gran Canaria Island. *Am. J. Respir. Crit. Care Med.* 164 (7), 1165–1170. doi: 10.1164/ajrccm.164.7.2101031

Casali, N., Broda, A., Harris, S. R., Parkhill, J., Brown, T., and Drobniewski, F. (2016). Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PloS Med.* 13 (10), 1–18. doi: 10.1371/journal.pmed.1002137

Cirillo, D. M., Cabibbe, A. M., De Filippo, M. R., Trovato, A., Simonetti, T., Rossolini, G. M., et al. (2016). Use of WGS in Mycobacterium Tuberculosis Routine Diagnosis. *Int. J. Mycobacteriol* 5 (Suppl 1), S252–S253. doi: 10.1016/j.ijmyco.2016.09.053

Coll, F., McNerney, R., Guerra-Assunção, J. A., Glynn, J. R., Perdigão, J., Viveiros, M., et al. (2014). A Robust SNP Barcode for Typing Mycobacterium Tuberculosis Complex Strains. *Nat. Commun.* 5, 4–8. doi: 10.1038/ncomms5812

Comín, J., Cebollada, A., Ibarz, D., Viñuelas, J., Vitoria, M. A., Iglesias, M. J., et al. (2021). A Whole-Genome Sequencing Study of an X-Family Tuberculosis Outbreak Focus on Transmission Chain Along 25 Years. *Tuberculosis* 126, 102022. doi: 10.1016/j.tube.2020.102022

Comin, J., Chaure, A., Cebollada, A., Ibarz, D., Viñuelas, J., Vitoria, M. A., et al. (2020). Investigation of a Rapidly Spreading Tuberculosis Outbreak Using Whole-Genome Sequencing. *Infect. Genet. Evol.* 81, 104184. doi: 10.1016/j.meegid.2020.104184

Constant, P., Perez, E., Malaga, W., Lanéelle, M.-A., Saurel, O., Daffé, M., et al. (2002). Role of the Pks15/1 Gene in the Biosynthesis of Phenolglycolipids in the Mycobacterium Tuberculosis Complex. Evidence That All Strains Synthesize Glycosylated P-Hydroxybenzoic Methyl Esters and That Strains Devoid of Phenolglycolipids Harbor a Frameshift. *J. Biol. Chem.* 277 (41), 38148–38158. doi: 10.1074/jbc.M206538200

DeJesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., et al. (2017). Comprehensive Essentiality Analysis of the Mycobacterium

Tuberculosis Genome via Saturating Transposon Mutagenesis. *MBio* 8 (1), e02133–16. doi: 10.1128/mBio.02133-16

Forrellad, M. A., Klepp, L. I., Gioffré, A., García, J. S., Morbidoni, H. R., de la Paz Santangelo, M., et al. (2013). Virulence Factors of the Mycobacterium Tuberculosis Complex. *Virulence* 4 (1), 3–66. doi: 10.4161/viru.22329

Freese, N. H., Norris, D. C., and Loraine, A. E. (2016). Integrated Genome Browser: Visual Analytics Platform for Genomics. *Bioinformatics* 32 (14), 2089–2095. doi: 10.1093/bioinformatics/btw069

Friedman, C. R., Quinn, G. C., Kreiswirth, B. N., Perlman, D. C., Salomon, N., Schluger, N., et al. (1997). Widespread Dissemination of a Drug-Susceptible Strain of Mycobacterium Tuberculosis. *J. Infect. Dis.* 176 (2), 478–484. doi: 10.1086/514067

García-Rodríguez, J. F., Álvarez-Díaz, H., Lorenzo-García, M. V., Mariño-Callejo, A., Fernández-Rial, Á., and Sesma-Sánchez, P. (2011). Extrapulmonary Tuberculosis: Epidemiology and Risk Factors. *Enferm. Infecc. Microbiol. Clin.* 29 (7), 502–509. doi: 10.1016/j.eimc.2011.03.005

Gonzalo-Asensio, J., Pérez, I., Aguiló, N., Uranga, S., Picó, A., Lampreave, C., et al. (2018). New Insights Into the Transposition Mechanisms of IS6110 and Its Dynamic Distribution Between Mycobacterium Tuberculosis Complex Lineages. *PloS Genet* 14 (4), e1007282. doi: 10.1371/journal.pgen.1007282

Hamblion, E. L., Le Menach, A., Anderson, L. F., Lalor, M. K., Brown, T., Abubakar, I., et al. (2016). Recent TB Transmission, Clustering and Predictors of Large Clusters in London, 2010-2012: Results From First 3-Years of Universal MIRU-VNTR Strain Typing. *Thorax* 71 (8), 749–756. doi: 10.1136/thoraxjnl-2014-206608

Hatherell, H. A., Colijn, C., Stagg, H. R., Jackson, C., Winter, J. R., and Abubakar, I. (2016). Interpreting Whole Genome Sequencing for Investigating Tuberculosis Transmission: A Systematic Review. *BMC Med.* 14, 21. doi: 10.1186/s12916-016-0566-x

Iglesias, M. J., Ibarz, D., Cebollada, A., Comín, J., Jiménez, M. S., Vázquez, M. C., et al. (2020). The Value of the Continuous Genotyping of Multi-Drug Resistant Tuberculosis Over 20 Years in Spain. *Sci. Rep.* 10 (1), 20433. doi: 10.1038/s41598-020-77249-x

Ingmer, H., and Brøndsted, L. (2009). Proteases in Bacterial Pathogenesis. *Res. Microbiol.* 160 (9), 704–710. doi: 10.1016/j.resmic.2009.08.017

Isabel Millan-Lou, M., Isabel López-Calleja, A., Colmenarejo, C., Antonia Lezcano, M., Asunción Vitoria, M., Del Portillo, P., et al. (2013). Global Study of Is6110 in a Successful Mycobacterium Tuberculosis Strain: Clues for Deciphering its Behavior and for its Rapid Detection. *J. Clin. Microbiol.* 51 (11), 3631–3637. doi: 10.1128/JCM.00970-13

Kamerbeek, J., Schouls, L., Kolk, A., Van Agterveld, M., Van Soolingen, D., Kuijper, S., et al. (1997). Simultaneous Detection and Strain Differentiation of Mycobacterium Tuberculosis for Diagnosis and Epidemiology. *J. Clin. Microbiol* 35 (4), 907–914. doi: 10.1128/jcm.35.4.907-914.1997

Klepp, L. I., Sabio Y Garcia, J., and FabianaBigi, (2022). Mycobacterial MCE Proteins as Transporters That Control Lipid Homeostasis of the Cell Wall. *Tuberculosis (Edinb).* 132, 102162. doi: 10.1016/j.tube.2021.102162

Lalor, M. K., Casali, N., Walker, T. M., Anderson, L. F., Davidson, J. A., Ratna, N., et al. (2018). The Use of Whole-Genome Sequencing in Cluster Investigation of a Multidrug-Resistant Tuberculosis Outbreak. *Eur. Respir. J.* 51 (6), 1702313. doi: 10.1183/13993003.02313-2017

Lillebaek, T., Dirksen, A., Kok-Jensen, A., and Andersen, A. B. (2004). A Dominant Mycobacterium Tuberculosis Strain Emerging in Denmark. *Int. J. Tuberc. Lung Dis. Off. J. Int. Union. Against. Tuberc. Lung Dis.* 8 (8), 1001–1006.

López-Calleja, A. I., Gavín, P., Antonia, M. A., Vitoria, M. A., Iglesias, M. J., Guimbao, J., et al. (2009). Unsuspected and Extensive Transmission of a Drug-Susceptible Mycobacterium Tuberculosis Strain. *BMC Pulm. Med.* 9, 1–10. doi: 10.1186/1471-2466-9-3

López-Calleja, A. I., Lezcano, M. A., Vitoria, M. A., Iglesias, M. J., Cebollada, A., Lafoz, C., et al. (2007). Genotyping of Mycobacterium Tuberculosis Over Two Periods: A Changing Scenario for Tuberculosis Transmission. *Int. J. Tuberc. Lung Dis.* 11 (10), 1080–1086.

Madacki, J., Orgeur, M., Mas Fiol, G., Frigui, W., Ma, L., and Brosch, R. (2021). ESX-1-Independent Horizontal Gene Transfer by Mycobacterium Tuberculosis Complex Strains. *MBio* 12 (3), e00965–21. doi: 10.1128/mBio.00965-21

Morse, D., Brothwell, D. R., and Ucko, P. J. (1964). TUBERCULOSIS IN ANCIENT EGYPT. *Am. Rev. Respir. Dis.* 90, 524–541. doi: 10.1016/s0761-8425(07)78506-6

Nikolayevskyy, V., Niemann, S., Anthony, R., van Soolingen, D., Tagliani, E., Ködmön, C., et al. (2019). Role and Value of Whole Genome Sequencing in Studying Tuberculosis Transmission. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* 25 (11), 1377–1382. doi: 10.1016/j.cmi.2019.03.022

Pérez, I., Uranga, S., Sayes, F., Frigui, W., Samper, S., Arbués, A., et al. (2020). Live Attenuated TB Vaccines Representing the Three Modern Mycobacterium Tuberculosis Lineages Reveal That the Euro-American Genetic Background Confers Optimal Vaccine Potential. *EBioMedicine* 55, 102761. doi: 10.1016/j.ebiom.2020.102761

Podobnik, M., Tyagi, R., Matange, N., Dermol, U., Gupta, A. K., Mattoo, R., et al. (2009). A Mycobacterial Cyclic AMP Phosphodiesterase That Moonlights as a Modifier of Cell Wall Permeability. *J. Biol. Chem.* 284 (47), 32846–32857. doi: 10.1074/jbc.M109.049635

Poirier, V., Bach, H., and Av-Gay, Y. (2014). Mycobacterium Tuberculosis Promotes Anti-Apoptotic Activity of the Macrophage by PtpA Protein-Dependent Dephosphorylation of Host GSK3α. *J. Biol. Chem.* 289 (42), 29376–29385. doi: 10.1074/jbc.M114.582502

Ramage, H. R., Connolly, L. E., and Cox, J. S. (2009). Comprehensive Functional Analysis of Mycobacterium Tuberculosis Toxin-Antitoxin Systems: Implications for Pathogenesis, Stress Responses, and Evolution. *PloS Genet.* 5 (12), e1000767. doi: 10.1371/journal.pgen.1000767

Rengarajan, J., Bloom, B. R., and Rubin, E. J. (2005). Genome-Wide Requirements for Mycobacterium Tuberculosis Adaptation and Survival in Macrophages. *Proc. Natl. Acad. Sci. U. S. A.* 102 (23), 8327–8332. doi: 10.1073/pnas.0503272102

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative Genome Viewer. *Nat. Biotechnol.* 29 (1), 24–26. doi: 10.1038/nbt.1754

Safi, H., Barnes, P. F., Lakey, D. L., Shams, H., Samten, B., Vankayalapati, R., et al. (2004). IS6110 Functions as a Mobile, Monocyte-Activated Promoter in Mycobacterium Tuberculosis. *Mol. Microbiol.* 52 (4), 999–1012. doi: 10.1111/j.1365-2958.2004.04037.x

Salgame, P., Geadas, C., Collins, L., Jones-López, E., and Ellner, J. J. (2015). Latent Tuberculosis Infection - Revisiting and Revising Concepts. *Tuberculosis* 95 (4), 373–384. doi: 10.1016/j.tube.2015.04.003

Sampson, S. L., Warren, R. M., Richardson, M., van der Spuy, G. D., and van Helden, P. D. (1999). Disruption of Coding Regions by IS6110 Insertion in Mycobacterium Tuberculosis. *Tuber. Lung Dis. Off. J. Int. Union. Against. Tuberc. Lung Dis.* 79 (6), 349–359. doi: 10.1054/tuld.1999.0218

Sampson, S. L., Warren, R. M., Richardson, M., Victor, T. C., Jordaan, A. M., van der Spuy, G. D., et al. (2003). IS6110-Mediated Deletion Polymorphism in the Direct Repeat Region of Clinical Isolates of Mycobacterium Tuberculosis. *J. Bacteriol.* 185 (9), 2856–2866. doi: 10.1128/JB.185.9.2856-2866.2003

Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2003). Genes Required for Mycobacterial Growth Defined by High Density Mutagenesis. *Mol. Microbiol.* 48 (1), 77–84. doi: 10.1046/j.1365-2958.2003.03425.x

Schürch, A. C., Kremer, K., Daviena, O., Kiers, A., Boeree, M. J., Siezen, R. J., et al. (2010). High-Resolution Typing by Integration of Genome Sequencing Data in a Large Tuberculosis Cluster. *J. Clin. Microbiol* 48 (9), 3403–6. doi: 10.1128/JCM.00370-10

Senaratne, R. H., Sidders, B., Sequeira, P., Saunders, G., Dunphy, K., Marjanovic, O., et al. (2008). Mycobacterium Tuberculosis Strains Disrupted in Mce3 and Mce4 Operons Are Attenuated in Mice. *J. Med. Microbiol.* 57 (Pt 2), 164–170. doi: 10.1099/jmm.0.47454-0

Sharma, N., Aggarwal, S., Kumar, S., Sharma, R., Choudhury, K., Singh, N., et al. (2019). Comparative Analysis of Homologous Aminopeptidase PepN From Pathogenic and non-Pathogenic Mycobacteria Reveals Divergent Traits. *PloS One* 14 (4), e0215123. doi: 10.1371/journal.pone.0215123

Shenoy, A. R., Sreenath, N., Podobnik, M., Kovacevic, M., and Visweswariah, S. S. (2005). The Rv0805 Gene From Mycobacterium Tuberculosis Encodes a 3',5'-Cyclic Nucleotide Phosphodiesterase: Biochemical and Mutational Analysis. *Biochemistry* 44 (48), 15695–15704. doi: 10.1021/bi0512391

Singh, S., Goswami, N., Tyagi, A. K., and Khare, G. (2019). Unraveling the Role of the Transcriptional Regulator VirS in Low pH-Induced Responses of

Mycobacterium Tuberculosis and Identification of VirS Inhibitors. *J. Biol. Chem.* 294 (26), 10055–10075. doi: 10.1074/jbc.RA118.005312

Soto, C. Y., Menéndez, M. C., Pérez, E., Samper, S., Gómez, A. B., García, M. J., et al. (2004). IS6110 Mediates Increased Transcription of the phoP Virulence Gene in a Multidrug-Resistant Clinical Isolate Responsible for Tuberculosis Outbreaks. *J. Clin. Microbiol* 42 (1), 212–219. doi: 10.1128/JCM.42.1.212-219.2004

Usha, V., Lloyd, A. J., Lovering, A. L., and Besra, G. S. (2012). Structure and Function of Mycobacterium Tuberculosis Meso-Diaminopimelic Acid (DAP) Biosynthetic Enzymes. *FEMS Microbiol. Lett.* 330 (1), 10–16. doi: 10.1111/j.1574-6968.2012.02527.x

Van Embden, J. D. A., Cave, M. D., Crawford, J. T., Dale, J. W., Eisenach, K. D., Gicquel, B., et al. (1993). Strain Identification of Mycobacterium Tuberculosis by DNA Fingerprinting: Recommendations for a Standardized Methodology. *J. Clin. Microbiol.* 31 (2), 406–409. doi: 10.1128/jcm.31.2.406-409.1993

van Soolingen, D., Borgdorff, M. W., de Haas, P. E., Sebek, M. M., Veen, J., Dessens, M., et al. (1999). Molecular Epidemiology of Tuberculosis in the Netherlands: A Nationwide Study From 1993 Through 1997. *J. Infect. Dis.* 180 (3), 726–736. doi: 10.1086/314930

van Soolingen, D., de Haas, P. E., Hermans, P. W., and van Embden, J. D. (1994). DNA Fingerprinting of Mycobacterium Tuberculosis. *Methods Enzymol.* 235, 196–205. doi: 10.1016/0076-6879(94)35141-4

Wang, J., Li, B.-X., Ge, P.-P., Li, J., Wang, Q., Gao, G. F., et al. (2015). Mycobacterium Tuberculosis Suppresses Innate Immunity by Coopting the Host Ubiquitin System. *Nat. Immunol.* 16 (3), 237–245. doi: 10.1038/ni.3096

Wang, J., Teng, J. L. L., Zhao, D., Ge, P., Li, B., Woo, P. C. Y., et al. (2016). The Ubiquitin Ligase TRIM27 Functions as a Host Restriction Factor Antagonized by Mycobacterium Tuberculosis PtpA During Mycobacterial Infection. *Sci. Rep.* 6, 34827. doi: 10.1038/srep34827

Warren, R. M., Sampson, S. L., Richardson, M., van der Spuy, G. D., Lombard, C. J., Victor, T. C., et al. (2000). Mapping of IS6110 Flanking Regions in Clinical Isolates of Mycobacterium Tuberculosis Demonstrates Genome Plasticity. *Mol. Microbiol.* 37 (6), 1405–1416. doi: 10.1046/j.1365-2958.2000.02090.x

Watson, J. M., and Moss, F. (2001). TB in Leicester: Out of Control, or Just One of Those Things? *BMJ (Clin. Res. Ed.)* 322, 1133–1134. doi: 10.1136/bmj.322.7295.1133

Wong, D., Bach, H., Sun, J., Hmama, Z., and Av-Gay, Y. (2011). Mycobacterium Tuberculosis Protein Tyrosine Phosphatase (PtpA) Excludes Host Vacuolar-H +-ATPase to Inhibit Phagosome Acidification. *Proc. Natl. Acad. Sci. USA.* 108 (48), 19371–19376. doi: 10.1073/pnas.1109201108

World Health Organization (2020). *Global Tuberculosis Report*, Vol. 2020. (Geneva: World Health Organization).

Yesilkaya, H., Dale, J. W., Strachan, N. J. C., and Forbes, K. J. (2005). Natural Transposon Mutagenesis of Clinical Isolates of Mycobacterium Tuberculosis: How Many Genes Does a Pathogen Need? *J. Bacteriol.* 187 (19), 6726–6732. doi: 10.1128/JB.187.19.6726-6732.2005

Zhang, Y., Li, J., Li, B., Wang, J., and Liu, C. H. (2018). Mycobacterium Tuberculosis Mce3C Promotes Mycobacteria Entry Into Macrophages Through Activation of β2 Integrin-Mediated Signalling Pathway. *Cell. Microbiol* 20 (2). doi: 10.1111/cmi.12800

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Table S1. SNPs of MtZ strain when compared to H37Rv reference strain. If the SNP was not present in other L4.8 strains, it was considered as specific of MtZ strain.

Mutation point	Change	Gene	Funcional category	Mutation effect	Specific of the MtZ
371	C/T	*dnaA*	Information pathways	Non-synonymous	Yes
1584	T/C	Intergenic region			Yes
5472	C/T	*gyrB*	Information pathways	Non-synonymous	Yes
24445	C/G	*fhaA*	Regulatory proteins	Non-synonymous	Yes
28640	T/C	*Rv0024*	Virulence, detoxification, adaptation	Synonymous	No
33670	C/T	Intergenic region			Yes
69194	T/G	*Rv0064*	Cell wall and cell processes	Non-synonymous	Yes
80541	C/G	Intergenic region			Yes
81649	C/G	*Rv0072*	Cell wall and cell processes	Synonymous	Yes
101989	G/A	*ctpA*	Cell wall and cell processes	Synonymous	Yes
113371	A/G	*nrp*	Lipid metabolism	Non-synonymous	Yes
117880	C/G	*Rv0102*	Cell wall and cell processes	Non-synonymous	Yes
143262	C/T	*oxcA*	Intermediary metabolism and respiration	Synonymous	Yes
151169	T/G	*pepA*	Intermediary metabolism and respiration	Non-synonymous	Yes
212244	G/A	*Rv0180c*	Cell wall and cell processes	Synonymous	No
219104-219105	GA/G	*Rv0187*	Intermediary metabolism and respiration	Reading frame alteration	Yes
220134	G/A	*ilvD*	Intermediary metabolism and respiration	Synonymous	Yes
226475	C/T	*Rv0193c*	Conserved hypotheticals	Non-synonymous	Yes
232361	A/G	*Rv0197*	Intermediary metabolism and respiration	Non-synonymous	Yes
284047	C/G	*Rv0236c*	Cell wall and cell processes	Synonymous	Yes
286174	C/T	*Rv0236c*	Cell wall and cell processes	Synonymous	Yes
317300	G/A	*Rv0265c*	Cell wall and cell processes	Synonymous	Yes

365513	G/A	*Rv0303*	Intermediary metabolism and respiration	Non-synonymous	Yes
380245	G/A	*Rv0311*	Conserved hypotheticals	Synonymous	Yes
397910	G/C	*Rv0332*	Conserved hypotheticals	Non-synonymous	Yes
411759-411767	Del	*iniA*	Cell wall and cell processes	Reading frame alteration	Yes
444309	C/T	*Rv0366c*	Conserved hypotheticals	Synonymous	Yes
452550	G/C	*Rv0375c*	Intermediary metabolism and respiration	Non-synonymous	Yes
454253	C/T	*Rv0376c*	Conserved hypotheticals	Synonymous	Yes
505454	C/T	*lpqM*	Cell wall and cell processes	Synonymous	Yes
546865	C/T	Intergenic region			Yes
587065	G/C	*Rv0496*	Conserved hypotheticals	Synonymous	Yes
631002	G/T	*Rv0538*	Cell wall and cell processes	Synonymous	Yes
688973	T/G	Intergenic region			Yes
718799	G/A	*Rv0628c*	Conserved hypotheticals	Non-synonymous	Yes
793421	T/G	*lldD1*	Intermediary metabolism and respiration	Non-synonymous	Yes
884985	C/T	*Rv0791c*	Conserved hypotheticals	Non-synonymous	Yes
905190	C/T	Intergenic region			Yes
966808	C/T	*moaA2*	Intermediary metabolism and respiration	Synonymous	Yes
972808	T/C	*Rv0874c*	Conserved hypotheticals	Non-synonymous	Yes
976375	C/T	*Rv0877*	Conserved hypotheticals	Synonymous	Yes
988310	C/T	*Rv0888*	Cell wall and cell processes	Non-synonymous	Yes
1030214	C/G	*Rv0923c*	Conserved hypotheticals	Non-synonymous	Yes
1039931	C/T	Intergenic region			Yes
1054646	C/T	*Rv0945*	Intermediary metabolism and respiration	STOP	Yes
1061770	A/C	Intergenic region			Yes
1082445	T/A	*echA7*	Lipid metabolism	Non-synonymous	No
1086556	G/A	*accD2*	Lipid metabolism	STOP	Yes

1090278	T/G	Intergenic region			Yes
1130526	G/A	*ispE*	Intermediary metabolism and respiration	Synonymous	No
1207386	A/G	*mca*	Virulence, detoxification, adaptation	Not STOP codon	Yes
1210525	T/G	Intergenic region			Yes
1227217	C/T	*fum*	Intermediary metabolism and respiration	Non-synonymous	No
1266091	A/G	*Rv1138c*	Intermediary metabolism and respiration	Non-synonymous	Yes
1273071	T/C	*mmpL13a*	Cell wall and cell processes	Non-synonymous	No
1289307	C/T	*narG*	Intermediary metabolism and respiration	Synonymous	Yes
1358042	G/A	*Rv1215c*	Conserved hypotheticals	Non-synonymous	Yes
1358411	T/C	*Rv1215c*	Conserved hypotheticals	Non-synonymous	Yes
1413109	G/A	*mcr11/ MTB000063*	Stable RNAs		Yes
1443365	G/C	*Rv1289*	Conserved hypotheticals	Non-synonymous	Yes
1460992	A/C	*atpB*	Intermediary metabolism and respiration	Non-synonymous	Yes
1474001	C/T	*rrl*	Stable RNAs		No
1478182	G/A	*alkA*	Information pathways	Non-synonymous	Yes
1485501	C/T	*Rv1322A*	Conserved hypotheticals	Non-synonymous	Yes
1492138	C/G	*glgB*	Intermediary metabolism and respiration	Non-synonymous	Yes
1504420	G/A	*Rv1337*	Cell wall and cell processes	Non-synonymous	Yes
1525502	C/T	*Rv1357c*	Conserved hypotheticals	Non-synonymous	Yes
1529677	A/C	*Rv1358*	Regulatory proteins- Transcription	Synonymous	Yes
1541209	G/A	*lprF*	Cell wall and cell processes	Non-synonymous	Yes
1558227	C/T	*carB*	Intermediary metabolism and respiration	Non-synonymous	Yes
1710786	T/G	*Rv1519*	Conserved hypotheticals	Synonymous	Yes
1782597	G/A	Intergenic region			Yes

1860228	G/T	*pheT*	Information pathways	Synonymous	Yes
1884443-1884444	GT/TC	*pks8*	Lipid metabolism	Non-synonymous	Yes
1912539	C/T	*Rv1687c*	Cell wall and cell processes	Non-synonymous	Yes
1914977	A/C	*G2*	Stable RNAs		Yes
1916211	C/T	*Rv1691*	Conserved hypotheticals	Non-synonymous	Yes
1944933	C/T	*Rv1718*	Conserved hypotheticals	Non-synonymous	Yes
1952316	A/C	*Rv1726*	Intermediary metabolism and respiration	Synonymous	Yes
1953599	G/A	*Rv1727*	Conserved hypotheticals	Synonymous	Yes
1974661	C/T	*Rv1747*	Cell wall and cell processes	Synonymous	Yes
1976857	T/G	*Rv1748*	Conserved hypotheticals	Synonymous	Yes
1981095-1981096	GT/G	Intergenic region			Yes
2071567	G/A	*Rv1825*	Conserved hypotheticals	Non-synonymous	Yes
2094479	T/C	*gnd1*	Intermediary metabolism and respiration	Non-synonymous	Yes
2099039	C/A	*ureC*	Intermediary metabolism and respiration	Non-synonymous	Yes
2173090	C/G	*lppF*	Cell wall and cell processes	Non-synonymous	Yes
2175826	C/G	*lipD*	Intermediary metabolism and respiration	Synonymous	Yes
2184176	G/A	*fadE18*	Lipid metabolism	Non-synonymous	No
2184715-2184716	TG/T	*fadE18*	Lipid metabolism	Reading frame alteration	No
2199616	G/A	Intergenic region			Yes
2213395	A/G	*mce3D*	Virulence, detoxification, adaptation	Non-synonymous	Yes
2216011-2216012	TG/T	*mce3F*	Virulence, detoxification, adaptation	Reading frame alteration	Yes
2242928	A/C	Intergenic region			Yes
2246603	C/T	*Rv2000*	Conserved hypotheticals	Synonymous	No
2268952	A/C	*Rv2024c*	Conserved hypotheticals	Non-synonymous	Yes
2293955	A/C	*Rv2047c*	Conserved hypotheticals	Non-synonymous	Yes
2298776	C/A	*pks12*	Lipid metabolism	Synonymous	Yes

2356027	C/T	*Rv2097c*	Intermediary metabolism and respiration	Synonymous	Yes
2373539	C/T	*Rv2113*	Cell wall and cell processes	Non-synonymous	Yes
2373696	ins varias pb	*Rv2113*	Cell wall and cell processes	Reading frame alteration	Yes
2397635	C/A	*lppL*	Cell wall and cell processes	Non-synonymous	Yes
2410390	C/T	*ftsQ*	Cell wall and cell processes	Synonymous	Yes
2410831	T/C	*murC*	Cell wall and cell processes	Non-synonymous	Yes
2436481	G/C	*mptA*	Cell wall and cell processes	Synonymous	Yes
2439204	A/G	*Rv2177c*			Yes
2448966	G/A	*fadD15*	Lipid metabolism	Synonymous	Yes
2480796	G/A	*ephD*	Virulence, detoxification, adaptation	Synonymous	Yes
2489669	A/C	*glnE*	Intermediary metabolism and respiration	Synonymous	Yes
2507254	G/A	*ptpA*	Regulatory proteins	Non-synonymous	No
2565679	C/A	*Rv2294*	Intermediary metabolism and respiration	Non-synonymous	Yes
2596092	G/C	*Rv2323c*	Conserved hypotheticals	Non-synonymous	Yes
2729103	T/C	*Rv2433c*	Conserved hypotheticals	Non-synonymous	Yes
2742477	T/C	*rne*	Information pathways	Synonymous	Yes
2759943	C/T	*mmuM*	Intermediary metabolism and respiration	Synonymous	Yes
2762592	T/G	*clpP2*	Intermediary metabolism and respiration	Non-synonymous	Yes
2788195	C/T	*plsB2*	Lipid metabolism	Synonymous	Yes
2792763	C/T	*lipQ*	Intermediary metabolism and respiration	Non-synonymous	Yes
2809250	C/T	*pdhC*	Intermediary metabolism and respiration	Synonymous	No
2864227	C/T	*aroF*	Intermediary metabolism and respiration	Synonymous	Yes
2884398	C/G	Intergenic region			No
2899426	G/A	*Rv2575*	Cell wall and cell processes	Non-synonymous	Yes
2903169	G/A	*Rv2578c*	Conserved hypotheticals	Synonymous	Yes

2930254	G/A	*Rv2601A*	Virulence, detoxification, adaptation	Non-synonymous	Yes
2955061	T/C	*Rv2628*	Conserved hypotheticals	Non-synonymous	No
3003123	G/A	*arsB1*	Cell wall and cell processes	Non-synonymous	Yes
3005789	T/C	Intergenic region			No
3017134	G/T	*ppgK*	Intermediary metabolism and respiration	Non-synonymous	No
3062858	G/T	*Rv2750*	Intermediary metabolism and respiration	Non-synonymous	Yes
3089689	C/G	*pepR*	Intermediary metabolism and respiration	Non-synonymous	Yes
3117022	T/G	*Rv2812*	Insertion seqs and phages	Non-synonymous	Yes
3164260	G/A	*Rv2854*	Conserved hypotheticals	Non-synonymous	Yes
3168967	C/A	*Rv2857c*	Intermediary metabolism and respiration	Synonymous	No
3190721	C/T	*cdsA*	Lipid metabolism	Non-synonymous	No
3207052	A/G	*Rv2897c*	Conserved hypotheticals	Synonymous	Yes
3207640	C/T	*Rv2897c*	Conserved hypotheticals	Synonymous	Yes
3208502	C/T	Intergenic region			Yes
3232108	C/T	*amt*	Cell wall and cell processes	Synonymous	Yes
3275416	C/G	*papA5*	Lipid metabolism	Synonymous	Yes
3296843	A/G	*pks15*	Lipid metabolism	Non-synonymous	Yes
3296972	C/A	*pks15*	Lipid metabolism	Non-synonymous	Yes
3349758	G/T	*gltS*	Information pathways	Non-synonymous	Yes
3359150	T/C	*Rv3000*	Cell wall and cell processes	Non-synonymous	Yes
3414557	G/C	*nrdI*	Information pathways	Non-synonymous	Yes
3419467	C/T	Intergenic region			No
3427191-3427194	DEL	Intergenic region			Yes
3428818	A/C	*cstA*	Virulence, detoxification, adaptation	Synonymous	Yes
3434713	T/G	*Rv3071*	Conserved hypotheticals	Non-synonymous	Yes
3445130	G/A	*pknK*	Regulatory proteins	Non-synonymous	Yes

3447666	T/C	*virS*	Virulence, detoxification, adaptation	Non-synonymous	Yes
3468628-3468629	AC/GA	*Rv3099c*	Conserved hypotheticals	Non-synonymous	Yes
3469024	G/A	*Rv3099c*	Conserved hypotheticals	Non-synonymous	Yes
3496931	T/C	*Rv3131*	Conserved hypotheticals	Synonymous	Yes
3518501	C/T	*nuoG*	Intermediary metabolism and respiration	Synonymous	Yes
3544390	G/A	*mesT*	Virulence, detoxification, adaptation	Non-synonymous	Yes
3545175	G/A	*mesT*	Virulence, detoxification, adaptation	Synonymous	Yes
3551134-3551144	DEL	Intergenic region			Yes
3563387	C/G	*Rv3194c*	Cell wall and cell processes	Non-synonymous	Yes
3564666	C/T	*Rv3195*	Conserved hypotheticals	Synonymous	Yes
3610610	G/A	*Rv3234c*	Lipid metabolism	STOP	Yes
3701000	C/A	*Rv3312A*	Cell wall and cell processes	Non-synonymous	Yes
3702632	A/C	*deoA*	Intermediary metabolism and respiration	Non-synonymous	Yes
3707594	G/A	Intergenic region			Yes
3719999	C/G	*Rv3333c*	Conserved hypotheticals	Non-synonymous	Yes
3777042	C/T	*Rv3365c*	Conserved hypotheticals	Non-synonymous	Yes
3836739	G/A	*groEl1*	Virulence, detoxification, adaptation	Synonymous	No
3840764	C/G	*alr*	Intermediary metabolism and respiration	Synonymous	Yes
3867085	C/T	*Rv3447c*	Cell wall and cell processes	Non-synonymous	Yes
3896705-3896726	DEL	*Rv3479*	Cell wall and cell processes	Reading frame alteration	Yes
4028292	C/T	*Rv3586*	Conserved hypotheticals	Synonymous	No
4040282	C/T	*clpC1*	Intermediary metabolism and respiration	Synonymous	Yes
4078051	T/C	*Rv3639c*	Conserved hypotheticals	Non-synonymous	Yes
4096712	G/C	*Rv3658c*	Cell wall and cell processes	Synonymous	Yes

4111107	A/G	*Rv3669*	Cell wall and cell processes	Non-synonymous	Yes
4126647	G/A	Intergenic region			Yes
4144477	G/A	*Rv3701c*	Conserved hypotheticals	Non-synonymous	Yes
4145009	G/A	*Rv3702c*	Conserved hypotheticals	Synonymous	Yes
4154165	G/A	*leuA*	Intermediary metabolism and respiration	Synonymous	Yes
4158865	G/A	*cobQ2*	Intermediary metabolism and respiration	Synonymous	Yes
4173128	G/A	*Rv3727*	Intermediary metabolism and respiration	Synonymous	Yes
4188020	A/G	*Rv3737*	Cell wall and cell processes	Non-synonymous	No
4197895	C/G	*Rv3749c*	Conserved hypotheticals	Non-synonymous	Yes
4200576	C/G	*tyrA*	Intermediary metabolism and respiration	Synonymous	Yes
4202069	T/G	*proZ*	Virulence, detoxification, adaptation	Non-synonymous	Yes
4230597	G/A	*Rv3784*	Intermediary metabolism and respiration	Synonymous	Yes
4234664	G/A	Intergenic region			Yes
4237297	G/A	*Rv3791*	Lipid metabolism	Non-synonymous	Yes
4241411	A/G	*embC*	Cell wall and cell processes	Non-synonymous	Yes
4273556	C/A	Intergenic region			Yes
4278358	C/G	Intergenic region			Yes
4284570	A/C	*papA2*	Lipid metabolism	Non-synonymous	No
4307292	C/T	*Rv3833*	Regulatory proteins	Synonymous	Yes
4370867	G/A	*Rv3887c*	Cell wall and cell processes	Non-synonymous	Yes
4388870	C/G	*Rv3903c*	Conserved hypotheticals	Synonymous	Yes

Table S2. Genetic distance based in the number of SNPs among the different MtZ isolates. The ones shadowed in green had ≤ 5 SNPs and therefore were considered recent contact. Case 4 was eliminated because it had too many SNPs.

Esta tabla es demasiado extensa para insertarla en un Word. Se encuentra accesible para su consulta aquí: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9173592/

Publicación 7

scientific reports

OPEN

Analysis of the twenty-six largest outbreaks of tuberculosis in Aragon using whole-genome sequencing for surveillance purposes

Jessica Comín[1]✉, Alberto Cebollada[1], Daniel Ibarz[2], Jesús Viñuelas[3,4], Juan Sahagún[5], Luis Torres[6], María-José Iglesias[2,7,8] & Sofía Samper[1,7,8]

The incidence of tuberculosis in Aragon, Spain, is around ten cases per 100,000 inhabitants. Since 2004, a molecular surveillance protocol has been carried out; therefore, all *M. tuberculosis* strains are genotyped. Recently, whole-genome sequencing has been implemented for relevant isolates. The aim of this work is to characterise at the molecular level the causative strains of the 26 largest outbreaks of the community (including ten or more cases), genotyped by IS*6110*-RFLP and causing 26% of tuberculosis cases. To achieve this objective, two or three isolates of each IS*6110*-cluster belonging to different years were selected for sequencing. We found that strains of lineages L4.8, L4.3 and L4.1.2 were the most frequent. The threshold of 12 SNPs as the maximum distance for confirming the belonging to an outbreak was met for 18 of the 26 IS*6110*-clusters. Four pairs of isolates with more than 90 SNPs were identified as not belonging to the same strain, and four other pairs were kept in doubt as the number of SNPs was close to 12, between 14 and 35. The study of Regions of Difference revealed that they are lineage conserved. Moreover, we could analyse the IS*6110* locations for all genome-sequenced isolates, finding some frequent locations in isolates belonging to the same lineage and certain IS*6110* movements between the paired isolates. In the vast majority, these movements were not captured by the IS*6110*-RFLP pattern. After classifying the genes containing SNP by their functional category, we could confirm that the number of SNPs detected in genes considered as virulence factors and the number of cases the strain produced were not related, suggesting that a particular SNP is more relevant than the number. The characteristics found in the most successful strains in our community could be useful for other researchers in epidemiology, virulence and pathogenesis.

Tuberculosis (TB) is the world's leading infectious disease killer, just surpassed by COVID-19 in 2020. In 2019, 10 million people fell ill with TB and 1.2 million died because of it[1]. The causative agent is *Mycobacterium tuberculosis*, with pulmonary TB being the most frequent presentation of the disease, although extrapulmonary forms can also occur[2].

M. tuberculosis belongs to the *M. tuberculosis* complex (MTBC), which includes eight phylogenetic lineages. L1, L5, L6 and L7 are considered ancient lineages, along with the animal branch, while L2, L3 and L4 are considered modern lineages[3]. Members of L2 and L4 are responsible for the majority of TB cases in the world, and particularly, L4 and its corresponding sub-lineages are the most widespread among our population[4].

Since 2004, a TB surveillance protocol has been carried out in Aragon, Spain, a low-incidence country with around ten cases per 100,000 inhabitants. All *M. tuberculosis* isolates are genotyped by IS*6110*-RFLP and

[1]Instituto Aragonés de Ciencias de la Salud, C/de San Juan Bosco, 13, 50009 Zaragoza, Spain. [2]Universidad de Zaragoza, C/Domingo Miral S/N, 50009 Zaragoza, Spain. [3]Hospital Universitario Miguel Servet, Paseo Isabel la Católica, 1-3, 50009 Zaragoza, Spain. [4]Grupo de Estudio de Infecciones por Micobacterias (GEIM), Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica, C/Agustín de Bentacourt, No 13, 28003 Madrid, Spain. [5]Hospital Clínico Universitario Lozano Blesa, C/ de San Juan Bosco, 15, 50009 Zaragoza, Spain. [6]Hospital San Jorge, Av. Martínez de Velasco, 36, 22004 Huesca, Spain. [7]Fundación IIS Aragón, C/de San Juan Bosco, 13, 50009 Zaragoza, Spain. [8]CIBER de Enfermedades Respiratorias, Av. Monforte de Lemos, 3-5. Pabellón 11, Planta 0, 28029 Madrid, Spain. ✉email: jcomin.iacs@aragon.es

Spoligotyping. As a result, we have a register of all *M. tuberculosis* genotypes and their relatedness (i.e. their belonging to some outbreak).

With the development of whole-genome sequencing (WGS), this technique is proposed to replace the standard molecular typing techniques as WGS has the highest resolution power[5] and is now becoming affordable for investigation laboratories[6].

To implement WGS in our routine laboratory, we planned to sequence two or three representative isolates of the 26 largest outbreaks in our community, which contained at least ten cases. The aim of this work is to characterise these strains at the molecular level and get a general view of the properties of these successful strains to be considered in future surveillance protocols.

Results

With the aim of characterising at the molecular level the responsible strains of the largest outbreaks in our community, comprising 665 out of the 2553 cases registered, we sequenced the genomes of the representative IS*6110*-clustered strains. The dendrogram based on their IS*6110*-RFLP patterns is shown in Fig. 1, and the MIRU-VNTR patterns and spoligotypes are detailed in Table S1.

General views. The 26 IS*6110*-outbreaks studied ranged from 10 to 178 cases since 2004. Seven outbreaks were caused by L4.8 strains producing 291 cases. These include CLS_7[7], the largest outbreak we ever had, with 242 cases from 1993 to 2020 (no data for 1996–2000). Eight outbreaks were caused by L4.1.2/Haarlem strains producing 170 cases. Six outbreaks were produced by L4.3/LAM strains (two by a L4.3.2 strain, three by a L4.3.3 strain and one by a L4.3.4 strain), producing 121 cases. These included CLS_217, which was independently studied[8]. The rest of the outbreaks were caused by strains from different lineages: one by a L4.1.1.3/X strain, with 21 cases[9]; one by a L4.7 strain, with 19 cases; one by a L4.4.1.1 strain, with 16 cases; one by a L4.9 strain, producing 13 cases; and one by a L4.6.1.1 strain, with 14 cases.

SNP distances. It has been proposed that a distance of ≤ 5 SNPs between two isolates is considered recent contact and that 12 SNPs should be the maximum distance to consider both isolates to be the same strain and therefore the same outbreak[10]. Eighteen out of the 26 IS*6110*-clusters studied fit this threshold, indicating that the sequenced isolates belong to the same WGS-outbreak. We found a distance of 1–10 SNPs among the two or three isolates sequenced for 18 of the IS*6110*-clusters studied (Table 1). In Table S2, a description of the SNPs found in these clusters (point, gene, type of mutation, effect of the mutation) can be found. For CLS_15, lineage identification revealed that one selected isolate belonged to L4.1.2.1 and the other to L4.8. A new revision of the genotype patterns confirmed that the correct one was L4.1.2.1, so the L4.8 isolate was a selection error. The SNP distance could not be studied for this IS*6110*-cluster. For three clusters, the SNP distance was large enough to guarantee that the selected isolates were not the same strain. CLS_2 had an SNP distance of 145, CLS_49 of 143 among the three sequenced isolates, and CLS_26 of 93. On the other hand, four clusters had a distance higher than 12 SNPs but close. CLS 157 had a distance of 34 SNPs between the two sequenced isolates. CLS_47 had an SNP distance of 35 among the three sequenced isolates; however, the majority of SNPs were due to one of the isolates, with the other two being more related (less than 12 SNPs). Finally, CLS_119 had an SNP distance of 18 and CLS_9 of 14 (Tables 1, Table S2).

Regions of Difference (RDs) study. We looked for large deletions (Regions of Difference or RDs)[11] to find differences between the clustered strains and non-clustered strains previously analysed in our laboratory of the same lineage. According to Coll et al.[12], RD182 is specific to L4.1.2.1, RD219 is specific to L4.8, RD115 is specific to L4.3.3 and RD724 is specific to L4.6.1.1. All clustered and non-clustered strains were concordant with these specific characteristics. No different RD was found among the clustered and non-clustered strains, with the majority of them being lineage conserved. In addition, the RDs of the isolates belonging to the same IS*6110*-outbreak were the same even in those with a high SNP distance. We only found one different large deletion, not previously described as an RD, between the isolates of CLS_119: one of the isolates had *Rv3054c-Rv3055-dinP-Rv3057* genes deleted, while the other conserved this region as the reference strain. The RDs of the different strains are shown in Table 2.

IS6110 locations. WGS allowed us to locate all the IS*6110* copies in the genomes of the clustered strains. The highest number of copies was found among the strains belonging to L4.3 (average number of IS*6110* copies = 16.3), followed by the L4.8 strains (12.6) and the L4.1.2.1 strains (11.4). The description with the exact locations in all the isolates studied is in Table S3. For the L4.3 strains, three copies were present in all the strains studied—*lpqQ:Rv0836c, Rv1754c* (RD152 area), and *Rv3113*. Moreover, copies located at *cut1*, *ppe38* and *MT3426:MT3427* were frequent. For L4.1.2.1, five copies located at *Rv0403c*, *Rv2336*, *Rv1754c*, *Rv0963c* and *MT3429* were present in all the strains studied. We observed the same locations for other sequenced non-clustered strains of the same lineages (L4.3 and L4.1.2.1). For the L4.8 strains, copies within *MT3429*, *Rv1668:Rv1669c* and *Rv1762c:Rv1763c* (RD152 area) were present in all the strains studied, while copies located at *ppe71* and *Rv0795-Rv0796* were frequent. A summary with the common and frequent IS*6110* copies in the different lineages is shown in Table 3.

We observed some IS*6110* movements among the clustered isolates studied. In five of them, additional copies were detected in the later isolate. In CLS_21, the isolate from 2018 had an extra IS in the DR region that was not present in the isolate from 2007. In CLS_13, the isolate from 2015 had two extra IS copies located in *ppe28:ppe29* and *Rv0756c* that were absent in the 2004 isolate. In CLS_93, the isolate from 2019 had an extra IS copy located in *Rv3177*, absent in the isolate from 2008. In CLS_71, the isolate from 2020 had two extra IS copies located in *Rv1371* and *phoT* that were not present in the isolate from 2007. Finally, in CLS_152, the isolate from 2020 had

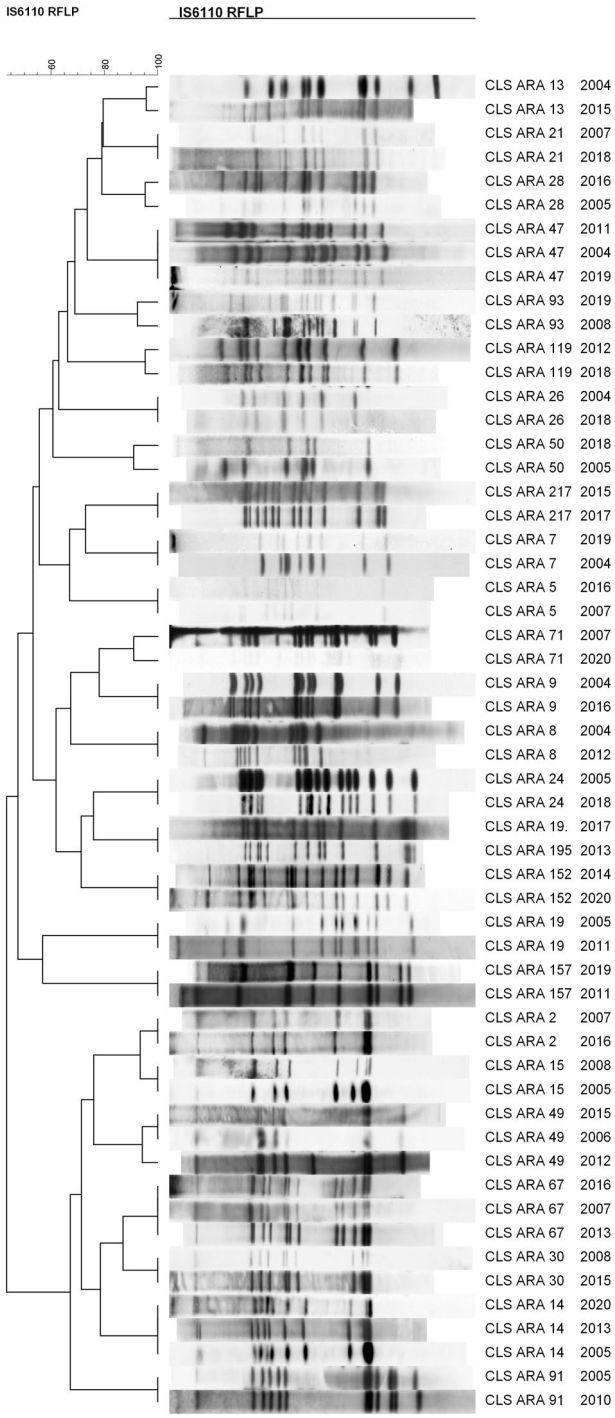


Figure 1. Dendrogram showing the IS*6110*-RFLP patterns of the selected isolates of the different outbreaks.

an extra IS*6110* in *Rv1730c:gabD2* that was absent in the 2014 isolate. On the other hand, in four cases, an extra IS*6110* was detected in the earlier isolate. The CLS_2 isolate of 2007 had an extra IS*6110* located in *Rv3183:Rv3184* that was not in the later isolate of 2016. In CLS_24, the isolate from 2005 had an extra IS*6110* located at *MT3426* that was absent in the 2018 isolate. In CLS 50, the isolate from 2005 had an extra IS copy located in *MT3427*. Lastly, in CLS_91, the isolate from 2005 had an IS*6110* located at the *plcA* gene that was absent in the isolate from 2010. CLS_49 was a special case: the isolate from 2012 had an extra IS inserted within the *Rv1765c* gene (or its homologous *Rv2015c*) that was not present in the other two isolates studied, and the isolate from 2015 had an extra IS located at *ppe49* that was absent in the other two isolates of this cluster. This is in accordance with the large number of SNPs observed among these isolates (143 SNPs). The same could be applied to CLS_2, with an SNP distance of 145. On the other hand, we found identical number of IS*6110* copies and locations for the isolates

Cluster	Lineage	N	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	SNPs
CLS_2	4.1.2.1	29	3	5	2	**7**	5	1	1	1	2	0	0	0	**2**	0	0	0	0	>90
CLS_5	4.4.1.1	16	2	2	0	**5**	0	1	0	1	1	0	0	1	**1**	0	0	2	0	≤5
CLS_7*	4.8	178	**34**	14	18	22	20	9	12	11	5	4	7	8	3	3	5	**2**	1	<12
CLS_8	4.3.4.1	19	**6**	2	2	3	1	1	1	1	**1**	1	0	0	0	0	0	0	0	<5
CLS_9	4.3.3	15	**3**	3	1	2	2	0	0	2	0	0	0	0	**2**	0	0	0	0	14–35
CLS_13	4.8	20	**6**	6	2	1	0	0	1	0	1	1	0	**1**	0	1	0	0	0	≤5
CLS_14	4.1.2.1	42	2	**13**	2	4	7	3	3	1	2	**3**	0	0	0	0	0	0	**2**	<12
CLS_15	4.1.2.1	12	2	**6**	1	1	**1**	0	0	0	1	0	0	0	0	0	0	0	0	>90
CLS_19	4.9	13	1	**5**	0	1	1	1	1	**2**	0	0	0	0	0	1	0	0	0	≤5
CLS_21	4.8	35	0	4	3	**8**	0	0	6	1	1	5	2	2	1	0	**1**	0	1	≤5
CLS_24	4.3.2	39	8	**9**	1	3	0	4	1	1	3	1	1	1	0	4	**1**	1	0	<12
CLS_26	4.7	19	**2**	3	1	1	0	1	0	1	1	1	2	1	0	1	**3**	0	1	>90
CLS_28	4.8	13	0	**4**	1	1	0	1	0	1	1	0	1	1	**1**	0	1	0	0	≤5
CLS_30	4.1.2.1	35	2	2	1	1	**10**	1	3	5	2	2	0	**4**	0	0	0	2	0	≤5
CLS_47	4.8	19	**2**	0	2	1	2	1	1	**1**	1	0	2	1	1	1	0	**3**	0	14–35
CLS_49	4.1.2.1	11	0	1	**2**	1	2	0	1	1	**2**	0	0	**1**	0	0	0	0	0	>90
CLS_50*	4.1.1.3	21	0	**2**	6	0	4	2	1	1	0	1	2	0	0	0	**2**	0	0	<12
CLS_67	4.1.2.1	19	0	2	0	**3**	3	1	1	0	1	**2**	0	3	**1**	1	1	0	0	≤5
CLS_71	4.3.3	19	0	0	1	**2**	3	2	0	0	0	0	3	0	2	1	0	3	**2**	<12
CLS_91	4.1.2.1	12	0	**1**	0	0	2	0	**3**	3	1	0	1	1	0	0	0	0	0	<12
CLS_93	4.8	14	0	0	0	0	**3**	2	1	0	0	2	0	0	1	2	0	**3**	0	<12
CLS_119	4.8	12	0	0	0	0	0	2	0	1	**4**	1	0	0	2	0	**1**	1	0	14–35
CLS_152	4.6.1.1	14	0	0	0	0	0	0	2	1	1	0	**5**	0	0	0	0	2	**3**	≤5
CLS_157	4.1.2	10	0	0	0	0	0	0	1	**3**	1	1	1	1	0	0	0	**1**	1	14–35
CLS_195	4.3.2	15	0	0	0	0	0	0	0	0	0	**6**	2	1	5	**1**	0	0	0	≤5
CLS_217*	4.3.3	14	0	0	0	0	0	0	0	0	0	0	1	**3**	6	**3**	0	1	0	≤5

Table 1. Lineage and number of cases (N) distributed per year of each IS*6110*-cluster. The bold indicates the year of the selected isolates of each outbreak for this genomic study. *Cluster independently studied; therefore, a greater number of cases distributed throughout the duration of the outbreak were sequenced. The column "SNPs" shows the SNP distance between the selected isolates of each cluster.

that apparently were of the same RFLP cluster but, according to the SNP distance, would not be. This occurred for CLS_9, CLS_26, CLS_47, CLS_119 and CLS_157, although the SNP distance for CLS_9 and CLS_119 was very close to 12 (14 and 18 SNPs, respectively). The IS*6110* movements are shown in Table S3.

Some of the deletions observed could be explained by a homologous recombination between two close IS*6110* resulting in the loss of the intermediate genes as one IS*6110* was found interrupting the genes involved. This was observed for CLS_30 isolates, showing the loss of *Rv2817c:Rv2819c* genes; CLS_157 and the deletion of *plcC-plcB-plcA* genes; CLS_13 and the deletion between *plcC* and *ppe40*; CLS_5, CLS_7, CLS_28, CLS_47 and CLS_119, and the deletion between *cut1* and *Rv1765c*; CLS_93 and the deletion between *ppe38* and *ppe71*; CLS_8 and CLS_217, and the deletion of *plcD* gene; CLS_26 and the deletion between *plcD* and *Rv1765c*; CLS_9 and the *Rv2810c:Rv2813* deletion; and CLS_152 and the deletion of *Rv2018* and *Rv2019*.

SNPs in virulence factors. All the L4.1.2.1 strains, including the non-clustered strains reviewed, had 14 common SNPs in genes considered as virulence factors (*mce2D*, *ctpV*, *Rv1290c*, *Rv1505c*, *pks5*, *mgtC*, *secA2*, *Rv1915*, *Rv1915*, *Rv1982c*, *pks12*, *Rv2494*, *cyp125* and *esxB*)[13]. The L4.3 strains had four common SNPs (*mce1D*, *Rv0990c*, *mce3F* and *fadE28*). In addition, there were four more common SNPs for the L4.3.2 strains (*Rv0204c*, *Rv1939*, *Rv3085* and *Rv3871*) and three for the L4.3.3 strains (*mmaA4*, *pks12* and *Rv3088*). The L4.8 strains shared only one common SNP in a gene considered as a virulence factor (*ptpA*). Looking into the isolates studied in the rest of the sub-lineages, the average SNP number was eight, all being unique SNPs. However, there was no relation between the number of SNPs in the genes considered as virulence factors and the number of cases of the IS*6110*-cluster (p-value = 0.8) (Fig. 2).

As these common sub-lineage SNPs were present in the clustered strains as well as non-clustered strains, we also decided to focus on the specific SNPs of each WGS-cluster strain. The L4.1.2.1 strains turned out to have less specific SNPs (an average of 3.6), followed by the L4.3 (4.7) and L4.8 (5.9) strains. Graphics with the SNPs classified according to the gene categories that Forrellad et al. and Ramage et al.[13,14] described are shown in Fig. 3. The L4.8 strains had more SNPs in the cell wall proteins (usually *mce* genes), synthesis of the complex lipids and toxin/antitoxin systems categories. For the L4.3 strains, the cell wall proteins category stands out above the rest. For the L4.1.2.1 strains, the secretion systems, cell wall proteins and toxin/antitoxin systems categories have more SNPs. A detailed list of the specific SNPs and descriptions of the genes can be found in Table S4. Regarding the

		RD109c	RD115	RD145	RD149	RD152	RD168	RD174 (Rio)	RD178	RD182	RD188	RD193	RD207	RD219	RD252	RD724
L4.1.2	CLS_2			X					X	X						
	CLS_14			X					X	X						
	CLS_15			X					X	X						
	CLS_30			X	X				X	X			X			
	CLS_49			X	X				X	X						
	CLS_67			X					X	X						
	CLS_91			X					X	X						
	CLS_157				X						X					
L4.8	CLS_7					X								X		
	CLS_13										X			X		
	CLS_21													X		
	CLS_28					X	X							X		
	CLS_47					X								X		
	CLS_93													X	X	
	CLS_119					X								X		
L4.3.2	CLS_24				X											
	CLS_195				X	X										
L4.3.3	CLS_9	X	X		X	X			X							
	CLS_71	X	X		X	X			X							
	CLS_217	X	X		X	X			X							
L4.3.4.1	CLS_8				X	X		X								
L4.1.1.3	CLS_50											X			X	
L4.7	CLS26				X	X										
L4.4.1.1	CLS_5					X										
L4.9	CLS_19															
L4.6.1.1	CLS_152															X

Table 2. RDs of the different IS*6110*-clustered strains studied. X implies the deletion of the genes.

	5'-IS point	3'-IS point	Direct repeat	Gene	Direction	
L4.3	**3,480,373**	**3,480,371**	**cag**	***Rv3113***	**Forward**	100% strains
	932,202	**932,204**	**aac**	***lpqQ:Rv0836c***	**Reverse**	
	1,987,457	**1,986,625**	**–**	***plcD/Rv1754c***	**Reverse**	
	1,989,080	**1,986,625**	**–**	***cut1/Rv1754c***	**Reverse**	
	1,986,623	**1,986,625**	**aac**	***Rv1754c***	**Reverse**	
	1,989,080	*1,979,923**	*cgc*	*cut1*	*Reverse*	81.8% strains
	1,989,080	*1,986,625*	*–*	*cut1/Rv1754c*	*Reverse*	
	2,633,843	*2,633,841*	*ctc*	*ppe38*	*Reverse*	45.5% strains
	*3,665,157**	*3,665,159**	*caa*	*MT3426:MT3427*	*Reverse*	91% strains
	*3,665,159**	*3,668,981**	*–*	*MT3426:MT3427/IS1547*	*Forward*	
L4.8	**3,668,619***	**3,668,723***	**–**	***MT3429***	**Forward**	100% strains
	1,895,651	**1,895,654**	**ccta**	***Rv1668c:Rv1669***	**Reverse**	
	1,986,939*	**1,986,937***	**acc**	***Rv1762c:Rv1763***	**Forward**	
	1,979,901*	**1,986,937***	**–**	***cut1/Rv1762c:Rv1763***	**Forward**	
	*2,604,207**	*2,604,210**	*gaaa*	*ppe71*	*Reverse*	85.7% strains
	889,020	*890,375*	*gagg*	*Rv0795-Rv0796*	*Forward*	85.7% strains
L4.1.2.1	**1,075,948**	**1,075,950**	**acc**	***Rv0963c***	**Reverse**	100% strains
	1,986,626	**1,986,622**	**tgttc**	***Rv1754c***	**Forward**	
	483,296	**483,298**	**agg**	***Rv0403c***	**Reverse**	
	2,610,863	**2,610,861**	**gcc**	***Rv2336***	**Forward**	
	3,668,575*	**3,668,756***	**–**	***MT3429***	**Forward**	

Table 3. Common and frequent IS*6110* locations in the strains studied, grouped by lineage. The IS*6110* locations present in all the strains studied belonging to that lineage are in bold, and the frequent locations found are in italic. *Points referred to the *M. bovis* genome.

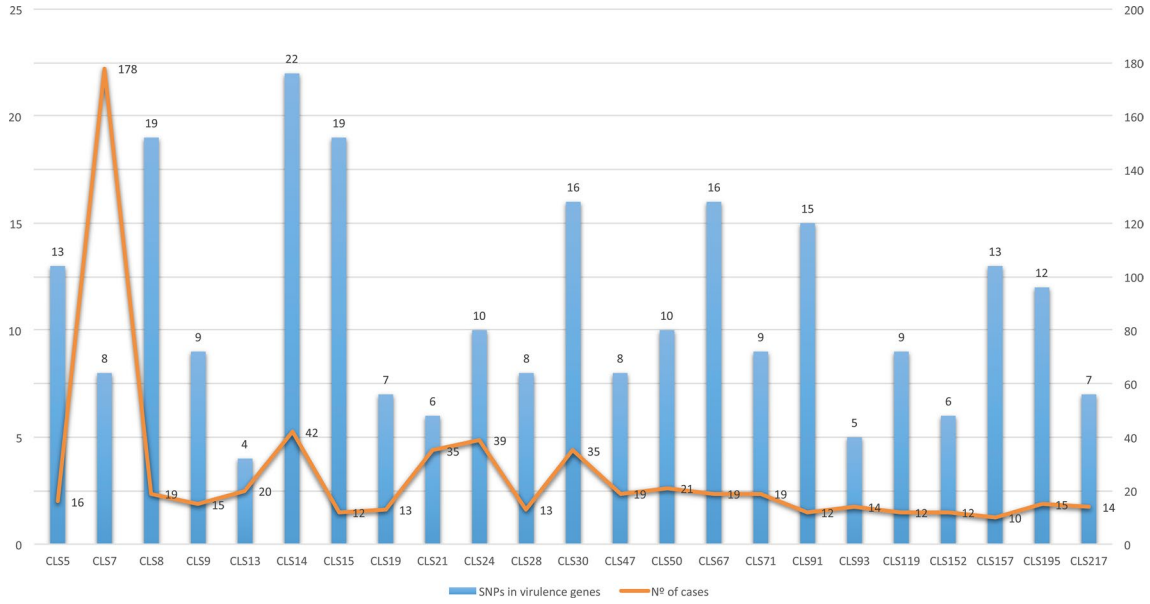


Figure 2. Number of cases of each IS*6110*-cluster vs. SNP number in genes considered as virulence factors. No cause–effect relationship is observed (p-value = 0.8). Clusters with more than 90 SNPs were not considered.

WGS-clusters from other lineages, the synthesis of complex lipids, cell wall proteins and toxin/antitoxin systems categories had more SNPs.

Discussion

We describe in this work the molecular characteristics of *M. tuberculosis* strains that produced the largest outbreaks in our community. Three of these outbreak strains were previously characterised[7–9] (CLS_ 217, CLS_50 and CLS_7). The majority of publications about TB outbreaks based on WGS focus on the power of this methodology to determine whether the isolates belong to the same cluster and/or attempt to elucidate the transmission chain[10,15]. However, few articles focus on the molecular characteristics of relevant strains using WGS[16].

A total of 665 TB cases had been caused by some of these studied strains since 2004, which represents 26% of the total TB cases in our community. As expected, all these strains belonged to L4, which is the lineage responsible for the majority of TB cases in our population[4]. The most widespread lineages were L4.8, L4.1.2.1/Haarlem and L4.3/LAM, in concordance with the dominant lineages in Spain described by Stucki et al.[4].

Regarding the SNP distance, the WGS results were in agreement with the RFLP-IS*6110* patterns for 18 outbreaks, including the three previously studied [7–9]: the genetic distance was between 1 and 10 SNPs, so they were considered epidemiologically linked isolates[10,17,18]. However, the classical gold-standard RFLP technic failed to cluster the isolates of CLS_2, CLS_49 and CLS_26. For CLS_2, there was an extra IS*6110* in one of the isolates that was not captured in the RFLP patterns. For CLS_49, two isolates had an identical RFLP pattern, while the third had an extra band, although initially, it was considered the same strain that had evolved. WGS confirmed this extra IS*6110* for this isolate and a different one for the other isolates that RFPL did not show. As for CLS_26, the RFLP pattern was the same but also the IS*6110* locations found by WGS, a total of six, which is the threshold indicated for discrimination by RFLP pattern. Thanks to the greater resolution power of WGS[19], we deciphered that these IS*6110*-clustered isolates were not epidemiologically related as the SNP distance was more than 90 despite sharing similar IS locations.

On the other hand, WGS did not clarify the relatedness in CLS_9, CLS_47, CLS_119 and CLS_157 using only the SNP distance information. These IS*6110*-clusters have distances of more than 12 SNPs but are very close. Furthermore, the IS*6110* locations found by WGS were exactly the same, along with their RFPL patterns (16, 13, 12 and 11 IS*6110*, respectively). The number of bands is large/high enough to be considered a fluke, reflecting that both isolates corresponded to the same strain despite the number of SNPs being ≥ 12. There is a fact about CLS_119 that would support the idea of separating one isolate from its pair: the deletion of the *Rv3054:Rv3058* region in one of the isolates but not in the other. We considered this deletion, which could not be explained by an IS*6110* recombination, an independent event that took place in this particular isolate. It is important to consider that the traceability between the selected isolates is not defined, making the intermediate cases that have occurred between them unknown. Although this is not frequent, strains can evolve during the transmission process. If the isolates of CLS_119 are not epidemiologically linked, a genetic separation between them must have taken place recently. It would be interesting to analyse the complete isolates of this cluster to understand their evolution, as well as the rest of the clusters to track for recent transmission. However, the 12 SNP threshold should be polished to clarify these intermediate SNP distances as for now, the significance of their relatedness is unclear.

One important fact is that the mutation rate is not the same for all MTBC strains. In the independent studies we made for CLS_7 and CLS_50[7,9], we found that all the isolates studied had ≤ 12 SNPs with at least one isolate, confirming its belonging to the outbreak, although the strains had been circulating for more than 25 years.

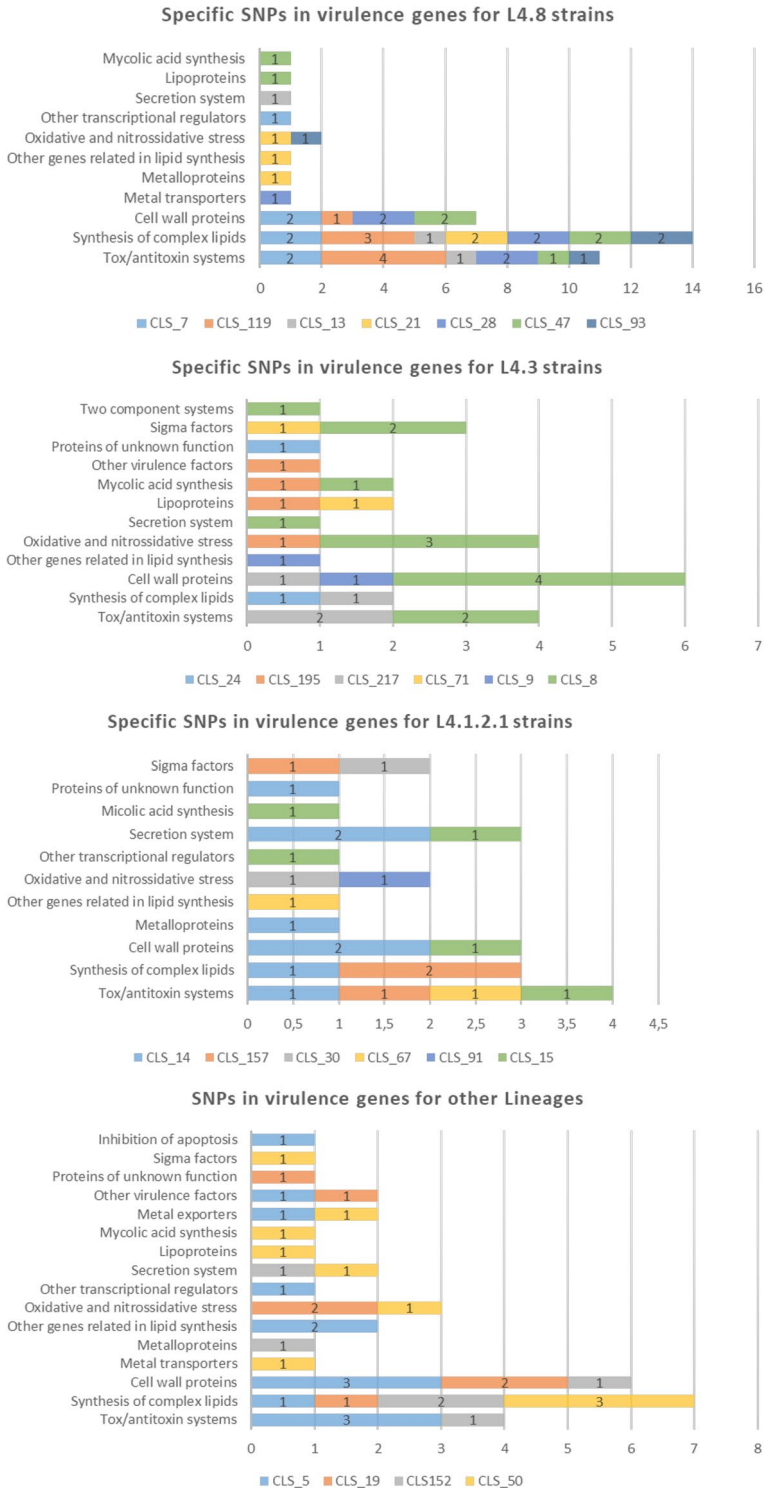


Figure 3. Specific SNPs in virulence genes for each WGS-cluster classified by lineage according to the categories of Forrellad et al. and Ramage et al.[13,14]. The number of categories remains constant among the three main lineages studied (L4.8, L4.3 and L4.1.2.1). The oxidative and nitrossidative stress, other genes related to lipid synthesis, cell wall proteins, synthesis of complex lipids and toxin/antitoxin systems categories are present in the three lineages. For clusters belonging to the L4.4.1.1, L4.9, L4.6.1.1 and L4.1.1.3/X families, there were no common and specific SNPs as only one cluster of each lineage had been sequenced, so comparison was not possible. The categories with more SNPs for these lineages were cell wall proteins, the synthesis of complex lipids and toxin/antitoxin systems. Clusters with more than 90 SNPs were not considered.

Some isolates presented several unique SNPs, which could be due to sequencing errors or favoured by the epidemiological characteristics of the patient, which could increase the SNP distance. For the outbreaks studied in this work, we only analysed two or three isolates; therefore, the SNP distance with some other isolates of the IS*6110*-cluster could be ≤ 12, indicating its membership to the outbreak. This could be the case of CLS_9, with an SNP distance of 14.

The RD study showed that the majority of RDs are lineage conserved, so they are present in clustered as well as non-clustered strains. Nevertheless, RD152 and RD188 are distributed among different sub-lineages, which may be due to both regions containing hot spots for IS*6110* insertions, causing later recombination events[20–22]. RD178 (a fragment of the *helZ* gene), RD252 and RD149 are also distributed among different lineages, especially RD149. This deletion has been associated with a reduction in growth and an increase in the induction of TNFα in host cells by CAS1 strains[23]. However, as this RD is also present in the non-clustered strains, this may not be responsible for the success in transmission of the outbreak strains.

WGS allowed an easier study of the IS*6110* location than the molecular techniques based on PCR, previously used. We found some lineage-conserved locations, in accordance with those described before[24]. It called our attention to the presence of an IS*6110* copy in the *ppe38/ppe71* locus in many of the strains studied (67.4%). The deletion of this locus has been related to virulence in L2/Beijing strains[25], so the disruption of these genes by an IS*6110* could be an advantage for the mycobacteria. This could be the reason why so much polymorphism is observed in this region. The study of the IS also allowed us to discover movements of this mobile element between the different isolates for 10 of the clusters studied. Some of them were captured by the RFLP pattern, but others were not. These observations lead us to believe that IS*6110* transposition could occur during infection, although majority of the time, it would fail, provoking some handicap to the bacteria.

We decided to study SNPs in genes considered virulence factors as we previously observed that a single SNP in one of these genes could be responsible for higher transmission[7,9]. We focused on the specific SNPs of each cluster strain as the ones present in both clustered and non-clustered strains would not be related to higher virulence or transmission. We observed no relationship between the number of SNPs in these genes and the number of cases the strain produced, so we concluded that a particular SNP is more important than the number. We describe these SNPs in Table S4, but more research is required to determine whether some of them are responsible for the increase of transmission of these outbreak strains. SNPs in the categories oxidative and nitrossidative stress, other genes related in lipid synthesis, cell wall proteins, synthesis of complex lipids and toxin/antitoxin systems are present in almost all the strains studied, so those genes may be important for the success of the outbreak strains.

In conclusion, we describe in this work the molecular characteristics—lineage, presence or absence of RD, IS*6110* locations and SNPs in virulence factors—of the most successful strains in our population. We give value to the classical techniques maintained along the time to track the pathway of these strains. We are deep into evolution and have identified potential outbreak features, shared by some of these strains, to develop surveillance actions.

Materials and methods

Clinical sample selection. All IS*6110*-RFLP and Spoligo patterns of the *M. tuberculosis* isolates from 2004 to 2020 were loaded in the Bionumerics database (v7.6, Applied Maths, Kortrijk, Belgium). Among the 2553 genotyped isolates, we selected two or three isolates belonging to each of the IS*6110*-clusters involving 10 or more cases of TB: CLS 2 (one isolate from 2007 and one from 2016), CLS 5 (2007/2016), CLS 7 (2004/2019), CLS 8 (2004/2012), CLS 9 (2004/2016), CLS 13 (2004/2015), CLS 14 (2005/2013/2020), CLS 15 (2005/2008), CLS 19 (2005/2011), CLS 21 (2007/2018), CLS 24 (2005/2018), CLS 26 (2004/2018), CLS 28 (2005/2016), CLS 30 (2008/2015), CLS 47 (2004/2011/2019), CLS 49 (2006/2012/2015), CLS 50 (2005/2018), CLS 67 (2007/2013/2016), CLS 71 (2007/2020), CLS 91 (2005/2010), CLS 93 (2008/2019), CLS 119 (2012/2018), CLS 152 (2014/2020), CLS 157 (2011/2019), CLS 195 (2013/2017) and CLS 217 (2015/2017). Three of these clusters were independently studied by WGS: CLS 7[7] (57 isolates sequenced), CLS 50[9] (32 isolates sequenced), and CLS 217[8] (13 isolates sequenced).

Thirteen genomes from different sub-lineages, six from L4.1.2 and seven from L4.3, not involved in any of the large outbreaks (non-clustered strains) and sequenced in previous studies, were used to make the comparisons with the clustered strains belonging to these sub-lineages.

DNA extraction and classical genotyping. DNA was extracted from bacterial cultures using the cetrimonium bromide method described by van Soolingen[26]. The IS*6110*-RFLP and Spoligo genotyping were made for all the isolates, as previously described[27,28]. An IS*6110*-cluster was defined as having the same or similar RFLP pattern (one extra IS*6110* band was accepted if some epidemiological link was found). Mycobacterial interspersed repetitive unit-variable number of tandem repeats (MIRU-VNTR) was performed for one isolate of each cluster[29]. The DNA samples were stored at − 20 °C until sequencing.

Whole-genome sequencing (WGS). The majority of DNA (50 isolates) were sequenced using Illumina technology. However, some DNA used in previous studies (six isolates) were sequenced using the IonTorrent sequencing platform. Both technologies were applied according to the manufacturer's instructions. After sequencing, the fastQ files obtained were mapped against the reference strain H37Rv (NC_000962.3) to obtain the Binary Aligned Map (bam) and the Variant Call Format (vcf) files. The SNP classification established by Coll et al.[12] was used for identifying the MTBC lineage of the outbreak strains. This classification is based on the specific SNPs of the different lineages.

Bioinformatics. The Bionumerics software was used for the SNP study and for the construction of the dendrograms using the UPGMA method. For greater accuracy, strict SNP filtering that removed positions with at least one ambiguous or unreliable base, gaps (maximun frequency 1%), non-discriminatory positions and *ppe* and *pgrs* genes, was applied. It was also considered that the retained SNP positions had a minimum 5× coverage and that the minimum distance between SNPs was at least 12 base pairs (bp). The Integrative Genomics Viewer (IGV, from the Broad Institute[30]) software was used for the RD study and the SNP study. GeneWise (https://www.ebi.ac.uk/Tools/psa/genewise/) and PROVEAN (http://provean.jcvi.org/seq_submit.php) platforms were used for the SNP study, predicting whether an SNP is synonymous or non-synonymous and the effect of a non-synonymous mutation is neutral or deleterious, respectively. All SNPs are referred to H37Rv genome (NC_000962.3), unless otherwise indicated.

From the fastQ files, the reads containing the first 30 bases of the IS*6110* and the ones with the last 30 were extracted. We used Tuberculist (http://genolist.pasteur.fr/TubercuList/) and Bovilist (http://genolist.pasteur.fr/BoviList/) to apply BLAST and find the IS*6110* insertion points. Mycobrowser (https://mycobrowser.epfl.ch/) and UniProtKB (https://www.uniprot.org/uniprot/) websites were used to find information on the genes and proteins with noteworthy SNPs.

Data availability

The fastq files of the selected isolates of each outbreak are uploaded in GenBank with the accession numbers SAMN26722357-SAMN26722406 (BioProject accession number PRJNA816739). https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA816739.

Received: 29 June 2022; Accepted: 30 October 2022
Published online: 05 November 2022

References

1. WHO. *WHO | Global tuberculosis report 2019. Geneva: World Health Organization*. Licence: CC BY-NC-SA 3.0 IGO (2019).
2. García-Rodríguez, J. F. *et al.* Extrapulmonary tuberculosis: Epidemiology and risk factors. *Enferm. Infecc. Microbiol. Clin.* **29**(7), 502–509. https://doi.org/10.1016/j.eimc.2011.03.005 (2011).
3. Gagneux, S. *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* https://doi.org/10.1073/pnas.0511240103 (2006).
4. Stucki, D. *et al. Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48**(12), 1535–1543. https://doi.org/10.1038/ng.3704 (2016).
5. Nikolayevskyy, V. *et al.* Role and value of whole genome sequencing in studying tuberculosis transmission. *Clin. Microbiol. Infect.* **25**(11), 1377–1382. https://doi.org/10.1016/j.cmi.2019.03.022 (2019).
6. Cirillo, D. M. *et al.* Use of WGS in *Mycobacterium tuberculosis* routine diagnosis. *Int. J. Mycobacteriol.* https://doi.org/10.1016/J.IJMYCO.2016.09.053 (2016).
7. Comín, J. *et al.* The MtZ strain: Molecular characteristics and outbreak investigation of the most successful *Mycobacterium tuberculosis* strain in Aragon using whole-genome sequencing. *Front. Cell. Infect. Microbiol.* **12**, 887134. https://doi.org/10.3389/fcimb.2022.887134 (2022).
8. Comin, J. *et al.* Investigation of a rapidly spreading tuberculosis outbreak using whole-genome sequencing. *Infect. Genet. Evol.* https://doi.org/10.1016/j.meegid.2020.104184 (2020).
9. Comín, J. *et al.* A whole-genome sequencing study of an X-family tuberculosis outbreak focus on transmission chain along 25 years. *Tuberculosis* https://doi.org/10.1016/j.tube.2020.102022 (2021).
10. Lalor, M. K. *et al.* The use of whole-genome sequencing in cluster investigation of a multidrug-resistant tuberculosis outbreak. *Eur. Respir. J.* **51**(6), 1702313. https://doi.org/10.1183/13993003.02313-2017 (2018).
11. Tsolaki, A. G. *et al.* Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains. *Proc. Natl. Acad. Sci.* https://doi.org/10.1073/pnas.0305634101 (2004).
12. Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* **5**, 4–8. https://doi.org/10.1038/ncomms5812 (2014).
13. Forrellad, M. A. *et al.* Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence* https://doi.org/10.4161/viru.22329 (2013).
14. Ramage, H. R., Connolly, L. E. & Cox, J. S. Comprehensive functional analysis of *Mycobacterium tuberculosis* toxin-antitoxin systems: Implications for pathogenesis, stress responses and evolution. *PLoS Genet.* **5**(12), e1000767. https://doi.org/10.1371/journal.pgen.1000767 (2009).
15. Walker, T. M. *et al.* Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: An observational study. *Lancet Respir. Med.* https://doi.org/10.1016/S2213-2600(14)70027-X (2014).
16. Folkvardsen, D. B. *et al.* Genomic epidemiology of a major *Mycobacterium tuberculosis* outbreak: Retrospective cohort study in a low-incidence setting using sparse time-series sampling. *J. Infect. Dis.* **216**(3), 366–374. https://doi.org/10.1093/infdis/jix298 (2017).
17. Casali, N. *et al.* Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: A retrospective observational study. *PLoS Med.* **13**(10), 1–18. https://doi.org/10.1371/journal.pmed.1002137 (2016).
18. Hatherell, H. A. *et al.* Interpreting whole genome sequencing for investigating tuberculosis transmission: A systematic review. *BMC Med.* https://doi.org/10.1186/s12916-016-0566-x (2016).
19. Meehan, C. J. *et al.* The relationship between transmission time and clustering methods in *Mycobacterium tuberculosis* epidemiology. *EBioMedicine* **37**, 410–416. https://doi.org/10.1016/j.ebiom.2018.10.013 (2018).
20. Ho, T. B. L., Robertson, B. D., Taylor, G. M., Shaw, R. J. & Young, D. B. Comparison of *Mycobacterium tuberculosis* genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. *Yeast* **1**(4), 272–282. https://doi.org/10.1002/1097-0061(200012)17:4%3c272::AID-YEA48%3e3.0.CO;2-2 (2000).
21. Vera-Cabrera, L., Hernández-Vera, M. A., Welsh, O., Johnson, W. M. & Castro-Garza, J. Phospholipase region of *Mycobacterium tuberculosis* is a preferential locus for IS6110 transposition. *J. Clin. Microbiol.* https://doi.org/10.1128/JCM.39.10.3499-3504.2001 (2001).
22. Vera-Cabrera, L. *et al.* Genetic characterization of *Mycobacterium tuberculosis* clinical isolates with deletions in the plcA-plcB-plcC locus. *Tuberculosis* **87**(1), 21–29. https://doi.org/10.1016/j.tube.2006.01.023 (2007).
23. Kanji, A. *et al.* Presence of RD149 deletions in *M. tuberculosis* Central Asian Strain 1 isolates affect growth and TNFα induction in THP-1 monocytes. *PloS One* **6**(8), e24178. https://doi.org/10.1371/journal.pone.0024178 (2011).

24. Reyes, A. *et al.* IS-seq: A novel high throughput survey of in vivo IS6110 transposition in multiple *Mycobacterium tuberculosis* genomes. *BMC Genom.* https://doi.org/10.1186/1471-2164-13-249 (2012).
25. Ates, L. S. *et al.* Mutations in ppe38 block PE-PGRS secretion and increase virulence of *Mycobacterium tuberculosis*. *Nat. Microbiol.* **3**(2), 181–188. https://doi.org/10.1038/s41564-017-0090-6 (2018).
26. van Soolingen, D., de Haas, P. E., Hermans, P. W. & van Embden, J. D. DNA fingerprinting of *Mycobacterium tuberculosis*. *Methods Enzymol.* **235**, 196–205 (1994).
27 Van Embden, J. D. A. *et al.* Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: Recommendations for a standardized methodology. *J. Clin. Microbiol.* **31**, 406–409 (1993).
28. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* https://doi.org/10.1128/jcm.35.4.907-914.1997 (1997).
29. Supply, P. *et al.* Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* https://doi.org/10.1128/JCM.01392-06 (2006).
30. Robinson, J. T. *et al.* Integrative Genome Viewer. *Nat. Biotechnol.* **29**(1), 24–26. https://doi.org/10.1038/nbt.1754.Integrative (2011).

Acknowledgements

Authors would like to acknowledge the use of Servicio General de Apoyo a la Investigación-SAI, Universidad de Zaragoza (Servicio de Análisis Microbiológico), and Servicios Científico Técnicos, IACS (Servicio de Secuenciación y Genómica Funcional and Servicio de Biocomputación). We would like to thank the EPIMOLA group for supplying the genotyped bacterial DNA used in this work. This work was supported by the Carlos III Health Institute in the context of a Grant (FIS18/0336) and J.C. was awarded a scholarship by the Government of Aragon/European Social Fund, "Building Europe from Aragon".

Author contributions

S.S. "conceptualization, funding acquisition, writing the manuscript". J.C. "laboratory work, analysis the data, writing the manuscript". A.C. "statistical analysis, biocomputational work". M.J.I., D.I., J.V., L.T. & J.S. "genotyping surveillance, epidemiological support".

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-23343-1.

Correspondence and requests for materials should be addressed to J.C.

Reprints and permissions information is available at www.nature.com/reprints.

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

Table S1. MIRU-VNTR and Spoligotype patterns of the 26 outbreaks studied.

	MIRU-VNTR																							Spoligotype
CLS	154	424	577	580	802	960	1644	1955	2059	2163b	2165	2347	2401	2461	2531	2687	2996	3007	3171	3192	3690	4052	4156	4348	
2	2	2	3	2	3	5	3	2	2	4	3	4	4	2	5	1	5	3	3	3	3	6	3	2	1111111111111111111111111111110100001111111
5	3	3	4	3	2	3	3	1	2	4	4	4	2	2	5	1	5	3	3	2	4	5	2	1	1111000100111111111111111111111100001111111
7	2	2	4	2	2	3	4	2	2	3	2	4	2	2	5	1	5	3	3	3	3	5	2	2	11111111111111111111111000000000000000000111
8	2	1	4	2	1	4	2	3	2	2	2	4	1	2	5	1	6	3	3	3	3	5	2	2	11011111111101111111000011111111100001111111
9	1	3	4	2	5	4	2	3	2	2	2	4	1	2	6	1	5	3	3	3	2	7	3	2	11111111111111111111000000000000000000000000
13	2	2	4	2	2	3	4	2	2	4	2	4	2	2	6	1	5	4	3	3	3	5	2	2	11111100111101111111111111111111100001111111
14	2	2	3	2	4	5	3	3	1	6	3	4	4	2	5	1	5	3	3	3	3	7	3	2	11110111111111111111111110000001000011111111
15	2	2	3	2	3	6	3	2	2	6	3	4	4	2	5	1	5	3	3	3	3	1	3	2	11111111111111111111111111111110100001111111
19	2	2	3	2	4	5	3	3	1	6	3	4	4	2	5	1	5	3	3	3	3	7	3	2	11111111111111111111111111111100001111111
21	2	2	4	2	6	3	3	2	2	2	2	4	2	2	5	1	4	3	3	3	3	5	2	2	11011111111111111111111110000000000000111111
24	2	6	4	2	3	4	3	3	2	4	2	4	2	2	6	1	5	3	1	3	1	3	2	2	11111111000111111111000011111111100001111111
26	2	2	4	2	5	3	3	2	2	3	3	4	2	2	6	1	5	3	4	3	2	5	2	2	11111111111111111110110111111111100001111111
28	2	2	4	2	2	3	4	2	2	4	2	4	2	2	5	1	5	3	2	4	3	5	2	2	11111111111111111111111111111111100001111111
30	2	3	3	2	3	5	3	3	2	3	3	2	4	2	3	1	5	3	3	3	3	7	3	2	00000000000000000000000010000001000011111111
47	2	2	3	2	2	3	4	2	2	1	2	4	2	2	5	1	5	3	3	3	3	5	2	2	11111111111111111111111111111111100001111111
49	2	2	3	2	3	2	3	2	2	5	3	4	4	2	4	1	5	3	3	3	3	6	3	2	11111111111111111110011111110000000000000001
50	2	5	3	2	4	4	3	3	2	4	3	4	4	2	5	1	5	3	3	2	3	7	4	2	00011011111111111011111111111111100001111111
67	2	2	3	2	3	5	3	3	2	5	3	2	4	2	3	1	5	3	3	3	3	7	3	2	11111111111111111111111111111110100001111111
71	1	3	4	2	6	4	3	3	2	2	2	4	1	2	6	1	5	3	3	3	3	4	3	2	11111111111101111111000011111111100001111111
91	2	2	3	2	4	3	3	3	2	5	3	4	4	2	5	1	5	3	3	3	3	5	3	2	11111111000000000000000010000001000011111111
93	2	1	4	2	4	3	4	2	2	5	2	4	2	2	5	1	5	4	3	2	3	5	2	2	10111110111111111111111111111111100001111111
119	2	2	4	3	4	3	2	2	2	5	3	4	2	2	5	1	5	3	3	3	3	5	2	2	11111111111111111111111111111111100001111111
152	2	2	5	2	4	3	3	3	1	2	4	5	2	1	5	1	4	3	3	4	3	7	2	2	11001111111111010111111111111111100001110110
157	2	2	3	2	1	4	3	4	2	6	3	4	4	2	5	1	4	3	3	3	3	5	3	2	11011111000111111111000011111111100001111111
195	2	3	2	2	3	4	2	3	2	1	2	4	2	2	6	1	5	3	1	3	3	7	2	2	11111111111000001110000010111111100001001111
217	1	2	4	2	2	4	3	4	2	2	2	4	1	2	6	1	5	3	3	2	2	7	2	2	11111111111011110000011111111011100001111111

Table S2. SNPs found among the isolates that belong to the same cluster. The isolate with the mutation is the one with the word "SNP". For each mutation point, the reference allele, the SNP, the affected gene, the type of mutation and the potential effect of the mutation are indicated.

CLS_5

	2007 isolate	2016 isolate	Reference	SNP	Gene	Type of mutation	Effect
459226		SNP	G	A	*Rv0383c*	Synonymous	
2120653		SNP	C	T	*Rv1869c*	Non-synonymous	Deleterious

CLS_7

	2004 isolate	2019 isolate	Reference	SNP	Gene	Type of mutation	Effect
307448	SNP		G	C	*cobQ1*	Synonymous	
896966		SNP	C	T	*Rv0805*	Synonymous	
1103129		SNP	T	C	*Rv0987*	Synonymous	
2770751		SNP	T	C	*pepN*	Non-synonymous	Deleterious
3070366		SNP	C	A	*Rv2757c*	Non-synonymous	Deleterious
3771954		SNP	G	C	*Rv3359*	Non-synonymous	Deleterious

CLS_8

	2004 isolate	2012 isolate	Reference	SNP	Gene	Type of mutation	Effect
590249		SNP	C	T	*proC*	Non-synonymous	Neutral
3650241		SNP	T	C	Intergenic	-	
3860087		SNP	G	A	*mrsA/glmM*	Synonymous	

CLS_9

	2004 isolate	2016 isolate	Reference	SNP	Gene	Type of mutation	Effect
2989		SNP	G	A	*dnaN*	Non-synonymous	Deleterious
81911	SNP		T	C	*Rv0073*	Non-synonymous	Deleterious
664875	SNP		A	C	Intergenic	-	
933938	SNP		A	G	*Rv0837c*	Non-synonymous	Neutral
955782		SNP	G	T	*fadA*	Non-synonymous	Neutral
1805118	SNP		C	T	*hisF*	Non-synonymous	Deleterious
2123145		SNP	C	T	*lldD2*	Non-synonymous	Neutral
2338294		SNP	G	A	*Rv2081c*	Non-synonymous	Neutral
2543890	SNP		G	A	*cyp128*	Synonymous	
2583464	SNP		G	C	*Rv2310*	Synonymous	
2817261		SNP	A	G	*accD1*	Non-synonymous	Deleterious
2908252		SNP	G	A	*relA*	Synonymous	
3354218	SNP		G	A	*serA1*	Synonymous	
3959556	SNP		G	C	*ltp3*	Non-synonymous	Deleterious

CLS_13

	2004 isolate	2015 isolate	Reference	SNP	Gene	Type of mutation	Effect
857321		SNP	G	C	*cyp51*	Synonymous	
4200145		SNP	G	A	*Rv3753c*	Non-synonymous	Deleterious
4353270		SNP	G	A	*Rv3876*	Synonymous	

CLS_14

	2005 isolate	2013 isolate	2020 isolate	Reference	SNP	Gene	Type of mutation	Effect
665		SNP		G	C	*dnaA*	Non-synonymous	Deleterious
1495326	SNP			A	G	*glgP*	Non-synonymous	Neutral

2552054			SNP	G	A	*Rv2280*	Synonymous	
2626600	SNP			G	A	Intergenic		
2641831			SNP	A	C	*furB/zur*	Non-synonymous	Deleterious
2976563	SNP			C	T	Intergenic		

CLS_19

	2005 isolate	**2011 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
559260	SNP		C	T	*fadB2*	Synonymous	
596606	SNP		G	C	Intergenic	-	
1120040	SNP		T	C	*Rv1003*	Synonymous	
3051874	SNP		C	T	*Rv2738c*	Non-synonymous	Deleterious
4186244	SNP		T	C	*Rv3735*	Synonymous	

CLS_21

	2007 isolate	**2018 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
346100		SNP	T	A	*Rv0284*	Non-synonymous	Neutral
997170		SNP	C	T	*Rv0894*	Non-synonymous	Neutral

CLS_24

	2005 isolate	**2018 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
1445894		SNP	C	T	Intergenic	-	
2535389		SNP	G	C	*Rv2262c*	Non-synonymous	Neutral
2978833	SNP		C	T	*Rv2656c*	Non-synonymous	Neutral
3139847	SNP		G	T	*ugpB*	Non-synonymous	Neutral
3769492		SNP	G	T	*Rv3354*	Non-synonymous	Neutral
3772423		SNP	C	T	*Rv3359*	Synonymous	

4373812		SNP	C	G	*esxC*	Non-synonymous	Neutral

CLS_28

	2005 isolate	2016 isolate	Reference	SNP	Gene	Type of mutation	Effect
2060855		SNP	C	T	Rv1817	Synonymous	

CLS_30

	2008 isolate	2015 isolate	Reference	SNP	Gene	Type of mutation	Effect
689828		SNP	C	T	*mce2C*	Non-synonymous	Deleterious
1417262		SNP	G	C	*embR*	Non-synonymous	Deleterious
3062615		SNP	G	A	*Rv2749*	Synonymous	

CLS_47

	2004 isolate	2011 isolate	2019 isolate	Reference	SNP	Gene	Type of mutation	Effect
137921		SNP		C	G	Intergenic	-	
579098	SNP		SNP	G	T	*gpmA*	Non-synonymous	Neutral
666421	SNP		SNP	C	G	*Rv0573c*	Synonymous	
719016		SNP		G	T	*Rv0628c*	Synonymous	
728775	SNP		SNP	G	T	*echA3*	Synonymous	
779238		SNP		C	G	*Rv0678*	Synonymous	
973146	SNP		SNP	G	A	*Rv0874c*	Synonymous	
1088556			SNP	C	T	*Rv0976c*	Synonymous	
1124814		SNP		G	A	*Rv1006*	Synonymous	
1333446		SNP		C	A	*Rv1190*	Non-synonymous	Neutral
1589012	SNP		SNP	C	T	*ribC*	Non-synonymous	Deleterious
1682910			SNP	G	A	*mutA*	Non-synonymous	Neutral

1913970			SNP	C	T	*tyrS*	Synonymous	
1926258		SNP		A	G	*Rv1701/xerD*	Synonymous	
1951410	SNP		SNP	C	T	*Rv1725c*	Synonymous	
2117433	SNP			T	C	*Rv1868*	Synonymous	
2202104	SNP		SNP	C	G	Intergenic	-	
2213005	SNP		SNP	G	T	*mce3D*	Non-synonymous	Deleterious
2229741	SNP		SNP	G	A	*Rv1985c*	Synonymous	
2315446		SNP		G	A	*Rv2059*	Non-synonymous	Neutral
2472012		SNP		G	A	*cobT*	Non-synonymous	Deleterious
2475938			SNP	T	C	*ilvE*	Synonymous	
2532616		SNP		G	A	*mscR*	Synonymous	
2602866			SNP	A	G	*narK1*	Non-synonymous	Deleterious
2625924		SNP		T	C	*esxO*	Synonymous	
2627922	SNP		SNP	T	C	*plcC*	Synonymous	
2720062		SNP		G	C	*Rv2423*	Non-synonymous	Deleterious
2940074			SNP	C	T	*pgsA1*	Non-synonymous	Neutral
3075613	SNP		SNP	C	T	*fabG/Rv2766c*	Non-synonymous	Neutral
3125808		SNP		A	G	*Rv2818c*	Non-synonymous	Neutral
3678687			SNP	A	G	*lhr*	Non-synonymous	Deleterious
4332175			SNP	T	C	*gltB*	Non-synonymous	Neutral

CLS_50

	2005 isolate	**2018 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
282798		SNP	C	T	*Rv0236c*	Non-synonymous	Neutral
1072864	SNP		C	T	*Rv0959*	Non-synonymous	Neutral
1392461		SNP	G	A	*kgd*	Non-synonymous	Deleterious
1825730		SNP	A	G	*cydA*	Non-synonymous	Deleterious
1826734		SNP	T	C	*cydA*	Non-synonymous	Deleterious
2212353		SNP	C	T	*mce3C*	Non-synonymous	Deleterious

2580479		SNP	G	T	*Rv2308*	Non-synonymous	Deleterious
2698983		SNP	A	G	*Rv2402*	Non-synonymous	Deleterious
4232392		SNP	G	C	*Rv3786c*	Non-synonymous	Neutral

CLS_67

	2007 isolate	**2013 isolate**	**2016 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
912342	SNP			A	G	*Rv0819*	Non-synonymous	Neutral
1011099			SNP	C	G	*Rv0907*	Non-synonymous	Deleterious
2155168	SNP			C	G	*katG*	Non-synonymous	Deleterious

CLS_71

	2007 isolate	**2020 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
124696		SNP	A	G	*Rv0106*	Non-synonymous	Deleterious
420253		SNP	A	G	*dnaK*	Non-synonymous	Neutral
906355		SNP	A	G	Intergenic	-	
1746600		SNP	A	T	*Rv1544*	Non-synonymous	Deleterious
2449707		SNP	G	A	*fadD15*	Synonymous	
3263698	SNP		A	C	*ppsD*	Non-synonymous	Deleterious
4088617		SNP	G	T	Intergenic	-	

CLS_91

	2005 isolate	**2010 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
79319		SNP	G	A	Intergenic	-	
401678		SNP	C	A	*Rv0336*	Non-synonymous	Neutral
557279	SNP		G	T	Intergenic	-	
723514		SNP	G	A	*recB*	Synonymous	

1272347		SNP	T	A	Intergenic	-	
1743286		SNP	G	C	*Rv1540*	Non-synonymous	Deleterious
3604255		SNP	G	A	*aroA*	Synonymous	
3708572		SNP	G	A	*Rv3322c*	STOP	
3717396	SNP		G	C	*sugI*	Non-synonymous	Deleterious
3945519	SNP		C	T	*fadD18*	Non-synonymous	Deleterious

CLS_93

	2019 isolate	**2008 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
505884	SNP		G	T	*lpqM*	Non-synonymous	Deleterious
1065437	SNP		C	T	*Rv0954*	Non-synonymous	Neutral
1088992	SNP		T	C	*Rv0976c*	Non-synonymous	Neutral
1587885	SNP		C	G	*lprG*	Non-synonymous	Deleterious
1971331	SNP		C	G	*Rv1744c*	Synonymous	
2093856	SNP		G	A	*gnd1*	Non-synonymous	Deleterious
2103577		SNP	T	G	*Rv1855c*	Synonymous	
2392482		SNP	G	A	Intergenic	-	
3017654		SNP	G	C	*ppgK*	Non-synonymous	no STOP codon
3653902	SNP		A	C	*Rv3272*	Non-synonymous	Deleterious

CLS_119

	2012 isolate	**2018 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
7997		SNP	A	C	*gyrA*	Synonymous	
26048	SNP		C	A	*Rv0021c*	Synonymous	
47615	SNP		C	T	*Rv0043c*	Non-synonymous	Neutral
289256		SNP	C	T	*Rv0239*	Synonymous	
311371		SNP	C	G	*Rv0259c*	Synonymous	

1226390-1226391	SNP		GG	AA	*fum*	Non-synonymous	Neutral
1491841		SNP	C	T	*glgB*	Non-synonymous	Deleterious
1691315	SNP		G	A	*Rv1500*	Non-synonymous	Neutral
1705281		SNP	C	T	*Rv1513*	Non-synonymous	Deleterious
1782835	SNP		G	A	*Rv1578c*	Non-synonymous	Deleterious
1980175	SNP		C	A	*Rv1751*	Synonymous	
2015875	SNP		G	A	*malQ*	Synonymous	
2228325		SNP	G	A	*cfp21*	Synonymous	
2446399		SNP	A	C	*Rv2184c*	Non-synonymous	Deleterious
2557572	SNP		G	A	*Rv2285*	Non-synonymous	Deleterious
2816962		SNP	T	C	*accD1*	Non-synonymous	Neutral
3852303	SNP		G	A	*Rv3433c*	Non-synonymous	Neutral
4409046	SNP		T	C	*Rv3921c*	Non-synonymous	Neutral

CLS_152

	2014 isolate	**2020 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
1456754		SNP	C	T	*hemK*	Synonymous	
1623201		SNP	C	T	Intergenic	-	
1822944		SNP	C	T	*cydD*	Synonymous	
2225217		SNP	T	C	Intergenic	-	

CLS_157

	2011 isolate	**2019 isolate**	**Reference**	**SNP**	**Gene**	**Type of mutation**	**Effect**
48510		SNP	G	T	*Rv0044c*	Non-synonymous	Deleterious
73085	SNP		G	A	*icd2*	Non-synonymous	Deleterious
129311	SNP		G	T	*ctpI*	Non-synonymous	Deleterious
157492		SNP	T	C	*fbpC*	Non-synonymous	Neutral

394412		SNP	C	T	*Rv0328*	Non-synonymous	Deleterious
472186	SNP		G	C	*ndhA*	Non-synonymous	Neutral
486141		SNP	T	C	*pks6*	Synonymous	
817100		SNP	C	T	*sppA*	Non-synonymous	Deleterious
818103	SNP		G	A	*Rv0725c*	Synonymous	
939395	SNP		G	A	*Rv0842*	Synonymous	
1168629	SNP		G	A	Intergenic	-	
1232323	SNP		A	G	*Rv1105*	Non-synonymous	Neutral
1341120	SNP		A	G	*esxL*	Non-synonymous	Neutral
1473933		SNP	C	T	*rrl*	-	
1814842		SNP	C	T	Intergenic	-	
1905459		SNP	C	A	*moeX*	Synonymous	
2306540		SNP	G	C	*pks12*	Non-synonymous	Deleterious
2426596		SNP	C	T	*pbpB*	Synonymous	
2443110		SNP	A	G	*Rv2180c*	Synonymous	
2832457	SNP		G	C	*Rv2515c*	Synonymous	
2929362	SNP		C	T	*speE*	Synonymous	
3031043	SNP		G	A	*Rv2719c*	Synonymous	
3117255		SNP	C	A	*IS1604/Rv2812*	-	
3164652	SNP		G	C	*Rv2854*	Synonymous	
3184957	SNP		C	T	*dipZ*	Synonymous	
3197896		SNP	G	C	*amiC*	Non-synonymous	Deleterious
3571213		SNP	A	G	Intergenic	-	
3609143	SNP		C	T	*pvdS*	Synonymous	
3644189		SNP	G	A	*Rv3263*	Non-synonymous	Deleterious
3983844		SNP	C	A	*fadE28*	Non-synonymous	Neutral
3989266	SNP		G	A	*echA20*	Synonymous	
4060588		SNP	T	C	*esxW*	Non-synonymous	Neutral
4366853		SNP	C	T	Intergenic	-	
4385534		SNP	C	A	*Rv3900c*	Non-synonymous	Neutral

CLS_195

	2013 isolate	2017 isolate	Reference	SNP	Gene	Type of mutation	Effect
761155		SNP	C	T	*rpoB*	Non-synonymous	Deleterious

CLS_217

	2015 isolate	2017 isolate	Reference	SNP	Gene	Type of mutation	Effect
1924619	SNP		A	G	*pyrG*	Non-synonymous	Deleterious

Table S3. IS*6110* data of each cluster. For each location, the number of reads found, the direct repeats, the gene involved and the direction of the IS*6110* are indicated. Extra IS*6110* in one of the isolates is shadowed in blue.

CLS_2							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
483296	14	483298	17	agg	*Rv0403c*	Reverse	
3121879	17	3120523	11	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse	
1075948	5	1075950	14	acc	*Rv0963c*	Reverse	
2610863	15	2610861	22	gcc	*Rv2336*	Forward	
1715972	8	1715974	15	acc	*mmpL12*	Reverse	
3076505 (*M. bovis*)	7	3076507 (*M. bovis*)	2	gtc	*Rv2813:Rv2814c*	Reverse	
3668575 (*M. bovis*)	13	3668756 (*M. bovis*)	9	-	*MT3429*	Forward	
1986626	3	1986622	10	tgttc	*Rv1754c*	Forward	
3551130	19	3551132	23	aag	*Rv3183:Rv3184*	Reverse	Extra IS*6110* in 2007 isolate

CLS_5						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
2637060	10	2634130	23	ccgt	*Rv2355:PPE40/PPE38:PPE39*	Reverse
80480	13	80482	11	gga	*Rv0071:Rv0072*	Reverse
3121879	11	3120523	14	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
2166573	8	2166575	8	ttg	*PPE34*	Reverse
1889066	17	1889068	21	ccg	*pks9*	Reverse
3668758 (*M. bovis*)	11	3668960 (*M. bovis*)	16	-	*MT3429/IS1547*	Forward
3125708	11	3125706	19	cgc	*Rv2818c*	Forward
1998849	13	1987457	32	-	*Rv1765c:Rv1765A/plcD*	Reverse

CLS_7						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
3606308	98	3606310	127	acg	*desA3*	Reverse
850539	171	850536	106	tgtc	*Rv0755A:thrV*	Forward
2627494	69	2627492	106	tca	*plcC*	Forward
2559568	135	2559570	108	gat	*Rv2286c*	Reverse
3129520	116	3129522	131	caa	*Rv2823c*	Reverse
1979901 (*M. bovis*)	84	1998483/1988923 (*M. bovis*)	107	-	*cut1/Rv1765c*	Forward
2633977	61	2633979	54	gcg	*PEE38*	Reverse
2604207 (*M. bovis*)	133	2604210 (*M. bovis*)	87	gaaa	*PPE71*	Reverse
1895651	90	1895654	88	ccta	*Rv1668c:Rv1669*	Reverse
889020	84	890376	131	gagg	*Rv0795-Rv0796*	Forward
3668725 (*M. bovis*)	42	3668723 (*M. bovis*)	53	gcc	*MT3429*	Forward
3121879	89	3119660	72	-	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
1627		1625			*dnaA:dnaN*	Forward

CLS_8						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
1481531	19	1481529	19	ggc	*Rv1319c*	Forward
3480373	23	3480371	21	cag	*Rv3113*	Forward
2047033	14	2047036	11	tcag	*Rv1804c*	Reverse
2165990	26	2165988	16	tgg	*PPE34*	Forward
3665157 (*M. bovis*)	22	3665159 (*M. bovis*)	12	caa	*MT3426:MT3427*	Reverse
2180843	21	2180845	19	ctt	*Rv1928c*	Reverse
1989080/1979921 (*M. bovis*)	13	1986625	31	-	*cut1/Rv1754c*	Reverse
932202	24	932204	7	aac	*lpqQ:Rv0836c*	Reverse
3078617 (*M. bovis*)	13	3120523	10	-	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
3493843	29	3493841	15	aac	*Rv3128c*	Forward
401268/607627	26	401270/607629	28	atc	*Rv0336/Rv0515*	Reverse

CLS_9						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
3480373	12	3480371	25	cag	*Rv3113*	Forward
3665157 (*M. bovis*)	18	3665159 (*M. bovis*)	10	caa	*MT3426:MT3427*	Reverse
1987457	13	1986625	31	-	*plcD/Rv1754c*	Reverse
932202	19	932204	13	aac	*lpqQ:Rv0836c*	Reverse
4077859	21	4077861	12	atc	*Rv3638:Rv3639c*	Reverse
483310	15	483308	27	tga	*Rv0403c*	Forward
2047368	19	2047370	24	att	*Rv1804c:Rv1805c*	Reverse
3083207	17	3083204	24	cagc	*Rv2775*	Forward
1995894	11	1995892	11	tgc	*Rv1762c:Rv1763*	Forward
1989080/1979921 (*M. bovis*)	15	1979923 (*M. bovis*)	18	cgc	*cut1*	Reverse
2807873	9	2807871	32	aga	*Rv2492*	Forward
3078617 (*M. bovis*)	11	3115687	36	-	*Rv2815c:Rv2816c/Rv2809*	Reverse
2555929	4	2555931	10	tca	*Rv2282c:Rv2283*	Reverse

2366892	14	2366896	8	tagtg	*Rv2106:PE22*	Reverse
3743341	34	3743344	42	gcgg	*Rv3346c*	Reverse
3493190	14	3493192	20	agg	*Rv3128c*	Reverse

CLS_13							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
1895651	7	1895654	5	ccta	*Rv1668c:Rv1669*	Reverse	
3668725 (*M. bovis*)	7	3668723 (*M. bovis*)	15	gcc	*MT3429*	Forward	
2964323	6	2964321	15	atc	*dedA:Rv2638*	Forward	
3547613	5	3547610	10	ttga	*Rv3178:Rv3179*	Forward	
3121879	9	3120523	8	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse	
850539	5	850536	6	tgtc	*Rv0755A:thrV*	Forward	
1979901 (*M. bovis*)	3	1979899 (*M. bovis*)	16	gag	*cut1*	Forward	
1996100/1986939 (*M. bovis*)	8	1986937 (*M. bovis*)	8	acc	*Rv1762c:Rv1763*	Forward	
2010649	17	2010647	14	cct	*Rv1776c:cyp144*	Forward	
1891152	8	1891150	6	atg	*pks9:pks11*	Forward	
888926	6	890376	22	-	*Rv0796:IS1547*	Forward	
2637903	6	2627510	32	-	*PPE40/plcC*	Reverse	
1982028	3	1982030	19	ctc	*PPE24*	Reverse	
2041739	1	2041741	4	cgc	*PPE28:PPE29*	Reverse	Extra IS*6110* in 2015 isolate
851324	1	851322	1	acc	*Rv0756c*	Forward	Extra IS*6110* in 2015 isolate

CLS_14						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
3042322	7	3042320	8	cag	*Rv2729c*	Forward

483296	5	483298	6	agg	*Rv0403c*	Reverse
1469505/4252993	7	1469503/4252995	1	cac	*IS1557*	Forward
1998698	4	1995808	4	-	*Rv1765c:Rv1765A/Rv1762c*	Reverse
2633701	4	2633699	6	ccg	*PPE38*	Forward
3486461	1	3486463	4	ggg	*Rv3120:cyp141*	Reverse
2610863	3	2610861	6	gcc	*Rv2336*	Forward
3121879	1	3076507 (*M. bovis*)	1	-	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
1986626	5	1986622	10	tgttc	*Rv1754c*	Forward
1075948	1	1075950	5	acc	*Rv0963c*	Reverse
2163473/2163404	5	2163402	9	cag	*PPE34*	Reverse
3668575 (*M. bovis*)	3	3668756 (*M. bovis*)	3	-	*MT3429*	Forward
2634049	2	2634051	8	gat	*PPE38*	Reverse

CLS_15						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
2610863	9	2610861	7	gcc	*Rv2336*	Forward
1986626	9	1986622	16	tgttc	*Rv1754c*	Forward
1075948	6	1075950	9	acc	*Rv0963c*	Reverse
3048327	3	3048324	11	taag	*Rv2735c*	Forward
3121879	8	3120523	5	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
3668575 (*M. bovis*)	3	3668829 (*M. bovis*)	6	-	*MT3429*	Forward
3120025/3076505 (*M. bovis*)	5	3076507 (*M. bovis*)	7	gtc	*Rv2813:Rv2814c*	Reverse
1715972	7	1715974	10	acc	*mmpL12*	Reverse
483296	1	483298	14	agg	*Rv0403c*	Reverse

CLS_19						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
482937	9	482935	20	gtc	*mmpL1*	Forward
1987129	15	1987131	20	aaa	*plcD*	Reverse

2785970	13	2784614	10	ctg	*Rv2479c-Rv2480c*	Reverse
889020	16	890375	35	gagg	*Rv0795-Rv0796*	Forward
2972108	11	2973464	11	tcg	*Rv2648-Rv2649*	Forward
1979901 (*M. bovis*)	8	1989058	22	gag	*cut1/Rv1756c-Rv1757c*	Reverse
1996100	19	1997456	17	-	*Rv1763-Rv1764*	Forward
1543307	7	1541951	10	ggc	*Rv1369c-Rv1370c*	Reverse
		2604210 (*M. bovis*)	22	-	*PPE71*	
3890778	7	3892134	8	gac	*Rv3474-Rv3475*	Forward
2365413	11	2366769	9	cga	*Rv2105-Rv2106*	Forward
3121879	7	3120523	11	ccc	*Rv2815c:Rv2816c/Rv2813:Rv1814c*	Reverse
		2636932/2605006 (*M. bovis*)	12	-	*Rv2355:PPE40/PPE71*	
3710381	7				*Rv3325-Rv3326*	
	320				*IS within a IS*	

CLS_21							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
3668619 (*M. bovis*)	14	3668723 (*M. bovis*)	14	-	*MT3429*	Forward	
1895651	8	1895654	13	ccta	*Rv1668c:Rv1669*	Reverse	
2629052	9	2629054	20	gtt	*plcB*	Reverse	
3121879	15	3120523	9	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse	
2604207 (*M. bovis*)	19	2604210 (*M. bovis*)	43	gaaa	*PPE71*	Reverse	
1979901 (*M. bovis*)	11	1979899 (*M. bovis*)	26	gag	*cut1*	Forward	
1996100/1986939 (*M. bovis*)	12	1986937 (*M. bovis*)	16	acc	*Rv1762c:Rv1763*	Forward	
889020	15	890375	17	gagg	*Rv0795-Rv0796*	Forward	
1998664	14	1998666	12	ttg	*Rv1765c:Rv1765A*	Reverse	
1158764	14	1158762	23	gca	*trcR:Rv1034c*	Reverse	

3122551	4	3122549	1	acg	*Rv2815c:Rv2816c*	Forward	Extra IS*6110* en 2018 isolate

CLS_24							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
3480373	9	3480371	15	cag	*Rv3113*	Forward	
3668572 (*M. bovis*)	13	3668756 (*M. bovis*)	10	-	*MT3429*	Forward	
3302967	9	3302969	6	gtg	*fadD29:Rv2951c*	Reverse	
2163303	13	2163305	18	-	*PPE34*	Reverse	
1986904	9	1986902	12	ggg	*plcD*	Forward	
3547511	9	3547509	8	gat	*Rv3178:Rv3179*	Forward	
1986623	8	1986625	11	aac	*Rv1754c*	Reverse	
2044835	10	2044832	10	agac	*PPE30:PE_PGRS32*	Forward	
2368139	10	2368142	22	ccac	*PPE36*	Reverse	
1997628/2262185	17	1997630/ 2262187	24	cgt	*Rv1765c/Rv2015c*	Reverse	
932202	11	932204	18	aac	*lpqQ:Rv0836c*	Reverse	
1989080/1979921 (*M. bovis*)	6	1979923 (*M. bovis*)	7	cgc	*cut1*	Reverse	
2010922	11	2010924	26	gag	*cyp144*	Reverse	
2633843	7	2633841	15	ctc	*PPE38*	Forward	
3078618 (*M. bovis*)	3	3120523	12	-	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse	
3664914 (*M. bovis*)	14	3664912 (*M. bovis*)	8	ggg	*MT3426*	Forward	Extra IS*6110* in 2005 isolate

CLS_26						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
1998793	9	1998795	8	gtg	*Rv1765c:Rv1765A*	Reverse
3121879	7	3120523	13	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse

1987085	19	1986937 (*M. bovis*)	8	-	*plcD/Rv1762c:Rv1763*	Forward
889020	13	890376	25	gagg	*Rv0795-Rv0796*	Forward
3668725 (*M. bovis*)	11	3668723 (*M. bovis*)	11	gcc	*MT3429*	Forward
2604207 (*M. bovis*)	7	2604210 (*M. bovis*)	13	gaaa	*PPE71*	Reverse

CLS_28						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
3121879	15	3120523	13	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
1895651	10	1895654	11	ccta	*Rv1668c:Rv1669*	Reverse
1998792	16	1998794	8	tga	*Rv1765c:Rv1765A*	Reverse
1979899 (*M. bovis*)	17	1986937 (*M. bovis*)	16	-	*cut1/Rv1762c:Rv1763*	Forward
3550870	24	3550868	17	-	*Rv3183*	Forward
3363263	25	3363260	9	cgt	*ilvB1:cfp6*	Forward
851412	20	851409	17	gagc	*Rv0756c*	Forward
3115129	8	3115131	26	tcg	*Rv2808*	Reverse
889020	11	890376	23	gagg	*Rv0795-Rv0796*	Forward
2604207 (*M. bovis*)	10	2604210	24	gaaa	*PPE71*	Reverse
3668725 (*M. bovis*)	13	3668723 (*M. bovis*)	4	gcc	*MT3429*	Forward
850539	11	850536	10	tgtc	*Rv0755A:thrV*	Forward

CLS_30						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
47041	23	47039	19	ccg	*Rv0042c*	Forward
1986626	10	1986622	39	tgttc	*Rv1754c*	Forward
3668575 (*M. bovis*)	10	3668756 (*M. bovis*)	20	-	*MT3429*	Forward
1075948	14	1075950	17	acc	*Rv0963c*	Reverse
2610863	19	2610861	31	gcc	*Rv2336*	Forward
1998793	10	1998795	13	gtg	*Rv1765c:Rv1765A*	Reverse
891081/3712442	18	891083/3712444	25	cag	*IS1547*	Reverse

3126538	7	3076507 (*M. bovis*)	11	-	*Rv2819c/Rv2815c:Rv2816c*	Reverse
3709336	13	3709338	15	aag	*moaX*	Reverse
3491909	18	3491907	22	cca	*Rv3126c*	Forward
2634049	10	2634051	13	gat	*PPE38*	Reverse
483296	10	483298	26	agg	*Rv0403c*	Reverse

CLS_47						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
2633979	22	2633977	24	cgc	*PPE38*	Forward
888787	13	888789	16	cgg	*Rv0794c:Rv0795*	Reverse
2604207 (*M. bovis*)	22	2604210 (*M. bovis*)	23	gaaa	*PPE71*	Reverse
483537	20	483535	30	gtg	*Rv0403c*	Forward
3494395	15	3494393	8	tac	*Rv3128c:Rv3129*	Forward
1986724	14	1986726	8	gat	*Rv1754c:plcD*	Reverse
3121879	14	3120523	15	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
2559703	13	2559701	15	gta	*Rv2282c:yjcE*	Forward
850539	24	850536	11	tgtc	*Rv0755A:thrV*	Forward
889020	11	890375	23	gagg	*Rv0795-Rv0796*	Forward
3668725 (*M. bovis*)	15	3668723 (*M. bovis*)	9	gcc	*MT3429*	Forward
1979901 (*M. bovis*)	15	1986937 (*M. bovis*)	8	-	*cut1/Rv1762c:Rv1763*	Forward
1895651	10	1895654	12	ccta	*Rv1668c:Rv1669*	Reverse

CLS_49							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
2610863	10	2610861	14	gcc	*Rv2336*	Forward	
1075948	14	1075950	24	acc	*Rv0963c*	Reverse	
1715972	10	1715974	15	acc	*mmpL12*	Reverse	
3550927	15	3550924	15	ggag	*Rv3183*	Forward	

483296	15	483298	22	agg	*Rv0403c*	Reverse	
1986626	9	1986622	11	tgttc	*Rv1754c*	Forward	
3121879	6	3120523	15	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse	
1987254	12	1987257	15	ccgg	*plcD*	Reverse	
3668575 (*M. bovis*)	6	3668756 (*M. bovis*)	20	-	*MT3429*	Forward	
1998564/2263121	14	1998562/2263119	21	cca	*Rv1765c/Rv2015c*	Forward	Extra IS*6110* in 2012 isolate
3491502	16	3491500	15	agc	*PPE49*	Forward	Extra IS*6110* in 2015 isolate

CLS_50							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
3377330	105	3377327	134	gcca	*PPE46*	Forward	
1989058/1979899 (*M. bovis*)	96	1979901 (*M. bovis*)	107	ctc	*cut1*	Reverse	
483296	99	483298	127	agg	*Rv0403c*	Reverse	
3121879	64	3120523	102	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse	
3123106	77	3123104	131	ttc	*Rv2815c:Rv2816c*	Forward	
3665330 (*M. bovis*)	110	3665328 (*M. bovis*)	139	ttgt	*MT3427*	Forward	Extra IS*6110* in 2005 isolate

CLS_67						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
3121879	9	3120523	14	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
1075948	13	1075950	21	acc	*Rv0963c*	Reverse
483296	16	483298	23	agg	*Rv0403c*	Reverse
3126529	16	3126532	18	cgtc	*Rv2819c*	Reverse
891081/3712442	18	891083/3712444	18	cag	*IS1547*	Reverse
3120025/3076505 (*M. bovis*)	10	3076507 (*M. bovis*)	11	gtc	*Rv2813:Rv2814c*	Reverse
2610863	9	2610861	36	gcc	*Rv2336*	Forward

1998793	15	1998795	10	gtg	*Rv1765c:Rv1765A*	Reverse
2222122	10	2222120	21	gca	*Rv1979c*	Forward
2634049	4	2634051	8	gat	*PPE38*	Reverse
3668575 (*M. bovis*)	7	3668756 (*M. bovis*)	13	-	*MT3429*	Forward
3709336	16	3709338	5	aag	*moaX*	Reverse
1982068	16	1982070	1	cct	*PPE24*	Reverse
3491909	17	3491907	31	cca	*Rv3126c*	Forward
1986626	9	1986622	37	tgttc	*Rv1754c*	Forward

CLS_71							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
3122594	12	3122592	14	tcc	*Rv2815c:Rv2816c*	Forward	
3078617 (*M. bovis*)	10	3120523	10	-	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse	
4077859	15	4077861	20	atc	*Rv3638:Rv3639c*	Reverse	
2366894	11	2366896	20	tag	*Rv2106:PE22*	Reverse	
2540720	13	2540718	30	ggc	*cyp124*	Forward	
2555929	9	2555931	13	tca	*Rv2282c:Rv2283*	Reverse	
932202	11	932204	28	aac	*lpqQ:Rv0836c*	Reverse	
3665157 (*M. bovis*)	20	3665159 (*M. bovis*)	8	caa	*MT3426:MT3427*	Reverse	
2047368	13	2047370	18	att	*Rv1804c:Rv1805c*	Reverse	
2057398	4	2057396	18	cgc	*erg3*	Forward	
1987457	18	1986625	28	-	*plcD/Rv1754c*	Reverse	
3119934	21	3119936	14	gag	*Rv2813:Rv2814c*	Reverse	
2174919	21	2174917	15	acc	*Rv1922*	Forward	
3551137	10	3551139	15	cat	*Rv3183:Rv3184*	Reverse	
1989080/1979921 (*M. bovis*)	10	1979923 (*M. bovis*)	11	cgc	*cut1*	Reverse	
2614572	22	2614570	14	gcc	*moeW:mmpL9*	Forward	

2807873	19	2807871	16	aga	*Rv2492*	Forward	
3494461	16	3494463	20	aca	*Rv3128c:Rv3129*	Reverse	
3480373	9	3480371	31	cag	*Rv3113*	Forward	
3083207	13	3083204	11	cagc	*Rv2775*	Forward	
2163459	10	2163461	32	tta	*PPE34*	Reverse	
1544804	21	1544806	26	aaa	*Rv1371*	Reverse	Extra IS*6110* in 2020 isolate
901302	13	901300	26	gac	*phoT*	Forward	Extra IS*6110* in 2020 isolate

CLS_91							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
3076505 (*M. bovis*)	18	3076507 (*M. bovis*)	18	gtc	*Rv2813:Rv2814c*	Reverse	
1075948	10	1075950	17	acc	*Rv0963c*	Reverse	
1986626	18	1986622	16	tgttc	*Rv1754c*	Forward	
2579817	14	2579819	31	tgc	*Rv2307B*	Reverse	
3486602	22	3486604	35	tac	*cyp141*	Reverse	
483296	7	483298	37	agg	*Rv0403c*	Reverse	
1997928/2262485	14	1997926/2262483	26	tca	*Rv1765c/Rv2015c*	Forward	
2610863	15	2610861	20	gcc	*Rv2336*	Forward	
2634049	9	2634051	27	gat	*PPE38*	Reverse	
3548385	25	3548387	25	cgg	*Rv3179*	Reverse	
2410484 (*M. bovis*)	15	2410486 (*M. bovis*)/2430094	13	gtc	*Rv2166c:Rv2167c*	Reverse	
3668575 (*M. bovis*)	11	3668756 (*M. bovis*)	11	-	*MT3429*	Forward	
2634523	9	2634522	39	cc	*PPE39*	Forward	
2630784	1	2630787	10	ggcg	*plcA*	Reverse	Extra IS*6110* in 2005 isolate

CLS_93							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
1996100/1986939 (*M. bovis*)	12	1986937 (*M. bovis*)	10	acc	*Rv1762c:Rv1763*	Forward	
850180	11	850182	18	aag	*PPE12:Rv0755A*	Reverse	
3709800	16	3709803	23	aaaa	*moaC3*	Reverse	
3668725 (*M. bovis*)	13	3668723 (*M. bovis*)	15	gcc	*MT3429*	Forward	
2199550	11	2199552	16	agc	*Rv1949c*	Reverse	
2604207 (*M. bovis*)	20	2633979	21	-	*PPE71/PPE38*	Reverse	
3494099	7	3494097	29	cac	*Rv3128c*	Forward	
3121879	15	3120523	18	-	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse	
3167895	18	3167897	12	atc	*nicT:Rv2857c*	Reverse	
3493592	8	3493595	12	ccat	*Rv3128c*	Reverse	
3122959	11	3122956	12	tttc	*Rv2815c:Rv2816c*	Forward	
1895651	23	1895654	17	ccta	*Rv1668c:Rv1669*	Reverse	
889020	22	890375	27	gagg	*Rv0795-Rv0796*	Forward	
1973344 (*M. bovis*)	8	1989058	22	gag	*plcD/cut1*	Forward	
3545711	12	3545713	18	ggt	*Rv3177*	Reverse	Extra IS*6110* in 2019 isolate

CLS_119						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
889020	5	890375	8	gagg	*Rv0795-Rv0796*	Forward
3130939	6	3130941	9	aca	*Rv2823c*	Reverse
986512	6	986515	7	gtag	*Rv0887c:Rv0888*	Reverse
3668725 (*M. bovis*)	3	3668723 (*M. bovis*)	4	gcc	*MT3429*	Forward
1979901 (*M. bovis*)	6	1987418 (*M. bovis*)	4	-	*cut1/Rv1762c:Rv1763*	Forward
1987006	2	1987008	11	ccg	*plcD*	Reverse

3665044 (*M. bovis*)	4	3665046 (*M. bovis*)	3	gag	*MT3426*	Reverse
2604207 (*M. bovis*)	6	2604210 (*M. bovis*)	2	gaaa	*PPE71*	Reverse
3121879	2	3120523	2	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse
1895651	1	1895654	4	ccta	*Rv1668c:Rv1669*	Reverse
2366861	1	2366859	2	ctt	*Rv2106:PE22*	Forward
3114400	1	3114403	8	tgat	*Rv2807*	Reverse

CLS_152							
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**	
3367040	13	3367038	16	ggc	*Rv3008*	Forward	
1523107	13	1523105	16	ctg	*moeY*	Forward	
2367677	8	2367674	18	cgag	*PE22:PPE36*	Forward	
1533689	13	1533687	6	acg	*PPE19:Rv1362c*	Forward	
3710381	8	890375/3711737	19	-	*Rv3324A/Rv3326:Rv3327*	Forward	
2038789	3	2038787	7	caa	*Rv1798:lppT*	Forward	
3378894	30	3378897	13	gcag	*esxR*	Reverse	
3773280	6	3773278	24	ttt	*Rv3361c*	Forward	
2245301	12	2245299	24	aag	*Rv2000*	Forward	
	73		17	agc	*IS1081*		
2163389	14	2163392	33	ttaat	*PPE34*	Reverse	
2634069	7	2639361	6	-	*PPE38/PPE40*	Forward	
888926	12	888947	12	-	*Rv0794c:Rv0795*	Forward	
4172770	10	4172768	12	cta	*Rv3726:Rv3727*	Forward	
2266241	10	2265111	7	-	*Rv2019/Rv2017:Rv2018*	Reverse	
1902356	8	1902354	10	acg	*dsbF:Rv1678*	Forward	
2039176	11	2039174	21	aag	*lppT*	Forward	
1987511	8	1989058	28	-	*plcD/cut1*	Forward	
3121879	4	3120523	14	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*	Reverse	
1957586	7	1957583	11	gaac	*Rv1730c:gabD2*	Forward	Extra IS*6110* in 2020 isolate

CLS_157						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
444616	12	444614	14	agt	*Rv0366c*	Forward
2610863	12	2610861	26	gcc	*Rv2336*	Forward
1987085	7	1987083	20	ctg	*plcD*	Forward
1900789	15	1900791	20	ttg	*Rv1675c*	Reverse
483296	6	483298	23	agg	*Rv0403c*	Reverse
3709624/3664146 (*M. bovis*)	11	3668756 (*M. bovis*)	9	-	*moaX/MT3429*	Forward
2541732	14	2541730	20	cat	*Rv2267c*	Forward
3555250	6	3555247	8	gcgc	*Rv3189*	Forward
1986626	9	1986622	22	tgttc	*Rv1754c*	Forward
2633682	10	2627272	12	-	*PPE38/plcC*	Reverse
1075948	5	1075950	18	acc	*Rv0963c*	Reverse

CLS_195						
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**	**Direction**
3377000/3379437	34	3377002/3379439	34	gcc	*PPE46/PPE47*	Reverse
3551173	8	3551171	14	aaa	*Rv3183:Rv3184*	Forward
1481531	12	1481529	28	ggc	*Rv1319c*	Forward
932186	17				*lpqQ:Rv0836c*	
2633843	6	2633841	23	ctc	*PPE38*	Forward
2263267	15	2262187	46	-	*Rv2015c:Rv2016/Rv2015c*	Reverse
3078617 (*M. bovis*)	16	3120523/3077033 (*M. bovis*)	15	-	*Rv2813:Rv2814c*	Reverse
890429/3711790	60	890426/3711787	41	cct	*IS1547*	Forward
1987300 (*M. bovis*)/1987302 (*M. bovis*)	27				*Rv1762c:Rv1763*	
2010922	11	2010924	29	gag	*cyp144*	Reverse
3480373	13	3480371	21	cag	*Rv3113*	Forward
2637033	6	2637031	9	gac	*Rv2355:PPE40*	Forward
2608664	9	2608662	19	gcg	*Rv2333c:cysK1*	Forward

3711789	28	3665159 (*M. bovis*)	13	-	*IS1547/MT3426:MT3427*	Reverse
1987302 (*M. bovis*)	13				*Rv1762c:Rv1763*	
		1992737	10	-	*wag22:Rv1760*	
		1986625	25	-	*Rv1754c*	

CLS_217					
pre-IS point	**Number of reads**	**post-IS point**	**Number of reads**	**Direct repeat**	**Gene**
3480373	60	3480371	30	cag	*Rv3113*
2555929	27	2555931	24	tca	*Rv2282c:Rv2283*
3366750	31	3366747	20	tggc	*Rv3008*
4077859	36	4077861	31	atc	*Rv3638:Rv3639c*
1889080	31	1986625	23	-	*cut1/Rv1754c*
610518	20	610516	16	atg	*Rv0518*
3124653	25	3124651	26	ctt	*Rv2817c*
3665159 (*M. bovis*)	30	3668981 (*M. bovis*)/3711737	18	-	*MT3426:MT3427/IS1547*
1519048	33	1519046	25	cct	*Rv1352*
932202	17	932204	30	aac	*lpqQ:Rv0836c*
2430094/2410486 (*M. bovis*)	23	2410484 (*M. bovis*)	20	gac	*Rv2166c:Rv2167c*
1998795	25	1998793	18	cac	*Rv1765c:Rv1765A*
2166090	33	2166092	21	acc	*PPE34*
3121879	28	3120523	25	ccc	*Rv2815c:Rv2816c/Rv2813:Rv2814c*

Table S4. SNPs in genes considered as virulence factors for the outbreaks studied. The first Table includes clusters of L4.1, L4.3 and L4.8, the most frequent. The second Table includes the rest of the clusters, caused by strains from other lineages.

	Point	Reference	Change	Gene	Mutation effect	Description	
CLS_14	234742	C	T	*Rv0198c*	Deleterious	Metalloproteases: hydrolyzes peptides and/or proteins	**L4.1.2.1**
	242026	C	T	*Rv0204c*	Neutral	Proteins of Unknown Function	
	689990	T	C	*mce2C*	Deleterious	Cell wall proteins: involved in host cell invasion	
	716428	C	A	*Rv0623*	Neutral	Antitoxin VapB30	
	2295649	T	G	*pks12*	Neutral	Synthesis of complex lipids: involved in biosynthesis of mannosyl-beta-1-phosphomycoketide (MPM)	
	3916251	C	A	*mce4C*	Neutral	Cell wall proteins: involved in host cell invasion	
	4343530	C	G	*Rv3868*	Neutral	Secretion system:ESX conserved component EccA1. ESX-1 type VII secretion system protein	
	4345780	C	G	*Rv3869*	Deleterious	Secretion system:ESX conserved component EccB1. ESX-1 type VII secretion system protein	
CLS_30	2188210 -218816	several bp	G	*Rv1936*	Reading frame alteration	Oxidative and nitrosative stresses: monooxygenase	
	3599387	C	G	*sigH*	Deleterious	Sigma factors:plays a role in the oxidative-stress response	
CLS_2	3834950	A	G	*whiB3*	Neutral	Other transcriptional regulators: involved in transcriptional mechanism (growth phase-dependent)	
	4095126	G	T	*Rv3655c*	Neutral	Inhibition of apoptosis:participates in the suppression of macrophage apoptosis by blocking the extrinsic pathway	
	4341789	G	A	*Rv3865*	Neutral	Secretion system:ESX-1 secretion-associated protein EspF	
CLS_67	694889	G	A	*Rv0595c*	STOP	Toxin VapC4	
	2629188	A	C	*plcB*	Neutral	Others genes related in lipid synthesis: implicated in the pathogenesis at the level of intracellular survival	
CLS_91	2155168	C	G	*katG*	Deleterious	Oxidative and nitrosative stresses: play a role in the intracellular survival of mycobacteria within macrophages	
CLS_157	1878769	C	T	*pks7*	Deletérea	Synthesis of complex lipids: involved in some intermediate steps for the synthesis of a polyketide molecule	
	2851709	G	A	*Rv2527*	Neutral	Toxin VapC17	
	3018330	G	A	*sigA*	Neutral	Sigma Factors: involved in the housekeeping regulons	
	3291611	G	GT	*pks1*	Reading frame alteration	Synthesis of complex lipids: polyketide synthase possibly involved in lipid synthesis	
CLS_15	1002966	G	A	*ompA*	Neutral	Cell wall proteins: behaved as a porin of low specific activity. Structural protein that may protect the integrity of the bacterium	
	2808542	G	A	*Rv2494*	Neutral	Toxin VapC38	
	3453733	C	T	*Rv3087*	Deletérea	Mycolic acid synthesis: involved in synthesis of triacylglycerol	
	3834950	A	G	*whiB3*	Neutral	Other transcriptional regulators: involved in transcriptional mechanism (growth phase-dependent)	
	4355769	C	G	*Rv3877*	Neutral	Secretion system: ESX conserved component EccD1. ESX-1 type VII secretion system protein	
CLS_24	236686	C	T	*Rv0199*	Deleterious	Others proteins with unknown function	**L4.3**

	1879240	C	G	*pks7*	Deleterious	Synthesis of complex lipids: involved in some intermediate steps for the synthesis of a polyketide molecule	
CLS_195	1377657	C	G	*lpqY*	Neutral	Lipoproteins:involved in active transport of sugar across the membrane	
	2183609	T	G	*tpx*	Deleterious	Oxidative and nitrosative stresses: has antioxidant activity and could remove peroxides	
	2279951	A	G	*acg*	Neutral	Other Virulence Factors: may have a role for bacteria within the host environment	
	3453381	G	A	*Rv3087*	Neutral	Mycolic acid synthesis: involved in synthesis of triacylglycerol	
CLS_8	199672	A	G	*mce1A*	Deleterious	Cell wall proteins: involved in host cell invasion	
	213483	C	T	*sigG*	Deleterious	Sigma Factors: promotes attachment of the RNA polymerase to specific initiation sites and then is released	
	841377-841378	GG	AA	*Rv0749*	Neutral	Toxin VapC31	
	1364434	C	T	*sigE*	Neutral	Sigma Factors: seems to regulate the heat-shock response	
	1765349	G	A	*Rv1561*	Neutral	Toxin VapC11	
	2183850	T	C	*tpx*	Deleterious	Oxidative and nitrosative stresses: has antioxidant activity and could remove peroxides	
	2210740	A	C	*mce3B*	Deleterious	Cell wall proteins: involved in host cell invasion	
	2273767	C	G	*dosT*	Neutral	Two component systems: sensor part of the two component regulatory system DEVR/DEVS/dost	
	2726323	C	G	*ahpC*	Neutral	Oxidative and nitrosative stresses: involved in oxidative stress response	
	2745430	C	G	*ndkA*	Neutral	Oxidative and nitrosative stresses: major role in the synthesis of nucleoside triphosphates other than ATP	
	3448598	T	C	*Rv3083*	Neutral	Mycolic acid synthesis: required for maintaining the appropriate mycolic acid composition and permeability of the envelope on its exposure to acidic pH	
	3612009	C	T	*Rv3236c*	Neutral	Cell wall proteins: involved in transport of undetermined substrate (possibly cations Na/H) across the membrane	
	3612768	C	G	*Rv3236c*	Deleterious	Cell wall proteins: involved in transport of undetermined substrate (possibly cations Na/H) across the membrane	
	4345420	G	T	*Rv3869*	Deleterious	Secretion system:ESX conserved component EccB1. ESX-1 type VII secretion system protein	
CLS_71	1587906	T	C	*lprG*	Deleterious	Lipoproteins: helps membrane protein Rv1410c (P55) transport triacylglycerides (TAG) across the inner cell membrane into the periplasm	
	2183137	C	G	*Rv1931c*	Neutral	Sigma Factors: controls the expression of genes important for virulence	
CLS_9	557618	C	A	*icl*	Deleterious	Others genes related in lipid synthesis: Involved in glyoxylate bypass (at the first step), an alternative to the tricarboxylic acid cycle	
	694465	C	T	*mce2F*	Neutral	Cell wall proteins: involved in host cell invasion	
CLS_217	200154	T	C	*mce1B*	Deleterious	Cell wall proteins: involved in host cell invasion	
	332735	A	C	*Rv0277c*	Deleterious	Toxin VapC25	
	711271	T	C	*Rv0617*	Deleterious	Toxin VapC29	
	4291418	G	C	*mmpL8*	Neutral	Synthesis of complex lipids: involved in the transport of lipids, it has been shown to be required in the production of sulfolipid-1 (SL-1)	
CLS_21	236167	G	C	*Rv0198c*	Neutral	Metalloproteases: hydrolyzes peptides and/or proteins	**L4.8**
	1986936	A	C	*plcD*	Neutral	Others genes related in lipid synthesis: implicated in the pathogenesis at the level of intracellular survival	

	2190353	C	G	*Rv1937*	Neutral	Oxidative and nitrosative stresses: involved in electron transfer	
	2302033	G	A	*pks12*	Deleterious	Synthesis of complex lipids: involved in biosynthesis of mannosyl-beta-1-phosphomycoketide (MPM)	
	3274759	T	G	*drrC*	Neutral	Synthesis of complex lipids: involved in active transport of antibiotic and phthiocerol dimycocerosate (dim) across the membrane	
CLS_13	2930254	G	A	*Rv2601A*	Neutral	Antitoxin VapB41	
	3296972	C	A	*pks15*	Deleterious	Synthesis of complex lipids: polyketide synthase possibly involved in lipid synthesis	
	4347411	G	A	*Rv3870*	Neutral	Secretion system: ESX conserved component EccCa1. ESX-1 type VII secretion system protein	
CLS_47	2930254	G	A	*Rv2601A*	Neutral	Antitoxin VapB41	
	3296972	C	A	*pks15*	Deleterious	Synthesis of complex lipids: polyketide synthase possibly involved in lipid synthesis	
	1378596	A	C	*lpqY*	Neutral	Lipoproteins:involved in active transport of sugar across the membrane	
	1722094	C	T	*pks5*	Neutral	Synthesis of complex lipids:involved in polyketide metabolism	
	2213005	G	T	*mce3D*	Deleterious	Cell wall proteins: involved in host cell invasion	
	2213395	A	G	*mce3D*	Neutral	Cell wall proteins: involved in host cell invasion	
	3454736	G	A	*Rv3088*	Neutral	Mycolic acid synthesis: involved in synthesis of triacylglycerol	
CLS_93	2930254	G	A	*Rv2601A*	Neutral	Antitoxin VapB41	
	3296972	C	A	*pks15*	Deleterious	Synthesis of complex lipids: polyketide synthase possibly involved in lipid synthesis	
	2190368	C	G	*Rv1937*	Neutral	Oxidative and nitrosative stresses: involved in electron transfer	
	3285926	C	A	*mmpL7*	Neutral	Synthesis of complex lipids: involved in translocation of phthiocerol dimycocerosate (dim) in the cell wall	
CLS_28	2930254	G	A	*Rv2601A*	Neutral	Antitoxin VapB41	
	3296972	C	A	*pks15*	Deleterious	Synthesis of complex lipids: polyketide synthase possibly involved in lipid synthesis	
	206407	C	T	*mce1F*	Deleterious	Cell wall proteins: involved in host cell invasion	
	332883	C	T	*Rv0277c*	Deleterious	toxin VapC25	
	2213395	A	G	*mce3D*	Neutral	Cell wall proteins: involved in host cell invasion	
	2673235	A	G	*mbtB*	Neutral	Metals-Transporter proteins: involved in the biogenesis of the hydroxyphenyloxazoline-containing siderophore mycobactins	
	3292441	G	A	*pks1*	Neutral	Synthesis of complex lipids: polyketide synthase possibly involved in lipid synthesis	
CLS_119	332760	T	C	*Rv0277c*	Deleterious	Toxin VapC25	
	332916	T	G	*Rv0277c*	Neutral	Toxin VapC25	
	2211477	G	C	*mce3B*	Deleterious	Cell wall proteins: involved in host cell invasion	
	2301787 - 2301788	GC	CT	*pks12*	Neutral	Synthesis of complex lipids: involved in biosynthesis of mannosyl-beta-1-phosphomycoketide (MPM)	
	2868793	C	T	*Rv2547*	Deleterious	Antitoxin VapB19	
	3070333	G	T	*Rv2757c*	Deleterious	Toxin VapC21	
	3292720	T	C	*pks1*	Deleterious	Synthesis of complex lipids: polyketide synthase possibly involved in lipid synthesis	

	4288850	C	T	*mmpL8*	Deleterious	Synthesis of complex lipids: involved in the transport of lipids, it has been shown to be required in the production of sulfolipid-1 (SL-1)	
CLS_7	2213395	A	G	*mce3D*	Neutral	Cell wall proteins: involved in host cell invasion	
	221601-2216012	TG	T	*mce3F*	Reading frame alteration	Cell wall proteins: involved in host cell invasion	
	2930254	G	A	*Rv2601A*	Neutral	VapB41	
	3296843	A	G	*pks15*	Neutral	Synthesis of complex lipids: polyketide synthase possibly involved in lipid synthesis	
	3296972	C	A	*pks15*	Deleterious	Synthesis of complex lipids: polyketide synthase possibly involved in lipid synthesis	
	3447666	T	C	*virS*	Neutral	Other transcriptional regulators: have a role in the regulation of proteins necessary for virulence	
	4197895	C	G	*Rv3749c*	Deleterious	VapC50	

	Point	**Reference**	**Change**	**Gene**	**Mutation effect**	**Description**	
CLS 5	252083	A	C	*pckA*	Deleterious	Other Virulence Factors: Rate-limiting gluconeogenic enzyme	**L4.4.1.1**
	753088	C	T	*Rv0656c*	Neutral	Toxin VapC6	
	2211714	A	G	*mce3C*	Deleterious	Cell wall proteins: involved in host cell invasion	
	2234475	C	T	*Rv1991c*	Deleterious	Toxin MazF6	
	2296297	C	T	*pks12*	Neutral	Synthesis of complex lipids: involved in biosynthesis of mannosyl-beta-1-phosphomycoketide (MPM)	
	2496830	G	A	*Rv2224c*	Neutral	Cell wall proteins: converts unknown esters to corresponding free acid and alcohol	
	2628914	A	G	*plcB*	Neutral	Others genes related in lipid synthesis: implicated in the pathogenesis at the level of intracellular survival	
	2631374	C	A	*plcA*	Deleterious	Others genes related in lipid synthesis: implicated in the pathogenesis at the level of intracellular survival	
	2868358	G	A	*Rv2546*	Neutral	Toxin VapC18	
	3518391	C	T	*nuoG*	Neutral	Inhibition of apoptosis: involved in aerobic|anaerobic respiration	
	3612665	A	G	*Rv3236c*	Deleterious	Cell wall proteins: involved in transport of undetermined substrate (possibly cations Na/H) across the membrane	
	3650530	C	A	*ctpC*	Neutral	Metal exporters: metal cation-transporting ATPase	
	4343048	G	C	*Rv3867*	Neutral	Secretion system: ESX-1 secretion-associated protein EspH	
CLS_19	565808	C	G	*hbhA*	Neutral	Cell wall proteins: Required for extrapulmonary dissemination	**L4.9**
	687191	C	G	*mce2A*	Neutral	Cell wall proteins: involved in host cell invasion	
	1107537	C	T	*Rv0990c*	Neutral	Other Virulence Factors: novel heat shock protein	
	1443970	G	C	*Rv1290c*	Deleterious	Others proteins with unknown function	
	2188920	A	G	*Rv1937*	Neutral	Oxidative and nitrosative stresses: involved in electron transfer	
	2190491	G	C	*Rv1937*	Neutral	Oxidative and nitrosative stresses: involved in electron transfer	
	2301115	C	T	*pks12*	Neutral	Synthesis of complex lipids: involved in biosynthesis of mannosyl-beta-1-phosphomycoketide (MPM)	

	566330-566360	several bases	C	*hbhA*	Reading frame alteration	Cell wall proteins: Required for extrapulmonary dissemination	
	1378118	G	T	*lpqY*	Neutral	Lipoproteins:involved in active transport of sugar across the membrane	
CLS_152	1880936	C	T	*pks7*	Deleterious	Synthesis of complex lipids: involved in some intermediate steps for the synthesis of a polyketide molecule	**L4.6.1.1**
	2297287	G	T	*pks12*	Neutral	Synthesis of complex lipids: involved in biosynthesis of mannosyl-beta-1-phosphomycoketide (MPM)	
	3175335	C	T	*Rv2863*	Deleterious	Toxin VapC23	
	3181029	A	C	*rip*	Deleterious	Metalloproteases: controls membrane composition	
	736873	G	T	*mmaA4*	Neutral	Mycolic acid synthesis: Involved in mycolic acids modification	
	1724722	T	C	*pks5*	Neutral	Synthesis of complex lipids: involved in polyketide metabolism	
	1877973-1877974	CG	C	*pks7*	Reading frame alteration	Synthesis of complex lipids: involved in some intermediate steps for the synthesis of a polyketide molecule	
	2156096	G	A	*katG*	Neutral	Oxidative and nitrosative stresses: play a role in the intracellular survival of mycobacteria within macrophages	
CLS_50	2674765	G	A	*mbtB*	Deleterious	Metals-Transporter proteins: involved in the biogenesis of the hydroxyphenyloxazoline-containing siderophore mycobactins	**L4.1.1.3**
	3285146	C	T	*mmpL7*	Neutral	Synthesis of complex lipids: involved in translocation of phthiocerol dimycocerosate (dim) in the cell wall	
	3599381	C	A	*sigH*	Deleterious	Sigma factors:plays a role in the oxidative-stress response	
	3652581	G	A	*ctpC*	Deleterious	Metal exporters: metal cation-transporting ATPase	
	3982149	T	C	*fadE29*	Deleterious	Catabolism of cholesterol: involved in lipid degradation	
	4209518	T	A	*lpqH*	Deleterious	Lipoproteins: shown to inhibit gamma interferon regulated HLA-DR protein and mRNA expression in human macrophages	

Trabajo 8

AmpliSeq technology for rapid lineage and drug-resistance identification in clinical samples of *Mycobacterium tuberculosis*

Jessica Comín[1*], Jesús Viñuelas[2,3], Carmen Lafoz[4], Alberto Cebollada[1], Daniel Ibarz[5], María-José Iglesias[5,6,7] and Sofía Samper[1,6,7]

[1]Instituto Aragonés de Ciencias de la Salud, Zaragoza, C/de San Juan Bosco, 13, 50009, Zaragoza, Spain
[2]Hospital Universitario Miguel Servet, Paseo Isabel la Católica, 1-3, 50009, Zaragoza, Spain
[3]Grupo de Estudio de Infecciones por Micobacterias (GEIM), Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica, C/Agustín de Bentacourt, No 13, 28003, Madrid, Spain
[4]Servicio General de Apoyo a la Investigación, Servicio de Análisis Microbiológico, Universidad de Zaragoza, C/Pedro Cerbuna, 12, 50009 Zaragoza, Spain
[5]Grupo de Genética de Micobacterias, Facultad de Medicina, Universidad de Zaragoza, C/Domingo Miral S/N, 50009, Zaragoza, Spain
[6]Fundación IIS Aragón, C/de San Juan Bosco, 13, 50009, Zaragoza, Spain
[7]CIBER de Enfermedades Respiratorias, Av. Monforte de Lemos, 3-5. Pabellón 11, Planta 0, 28029, Madrid, Spain
Correspondence*: Jessica Comín jcomin.iacs@aragon.es

Abstract

Background: *Mycobacterium tuberculosis* is a slow growing bacterium, which could delay the diagnosis. Whole genome sequencing allows for obtaining the complete drug-resistance profile of the strain; however, bacterial cultivation of clinical samples and complex processing is required. **Methods:** In this work, we explore the AmpliSeq, an amplicon-based enrichment method for preparing libraries for targeted next-generation sequencing, to identify lineage and drug resistance directly from clinical samples. **Results:** 111 clinical samples were tested. The lineage was identified in 100% of the culture-coming samples (52/52), in 95% of the smear (BK) positive clinical samples (38/40) and in 42.1% of the BK negative clinical samples (8/19). The drug-resistance profile was accurately identified in all but 11 samples, in which some phenotypic and genotypic discrepancies were found. In this respect, our panels were not certain in the detection of streptomycin resistance for isolates coming from clinical samples, as an extremely high number of SNPs in *rrs* and *rrl* genes were detected due to cross contamination. **Conclusion:** The technique has demonstrated high sensitivity, as even those samples with DNA concentrations below the detection limit of Qubit gave a result. Lastly, AmpliSeq technology is cheaper than whole genome sequencing and easy to perform by laboratory technicians.

1. Introduction

Mycobacterium tuberculosis, a slow growing bacterium, requires two to four weeks to grow in culture, which considerably delays diagnosis, especially the drug susceptibility tests (DST). Some rapid tests to detect *M. tuberculosis* and drug-resistances are available, nevertheless the number of mutations and resistances detected by them is limited, especially for second-line drugs and for the new drugs such as linezolid or bedaquiline. DNA can be extracted from liquid mycobacterial growth indicator tubes[1] or directly from clinical samples[2–5]. Nowadays, whole genome sequencing (WGS) has become affordable; it provides the most exhaustive information about strain drug-resistance profile[6]. In this way, the timing for diagnosis and the optimization of the treatment is considerably reduced from weeks to days. However, some difficulties are found for sequencing samples with low DNA concentrations. The knowledge provided by WGS allows to focus on specific targets, diminishing the amount of unnecessary data and making the analysis easier. In particular, Ion AmpliSeq™ technology can design a multiplex PCR amplifying hundreds of sites of interest. In addition, the amount of DNA required for AmpliSeq is lower than for WGS, improving the sensitivity, and is cheaper as the sequencing is targeted instead of being on a whole genome level.

The objective of this work was to test the efficiency of AmpliSeq, using DNA extracted directly from clinical samples of patients infected with *M. tuberculosis*, and the quality of the data obtained regarding lineage identification and drug-resistance profile.

2. Materials and Methods

2.1 Set-up of a DNA extraction method from clinical sample

Three DNA extraction methods from clinical samples were tested (Figure 1). The first method, based in a mechanical cell disruption, was previously described[1]. The second method uses a chemical disruption using the buffers and enzymes provided by the MolYsis Basic5 kit, following the manufacturer's instructions. For the third method, the human DNA was also eliminated using the MolYsis Basic5 kit and a different chemical cell disruption was used, based in NaCl/CTAB[7], the same used for extracting DNA from bacterial cultures. For testing the efficiency of the extraction, we used a VIASURE *Mycobacterium tuberculosis* complex Real time PCR detection kit.

2.2 Samples analysed by AmpliSeq technology

Since 2004, a surveillance protocol is being carried out in Aragon, Spain. As part of this protocol, all *M. tuberculosis* isolates are IS*6110*-RFLP and Spoligotyping genotyped. DSTs

are performed by the microbiologist in the clinical laboratories. Sputum samples (54), biopsies (5), bronchoalveolar lavages (BAL) (3), bronchial aspirates (6), pleural fluids (4) and others (3) were collected and stored during the period 2018-2022 in the Hospital Miguel Servet in Zaragoza, Aragon, Spain. *M. tuberculosis* DNA was extracted from the samples and a standard PCR amplifying a 531 bp-region of *dnaA* gene (dnaAj-F: CAATCGACAAAGCGCTGGC; dnaAj-R: TGGGGTGTGTGTTGGGT) was performed to confirm *M. tuberculosis* DNA presence. A positive result was obtained in 42 of the 75 samples (Table 1). Seventeen samples with a negative result were included to test the sensitivity of the technique, three with a positive and fourteen with a negative BK result. We included 52 additional DNA isolated from bacterial cultures to compare the quality of the lineage and drug-resistance results to the DNA obtained directly from clinical samples. A total of 111 samples were included in the study. The DNA was quantified using Qubit. All patients' data remained anonymous. Our regional ethical committee (CEICA, Record No. 20/2018) approved the methodology, as detailed in 18/0336 project.

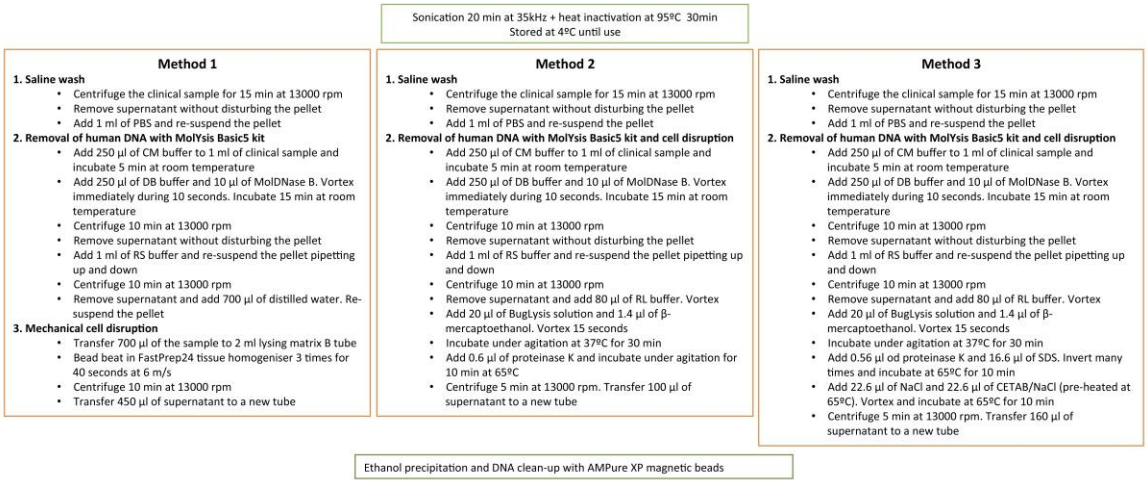


Figure 1. Description of the three protocols tested for DNA extraction directly from clinical samples. Samples with more than 1 ml were worked in different vials and the pellets were collected together after the saline washing step. Method 1 was described in Votintseva et al. (3).

2.3 Ampliseq technology

Two different AmpliSeq panels were designed for lineage and drug-resistance identification using the Ion AmpliSeq Designer tool of Thermo Fisher Scientific platform. The first version (IAD88259) was designed based on the Ion AmpliSeq™ TB Research Panel developed for the *M. tuberculosis* drug-resistance diagnosis, adding new targets for a wider resistance profile and to provide lineage information of *M. tuberculosis* complex.

There were 29 targets, with 210 amplicons of a range of 125-375 bp covering about 20 kb. Two years later, a new version of the panel (IAD137392) was developed in order to complement the former version with updated targets. It included 68 targets, with 268 amplicons of a range of 125-375 bp and a coverage of the 98.13%. The estimation of the total number of bases covered in the design was 52.65 kb. The targets included in both panels are listed in Table S1. The library preparation was carried out in the Ion Chef and quantification was performed using the Ion library Taqman quantitation kit. The AmpliSeq was carried out in the IonGene Studio™ S5 System using Ion 530™ Chip. The sequences obtained were mapped against the reference strain H37Rv (NC_000962.3) and a coverage analysis, followed by a variant calling, were performed.

2.4 Bioinformatic analysis

A pipeline in R software to obtain an automatic result was developed. In this way, the mutations of interest were marked with their significance, and automatically, the variants detected in the different samples were associated. For the drug-resistances, the catalogue of mutations from the World Health Organization and PhyResSE website were considered[8].

3. Results

3.1 Set-up of a DNA extraction method from clinical sample

The first step was to achieve a good quality DNA and an efficient extraction, as many clinical samples had a low bacterial load. Three different protocols were tested with two clinical samples (Figure 1). Before DNA extraction, the clinical samples were sonicated and heat inactivated. Extracted DNA was cleaned using magnetic beads (AMPure XP). We carried out a real time PCR using one VIASURE *M. tuberculosis* complex diagnostic kit. The PCR revealed that Method 1 had the highest yield (data not shown), and was chosen for extracting the DNA from the remaining clinical samples. The amount of DNA obtained in many of the clinical samples was below the detection limit of Qubit, but we included them in the study to test the sensitivity of the AmpliSeq technique, as some of those samples amplified by conventional PCR (Table 1).

Sample	ng/µl	Kind of sample	BK	PCR	Sample	ng/µl	Kind of sample
96	LOW	Sputum	+	Yes	MTB-1	12,5	Bacterial culture
91	LOW	Sputum	+	Yes	MTB-2	37,6	Bacterial culture
955	8,14	Sputum	+	Yes	MTB-3	23,9	Bacterial culture
52	2,28	Sputum	+	Yes	MTB-4	40,6	Bacterial culture

344	0,488	Sputum	+	Yes	MTB-5	55,4	Bacterial culture
692	0,598	Sputum	+	Yes	MTB-6	43	Bacterial culture
785	3,12	Sputum	+	Yes	MTB-7	11,2	Bacterial culture
532	0,162	Sputum	-	Yes	MTB-8	326	Bacterial culture
952	6	Sputum	+	Yes	MTB-9	35,8	Bacterial culture
684	8,5	Sputum	+	Yes	MTB-10	8,04	Bacterial culture
879	LOW	Pleural fluid	+	Yes	MTB-11	9,98	Bacterial culture
635	2,52	Biopsy	+	Yes	MTB-12	7,74	Bacterial culture
942	LOW	Biopsy	-	No	MTB-13	6,56	Bacterial culture
217	LOW	Sputum	+	Yes	MTB-14	2,66	Bacterial culture
659	5	Sputum	+	Yes	MTB-15	2	Bacterial culture
388	2,18	Sputum	+	Yes	MTB-16	1,56	Bacterial culture
658	LOW	Sputum	+	Yes	MTB-17	20,6	Bacterial culture
275	LOW	Sputum	+	Yes	MTB-18	11,8	Bacterial culture
698	5,72	Sputum	+	Yes	MTB-19	51,6	Bacterial culture
52	LOW	Sputum	+	Yes	MTB-20	2,86	Bacterial culture
542	1,67	Sputum	+	Yes	MTB-21	17,6	Bacterial culture
315	0.68	Sputum	+	Yes	MTB-22	10,3	Bacterial culture
40	LOW	Sputum	+	Yes	MTB-23	20,7	Bacterial culture
988	0.254	Sputum	-	Yes	MTB-24	8,4	Bacterial culture
212	0.368	Sputum	+	Yes	MTB-25	2,46	Bacterial culture
140	LOW	Sputum	+	Yes	MTB-26	40	Bacterial culture
381	LOW	Sputum	+	Yes	MTB-27	23	Bacterial culture
786	LOW	Sputum	+	Yes	MTB-28	2,96	Bacterial culture
644	LOW	Sputum	+	Yes	MTB-29	6,2	Bacterial culture
912	0.482	Sputum	+	Yes	MTB-30	43	Bacterial culture
120	LOW	Sputum	+	Yes	MTB-31	4,82	Bacterial culture
270	LOW	Biopsy	-	Yes	MTB-32	15,1	Bacterial culture
178	0,422	Sputum	+	Yes	MTB-33	472	Bacterial culture
537	0,454	Sputum	-	Yes	MTB-34	14,8	Bacterial culture
916	0,138	Sputum	+	Yes	MTB-35	139	Bacterial culture
521	0,122	Sputum	-	Yes	MTB-36	32,2	Bacterial culture
736	1,92	Sputum	+	Yes	MTB-37	60,6	Bacterial culture
69	LOW	Sputum	+	Yes	MTB-38	111	Bacterial culture
263	0,19	Sputum	+	Yes	MTB-39	21,4	Bacterial culture
453	7,22	Sputum	+	Yes	MTB-40	104	Bacterial culture
163	0,106	Aspirate	+	No	MTB-41	70,2	Bacterial culture
667	0,162	Aspirate	+	Yes	MTB-42	48,2	Bacterial culture
882	0,328	Sputum	+	Yes	MTB-43	3,92	Bacterial culture
007	LOW	Sputum	+	No	MTB-44	8,08	Bacterial culture
640	LOW	Lavage	-	No	MTB-45	108,7	Bacterial culture
716	LOW	Aspirate	-	No	MTB-46	6,12	Bacterial culture
907	LOW	Pleural fluid	-	No	MTB-47	8,42	Bacterial culture
100	0,964	Sputum	-	No	MTB-48	6,24	Bacterial culture

561	LOW	Adeno puncture	-	No	MTB-49	2	Bacterial culture
327	LOW	Aspirate	-	No	MTB-50	14,3	Bacterial culture
343	LOW	Aspirate	-	No	MTB-51	0,744	Bacterial culture
590	0,114	Sputum	-	No	MTB-52	1,292	Bacterial culture
169	LOW	Pleural fluid	-	No			
273	1,54	Sputum	+	Yes			
884	0,152	Lavage	-	No			
295	0,202	Sputum	-	No			
318	LOW	Aspirate	+	No			
473	0,84	Sputum	-	No			
366	LOW	Pleural fluid	-	No			

Table 1. Characteristics of the samples and DNA concentration of clinical samples (after the extraction) and bacterial cultures (diluted from the original stock) used in the study. Fifty-five of the samples were analysed using the first version of the kit and 56 samples with the second version. LOW means that the concentration was below the detection limit of Qubit (0.1 ng).

3.2 AmpliSeq results

The genomic lineage was correctly assigned in all the samples coming from bacterial culture. Regarding the clinical samples, the lineage was identified in 95% of the BK positive samples (38/40) and in 42.1% in the BK negative samples (8/19). In this sense, AmpliSeq provided information for some strains with an unknown or undetermined spoligo-family, complementing their genotype (Table S2).

The drug-resistance genotypes were correctly identified in 90.1% of the samples, considering the previous information of the strains. Among the 17 rifampicin resistant strains analysed, nine harboured an additional mutation in *rpoB* (52.9%). Besides, seven out of the 12 (58.3%) rifampicin resistant strains analysed with IAD137392 panel (which added the *rpoC* as a target) harboured one or more mutations in *rpoC* gene. Furthermore, fluoroquinolone resistance was unexpectedly identified for two susceptible samples, 658 and 659 (both belonging to the same patient), as second line antibiotic tests are not usually performed in the routine when the strains show susceptibility to the first line drugs. Conversely, some discrepancies were found in 11 samples (Table S2). Besides, 26 out of the 38 clinical samples analysed with the second version of the panel showed *rrs* gene erroneously sequenced, containing high number of SNPs. Despite this, none of these strains was identified as streptomycin resistant.

Among the clinical samples with low concentration in Qubit, the 100% of them with a BK positive result and the 50% of them with a BK negative result were correctly identified,

both lineage and resistances, meaning a high sensitivity of the AmpliSeq technique. The results are shown in Table S2. The SNPs used for lineage and drug-resistance identification can be found in Table S3.

4. Discussion

In this work we showed a fast *M. tuberculosis* resistance diagnostic method based on the AmpliSeq methodology. Attending the processing time of the technique, the DNA extraction can be made in 4-5 hours, and the different AmpliSeq steps in one or two days, what means that in three days a complete profile of the strains, both lineage and resistance profile, can be obtained. This technique is safer than methods based on bacterial cultures, which need a biosafety infrastructure that could be less strict when working with clinical samples as the bacterial load is considerably lower. The cost per sample in our laboratory is around 165€, including the kits used for the DNA extraction, while the cost of WGS with IonTorrent technology is around 220€. As a limitation, the samples have to be worked from eight to eight to optimize the price of library preparation, as the optimum number of samples to load on the chip in the Ion Torrent sequencer is 32, therefore it is necessary to wait until the required number of samples is available to adjust expenses. Deeplex® Myc-TB commercial kit has been developed showing accurate results and predicting strain resistance to 15 anti-TB drugs in two days. Our AmpliSeq panel could be an alternative for laboratories with access to Ion Torrent platforms, as Deeplex works with Illumina technology. Our accuracy regarding the drug-resistance is comparable to the obtained with Deeplex (90% and 73-96%, respectively), and are even better regarding the lineage identification (95% against 42-73%, respectively)[5]. Attending the timing to obtain results, both techniques are equivalent.

We have demonstrated that a minimum amount of DNA is required for this technique. Some *M. tuberculosis* SNPs were detected in 63.2% of the BK negative samples, and the lineage was correctly identified in 42.1% of them, including some with a negative result in the standard PCR done after the DNA extraction, which highlights the sensitivity of the technique. However, AmpliSeq is not trustworthy for drug-resistance detection in samples with negative BK, in which the bacterial concentration is quite low. This was observed in sample 942, phenotypically resistant to isoniazid but no SNP was detected in the related genes by the AmpliSeq technology. Another advantage of AmpliSeq is that it is a faster way to obtain a broad susceptibility profile of the strains. This is very important when starting treatment to avoid providing a wrong treatment regimen. Although there are some rapid kits to test resistances, as *Xpert MTB/RIF*[9] for identifiying rifampicin

resistance, or MTBDR*sl*[10] for second-line drug-resistance, there is not one kit that includes all drug-resistances. Regarding the traditional susceptibility tests based on MGIT, they work well for rifampicin, isoniazid, fluoroquinolones and aminoglicosides, but they are less reliable for pyrazinamide and ethambutol[11], and can take weeks as they requiere bacterial growth. With the AmpliSeq panel, drug-resistances can be studied all at once, and multiple targets of interest can be included at any time. Some discrepancies were found among the genotype and phenotype drug-resistance profiles, especially for pyrazinamide resistance. All possible scenarios were observed regarding pyrazinamide resistance: a strain with a confident SNP in *pncA* but phenotypically susceptible, two strains with no mutations in *pncA* but phenotypically resistant and a strain with a SNP not described as conferring resistance and phenotypically resistant. Two reasons can be exposed: the less reliable DST for pyrazinamide, and maybe that the resistant mechanisms of pyrazinamide are not completely understood yet. In this line, the mutation in *pncA* Lys48Asn (sample MTB-22) should be considered as conferring resistance, as we observed this fact. The same could be applied to the mutation *katG* Ser315Arg (sample MTB-16) and *rpsL* Lys88Thr (sample MTB-20), as both strains were resistant to isoniazid and stremtomycin, respectively. Regarding sample MTB-20, some mutations in *gid* gene could explain the streptomycin resistance, as well as for sample 120 (phenotypically resistant to streptomycin but without any mutation in *rrs* or *rpsL* genes) and sample MTB-22 (sequenced with the first version of the panel in which *rrs* was not included). However, *gid* gene is not included as a target in our AmpliSeq panels. One limitation of our panel is the lack of some genes associated with resistance, the majority of them for second-line and recently incorporated drugs. That could be the reason why we did not detect ethionamide resistance in sample MTB-14. Another limitation involves streptomycin resistance in clinical samples as *rrs* gene is wrongly sequenced, with a lot of SNPs, which could be due to the fact that this gene has conserved regions among bacteria[12,13] and in the clinical samples presence of DNA not belonging to *M. tuberculosis* could interfere in the aligment. We observed bad sequencing of *rrs* only in clinical samples, not in DNA coming from cultures, afecting 26 out of 38 samples analysed with the second version of the panel, the one that contained *rrs* gene as a target. Based on this, the panel is not reliable for streptomycin resistance when applied to DNA extracted directly from clinical samples. Sample MTB-10 was amikacin resistant in the DST but the only mutation found as possibly implicated in the resistance is in *rrs* gene, although it was not described as conferring amikacin resistance, except with streptomycin resistance. Probably this mutation is also implicated in the amikacin resistance of this strain. Regarding strain MTB-21, with a confident SNP in *embB* but susceptible to ethambutol in the DST, other authors have also

found this mutation in susceptible isolates as well as in resistant isolates[3,14]. On the other hand, the AmpliSeq identified pyrazinamide and fluoroquinolones resistance in strains for which the DST of these drugs were not carried out, as normally only first line drugs are tested.

Resistant strains have been described to have reduced fitness[15]. Trying to understand this, some compensatory mutations have been observed for rifampicin resistant strains harbouring *rpoB* mutations in *rpoA* and *rpoC* genes[16], both enconding other subunits of the RNA polimerase. In this sence, *rpoC* was added as a target in our second version of the panel hypothesizing that some of the mutations found in *rpoB* and *rpoC* in the strains analysed could be compensatory.

Our panels also identify lineage and family, which is useful for genotyping as is faster than traditional genotyping techniques[17]. It is important to remark that some families have clinical relevance, for example Beijing family, known for being hypervirulent[18] and with a high prevalence of multi-drug-resistance[19].

The main advantage of AmpliSeq is the low DNA concentration required to obtain information regarding the lineage and the drug-resistances, allowing to directly process clinical samples. Besides, this technic is a bit cheaper than WGS, and a lower volume of computer data is generated, which could be an advantage. Thanks to the designed algorithm, the data is shown automatically with the significance of each SNP, so the interpretation of the results is fast and automatic.

5. Acknowledgments

Authors would like to thank: Dessislava Morinova for proofreading, the EPIMOLA group for supplying the genotyped bacterial DNA and genotypes used in this work, Gema Robles Clerencia and Marta Guardingo de la Riva for providing us with the clinical samples, CERTEST S.L. for providing the VIASURE *M. tuberculosis* complex Real Time PCR Detection Kit and Servicios Científico Técnicos, IACS (Servicio de Secuenciación y Genómica Funcional and Servicio de Biocomputación).

6. Funding

This study has been funded by ISCIII FIS18/0336 (Co-funded by European Regional Development Fund/European Social Fund "A way to make Europe"/"Investing in your future") and MICINN/AEI (PTQ2018-009754). J.C. was awarded a scholarship by the Government of Aragon/European Social Fund, “Building Europe from Aragon”.

7. References

1 Votintseva A A, Pankhurst L J, Anson L W, et al. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. J Clin Microbiol 2015; 53(4): 1137–1143.

2 Votintseva A A, Bradley P, Pankhurst L, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol 2017; 55(5): 1285–1298.

3 Brown A C, Bryant J M, Einer-Jensen K, et al. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J Clin Microbiol 2015; 53(7): 2230–2237.

4 Doyle R M, Burgess C, Williams R, et al. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing. J Clin Microbiol; 56(8).

5 Bonnet I, Enouf V, Morel F, et al. A Comprehensive Evaluation of GeneLEAD VIII DNA Platform Combined to Deeplex Myc-TB(®) Assay to Detect in 8 Days Drug Resistance to 13 Antituberculous Drugs and Transmission of Mycobacterium tuberculosis Complex Directly From Clinical Samples. Front Cell Infect Microbiol 2021; 11: 707244.

6 Cirillo D M, Cabibbe A M, De Filippo M R, et al. Use of WGS in Mycobacterium tuberculosis routine diagnosis. Int J Mycobacteriology [Epub ahead of print].

7 van Soolingen D, de Haas P E, Hermans P W, van Embden J D. DNA fingerprinting of Mycobacterium tuberculosis. Methods Enzymol 1994; 235: 196–205.

8 Feuerriegel S, Schleusener V, Beckert P, et al. PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. J Clin Microbiol 2015; 53(6): 1908–1914.

9 Lawn S D, Nicol M P. Xpert® MTB/RIF assay: development, evaluation and implementation of a new rapid molecular diagnostic for tuberculosis and rifampicin resistance. Future Microbiol 2011; 6(9): 1067–1082.

10 Theron G, Peter J, Richardson M, et al. The diagnostic accuracy of the GenoType(®) MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs. Cochrane database Syst Rev 2014; (10): CD010705.

11 World Health Organization. TB Diagnostics and Laboratory Services. Information Note.

12 Cabibbe A M, Spitaleri A, Battaglia S, et al. Application of Targeted Next-

Generation Sequencing Assay on a Portable Sequencing Platform for Culture-Free Detection of Drug-Resistant Tuberculosis from Clinical Samples. J Clin Microbiol; 58(10).

13 Simner P J, Salzberg S L. The Human "Contaminome" and Understanding Infectious Disease. N Engl J Med 2022; 387(10): 943–946.

14 Bakuła Z, Napiórkowska A, Bielecki J, Augustynowicz-Kopeć E, Zwolska Z, Jagielski T. Mutations in the embB gene and their association with ethambutol resistance in multidrug-resistant Mycobacterium tuberculosis clinical isolates from Poland. Biomed Res Int 2013; 2013: 167954.

15 Andersson D I, Levin B R. The biological cost of antibiotic resistance. Curr Opin Microbiol 1999; 2(5): 489–493.

16 Comas I, Borrell S, Roetzer A, et al. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. Nat Genet 2011; 44(1): 106–110.

17 García De Viedma D, Pérez-Lago L. The Evolution of Genotyping Strategies To Detect, Analyze, and Control Transmission of Tuberculosis. Microbiol Spectr; 6(5).

18 Ates L S, Dippenaar A, Ummels R, et al. Mutations in ppe38 block PE-PGRS secretion and increase virulence of Mycobacterium tuberculosis. Nat Microbiol 2018; 3(2): 181–188.

19 Borrell S, Gagneux S. Infectiousness, reproductive fitness and evolution of drug-resistant Mycobacterium tuberculosis. Int J Tuberc lung Dis Off J Int Union against Tuberc Lung Dis 2009; 13(12): 1456–1466.

Table S1. Amplicons of the first (IAD88259) and the second (IAD137392) version of the panels.

IAD88259 kit			
Amplicon name	**Target Start**	**Target End**	**Target size (pb)**
gyrA	7302	9818	2516
rpoB	759807	763325	3518
rpsL	781560	781934	374
inhA_pro	1673303	1673440	137
inhA	1674102	1675011	909
katG	2153889	2156211	2322
pncA	2288681	2289341	660
eis	2714124	2715432	1308
embB	4246514	4249810	3296
Icd2_74092	74091	74093	2
Rv0095c_105139	105138	105140	2
Ag85_157292	157291	157293	2
Rv0197_232574	232573	232575	2
phoT_913274	913273	913275	2
mgtC_2053987	2053986	2053988	2
hsp65_529006	529005	529007	2
Rv3221c_3597682	3597681	3597683	2
Rv2952_3304966	3304965	3304967	2
Rv1234_1377185	1377184	1377186	2
Rv0309_378404	378403	378405	2
fbpA_4266647	4266646	4266648	2
fbpC_157113	157112	157114	2
Rv2962c_3314412	3314411	3314413	2
mtc28_42747	42746	42748	2
Rv1258c_1406761	1406758	1406765	7
phoR_852606	852605	852607	2
phoR_852909	852908	852911	3
leuB_3352919_32	3352910	3352940	30
phoP_pro	851561	851563	2

IAD137392 kit			
Amplicon name	**Target Start**	**Target End**	**Target size (pb)**
Rv0309_378404	378403	378405	2
phoT_913274	913273	913275	2
ddn_3986844_3987299	3986840	3987299	459
Rv3221c_3597682	3597681	3597683	2
embB	4246514	4249810	3296
eis	2714124	2715432	1308

whiB7_3568300_3568679	3568300	3568679	379
mutT4_4393449_4195	4393449	4394195	746
phoR_852606	852605	852607	2
mmpL11_239733	239732	239734	2
pncA	2288681	2289341	660
leuB_3352919_32	3352910	3352940	30
inhA	1674102	1675011	909
glmU_1137518	1137517	1137519	2
mgtA_648757_9602	648757	649602	845
mkl_751999	751998	752000	2
Rv0197_232574	232573	232575	2
ctpV_1079697	1079696	1079698	2
fbpC_157112_157293_Ag85	157112	157293	181
Rv3354_RDRio_3769111_3769500	3769111	3769500	389
ctpH_513789_514245	513788	514246	458
Rv2962c_3314412	3314411	3314413	2
phoR_852909	852908	852911	3
phoP_pro	851561	851563	2
ddlA_3336796_3337917	3336796	3337917	1121
alr_3840194_3841420	3840194	3841420	1226
mutT2_1286595_1287020	1286595	1287020	425
rv0678_778990_779487	778940	779487	547
mgtC_2053987	2053986	2053988	2
rpoB	759807	763325	3518
rpsL	781560	781934	374
pks13_4260268	4260267	4260269	2
kasA_2518115_2519365	2518115	2519365	1250
inhA_pro	1673303	1673440	137
rpoC_766645	763370	767320	3950
kdpA_1152701	1152700	1152702	2
rrs_1472307_1472752	1472306	1472753	447
Icd2_74092	74091	74093	2
tlyA_1917924_1918746	1917924	1918746	822
fgd1_490783_491793	490733	491793	1060
rrl_1473658_1476795	1473658	1476795	3137
atpE_1461030_1461290	1461030	1461290	260
Rv0095c_105139	105138	105140	2
mgtC_2053455_2053987	2053455	2053985	530
mpa_2376135	2376134	2376136	2
Rv2952_3304966	3304965	3304967	2
rpfB_1128815_9166	1128815	1129166	351
gyrB_5240_7267_gyrA	5240	9818	4578

Rv1258c_1406761	1406758	1406765	7
Rv0260c_311613	311612	311614	2
ribD_2986839_2987615	2986839	2987615	776
fbpA_4266647	4266646	4266648	2
Rv1234_1377185	1377184	1377186	2
mycP3_355181	355180	355182	2
pks15/1_3296385_364	3296363	3296386	23
ahpC_2726119_2726780	2726119	2726780	661
katG	2153889	2156211	2322
Rv2629_2955958_6732	2955958	2956732	774
hsp65_529006	529005	529007	2
mtc28_42747	42746	42748	2
ubiA_4268925_4269833	4268925	4269833	908
adhB_855199	855198	855200	2
dfrA_3073130_3073629_3074471	3073130	3074471	1341
guaA_3813242	3813240	3813244	4

Table S2. AmpliSeq results for the 111 samples analysed. For each sample, the spoligo-family previously assigned based on SITVIT database (http://www.pasteur-guadeloupe.fr:8081/SITVIT2/), the family/lineage/genetic group assigned by the AmpliSeq and the drug resistances found by the Ampliseq are provided. The samples made with the first version of the panel are shadowed in green and the samples made with the second version of the panel are shadowed in orange. RIF=rifampicin, INH=isoniazid, PZA=pyrazinamide, EMB=ethambutol, STM=streptomycin, FQ=fluoroquinolones, AMK=amikacin, KAN=kanamycin, ETO=ethionamide. *Means that a discrepancy between the genotype and the phenotype was found.

Sample	Spoligo - Family	AmpliSeq - Family	Drug Resistance - Genotype	Drug-phenotype	Comments
7	T4_CEU1	L4.9/L4.10			
52	T4_CEU1	SCG 3b no Haarlem/PPG2			
91	LAM9	LAM			
96	Unknown	LAM			
100	T5				
120	Unknown	SCG 3b no Haarlem/PPG2	-15 c/t *fabG1*	INHr, STMr*	No mutation in *rrs* or *rpsL* was found
140	Unknown	LAM (L4.3.2)			
163	H3	Haarlem			
169	T1				
178	Unknown	PPG3 (L4.10)			
212	T1	PPG3 (L4.10)			
217	H1	Haarlem	-15 c/t *fabG1*	INHr	
263	Unknown	PPG3 (L4.10)			
270	BOVIS1_BCG	*M. bovis*	*pncA* His57Asp	PZAr	
273		Unknown			
275	T4_CEU1	SCG 3b no Haarlem/PPG2			
295	Unknown	Haarlem			
315		Unknown			
318	T1	SCG 3b no Haarlem/PPG2			
327	LAM9				

340	Unknown	LAM (L4.3.2)	-15 c/t *fabG1*	INHr	
343	H1	Haarlem			
344	LAM 3	LAM	-15 c/t *fabG1*	INHr	
366	T1				
381	Unknown	LAM			
388	T1	PPG3 (L4.10)			
453	T1	L4.9/L4.10			
473	T4_CEU1				
521	L4.8/L4.10	PPG3 (L4.10)			
532		Unknown			
537	L4.8/L4.10	PPG3 (L4.10)		PZAr *	No mutation in *pncA* was found
542	Unknown	Haarlem			
561	LAM5				
590	T1				
635	T1	PPG3 (L4.10)	-15 c/t *fabG1*	INHr	
640	LAM	LAM			
644	LAM4	LAM			
658	U	SCG 5 no LAM	*gyrA* Asp94Asn	FQ unk	
659	U	SCG 5 no LAM	*gyrA* Asp94Asn	FQ unk	
667	Unknown	LAM (L4.3.2)	-15 c/t *fabG1*	INHr	
684	T1	PPG3 (L4.10)			
692	Unknown	LAM			
698	LAM6	LAM			
716	X				
736	Unknown	Haarlem			
785	T1 o L4.8/L4.10	PPG3 (L4.10)			
786	H3	Haarlem			
879	LAM9	LAM			

882	Unknown	PPG3 (L4.10)			
884	*M. kansasii*	Unknown			
907	T1	Cameroon			
912	Unknown	Haarlem			
916	Unknown	West African 2/L6			
942	Unknown	Haarlem		INHr*	No mutations found
952	LAM6	LAM			
955	T1	PPG3 (L4.10)			
988	Unknown	Unknown			
069	LAM12_MAD1	LAM (L4.3.2)			
MTB-1	U	L1.2.1/SCG 5 no LAM			
MTB-10	H3	Haarlem	*rpoB* His445Leu, -15 c/t *fabG1*, *katG* Ser315Thr, *pncA* Val163 deletion, *embB* Met306Ile, *rrs* 906 a/t	RIFr, INHr, PZAr, EMBr, STMr, AMKr*	The SNP found in *rrs* was not described as conferring amikacin resistance
MTB-11	S	S (L4.4.1.1)	*rpoB* (Asp435Glu, His445Asn), *katG* Ser315Thr	RIFr, INHr	
MTB-12	S	S (L4.4.1.1)	*rpoB* (His445Asn), *katG* Ser315Thr, *pncA* Gly17Ser	RIFr, INHr, PZAs*	Mutation in *pncA* but phenotipically susceptible
MTB-13	Unknown	PPG3 (L4.10)	*rpoB* Ser450Leu, *katG* Ser315Thr, *embB* Gln497Arg, *rpsL* Lys43Arg	RIFr, INHr, EMBr, STMr	
MTB-14	Beijing	Beijing	*rpoB* Ser450Leu, *katG* Ser315Thr, *pncA* His57Arg, *embB* (Tyr334His, Gly406Ala), *rpsL Lys88Arg*, *gyrA (Ala90Val*, *Asp94Gly)*, -14 c/t *eis*	RIFr, INHr, PZAr, EMBr, STMr, FQr, KNr, AMKr, ETOr*	No mutation in *ethA* was found
MTB-15	LAM9	LAM	*rpoB* (His445Asn), *katG* Asn138Ser	RIFr, INHr	
MTB-16	T1	PPG3 (L4.10)	*rpoB* Ser450Leu, *katG* Ser315Arg	RIFr, INHr*	The SNP in *katG* was not describe as conferring resistance
MTB-17	T2	LAM (L4.3.3)	*rpoB* Asp435Val, *katG* Ser315Thr, *pncA* Gln10Arg, *embB* Tyr319Ser	RIFr, INHr, PZAr, EMBr	
MTB-18	Unknown	LAM	*rpoB* Ser450Leu, -15 c/t *fabG1*	RIFr, INHr	

MTB-19	Beijing	Beijing	*rpoB* Ser450Phe, *katG* Ser315Thr, *rpsL* Lys43Arg,	RIFr, INHr, PZAr*, STMr	No mutation in *pncA* was found
MTB-2	Unknown	MAF II/L6			
MTB-20	LAM9	LAM (L4.3.3)	*rpoB* Ser450Leu, *katG* Ser315Thr, *pncA* Asp49Asn, *embB* Met306Ile, *rpsL* Lys88Thr	RIFr, INHr, PZAr, EMBr, STMr*	The SNP in *rpsL* was not described as conferring resistance
MTB-21	Beijing	Beijing	*rpoB* Ser450Leu, *katG* Ser315Thr, *pncA* Leu4Ser, *embB* Gln497Arg, *rpsL* Lys43Arg,	RIFr, INHr, PZA unk, EMBs*, STMr	Mutation in *embB* but phenotipically susceptible
MTB-22	Cameroon related	LAM	*rpoB* Ser450Leu, -15 c/t *fabG1*, pncA Lys48Asn, *embB* Met306Val	RIFr, INHr, PZAr*, EMBr, STMr*	The SNP in *pncA* was not described as conferring resistance and no mutation was found in *rpsL*
MTB-23	Haarlem	Haarlem	*rpoB* Ser450Leu, *katG* Ser315Thr	RIFr, INHr	
MTB-24	Cameroon related	LAM	*rpoB Ser450Leu, -15 c/t fabG1, pncA Leu120Pro, embB Met306Val, rpsL Lys43Arg, gyrA* Asp94Gly, *inhA* Ser94Ala, -10 g/a *eis*	RIFr, INHr,PZAr, EMBr, SMTr, FQ unk, ETOr, KAN unk	
MTB-25	AFRI	MAF I/L5			
MTB-26	U	LAM			
MTB-27	Unknown	PPG3/6a			
MTB-28	LAM3	LAM			
MTB-29	AFRI_1	MAF II/L6			
MTB-3	Unknown	L4.7/6a			
MTB-30	U	SCG 6a/PPG3/L4.10			
MTB-31	Unknown	Haarlem			
MTB-32	Unknown	SCG 6a/PPG3/L4.10			
MTB-33	AFRI_2	MAF I/L5			
MTB-34	Unknown	MAF I/L5			
MTB-35	H1	Haarlem			
MTB-36	Unknown	L1 Indo Oceanic EAI, MANU2			
MTB-37	T1	SCG 6a/PPG3/L4.10			

MTB-38	Unknown	Haarlem			
MTB-39	T3	SCG 6a/PPG3/L4.10			
MTB-4	Beijing	Beijing			
MTB-40	Unknown	SCG 6a/PPG3/L4.10			
MTB-41	T5_MAD2	L4.7/6a			
MTB-42	Unknown	Haarlem	-15 c/t *fabG1*	INHr	
MTB-43	Unknown	SCG 6a/PPG3/L4.10			
MTB-44	H4	SCG 5 no LAM			
MTB-45	Unknown	SCG 6a/PPG3/L4.10			
MTB-46	U	SCG 6a/PPG3/L4.10			
MTB-47	Beijing	Beijing	*rpoB* Ser450Leu, *katG* Ser315Thr, *embB* Met306Val, *rpsL* Lys43Arg	RIFr, INHr, EMBr, STMr	
MTB-48	Beijing	Beijing	*rpoB* Ser450Leu, *katG* insertion, *pncA* insertion	RIFr, INHr, PZAr	
MTB-49	CAPRAE	*M. caprae*			
MTB-5	CAS1_Delhi	CAS1_Delhi			
MTB-50	AFRI_3	MAF I/L5			
MTB-51	T1	SCG 6a/PPG3/L4.10			
MTB-52	Beijing	Beijing			
MTB-6	T2	Ghana			
MTB-7	EAI5	EAI			
MTB-8	H3	Haarlem			
MTB-9	U	PPG3/L4.10			
O52		LAM			

Table S3. Points of interest for identification of lineage/family.

Gene	Point	Ref	SNP	Family/L/Genetic group	Aminoacid affected
mgtC	2053987	G	A	Haarlem	182
fbpC_Ag85	157292	C	T	LAM	103
fbpC_Ag85	157129	C	T	Dheli/CAS	157
adhB	855199	C	T	Beijing Asia ancestral 2	210
kdpA	1152701	C	T	Beijing Central Asia outbreak	230
hsp65	529006	C	T	*M. canetti*	133
pks13	4260268	G	C	L4.6.1	293
mycP3	355181	G	A	L4.4.1.1/S	228
*phoP*_pro-*46*	851512	G	A	L6, *M. bovis*	-
mgtA	648990	G	C	Haarlem	152
pro-*13 mutT2*	1286582			Beijing	-
*mutT2*_pro *-215 narG*	1287213			Beijing	-
phoR	852910	C	T	No H37Rv	172
phoR	852606	G	A	L5, L6, animal lineage	71
Rv0260c	311613	C	A	No H37Rv	349
icd2	74092	G	A	*M. africanum*, *M. bovis*	140
mkl	751999	C	T	Beijing Europe/Russia W148 outbreak	161
Rv0095c	105139	C	A	Beijing	26
leuB	3352932	C	G	Haarlem, *M. bovis, M. africanum*, Beijing, CAS, EAI, Ghana	179
phoT	913274	C	G	L5, L6, *M. bovis*, EAI	183
Rv0309	378404	G	A	L6	158
glmU	1137518	G	A	L7	181
ctpH	513789	C	T	Beijing Asia/Africa clade 1	511
Rv2962c	3314412	A	G	PGG1 (*M. bovis*, EAI, CAS-Delhi, Beijing, *M. africanum*)	237

ctpV	1079697	G	C	Beijing Asia ancestral 3	318
mpa	2376135	A	G	Beijing Asia/Africa clade 2	52
mmpL11	239733			Beijing Asia ancestral 1	520
*Rv2629*_base 191	2955957	A	C	Beijing	64
*Rv2629*_base 695	2956731	C	T	Ghana	232
*pimB(mgtA)*_base 457	648992	C	G	S	153
*pimB(mgtA)*_ base 532	649067	C	G	Cameroon	178
*pimB(mgtA)*_base 221	648756	C	T	*M. africanum*	74
*pimB(mgtA)*_base 810	649345	C	T	EAI	270
*pimB(mgtA)*_base 1050	649585	G	A	*M. caprae*	350
*pimB(mgtA)*_base 1066	649601	C	T	*M. canettii*	356
rpfB_base 735	1128825	C	T	*M. microtti, M. pinipedii*	245
rpfB_base 1070	1129160	C	T	*M. bovis*	357
gyrA	7585	G	C	L no H37Rv/No PPG3	95
ctpH	514245	G	A	X	359
gyrA	8040	G	A	L4.3.3	246/247
icd2	74059	C	T	LAM, Cameroon	151/152
icd2	74092	G	A	*M. bovis, M. africanum*	140
rpoC	765150	G	A	L4.1	594/595
fgd1	491742	T	C	*M. africanum, M. bovis, M. canettii, M. caprae, M. microti, M. pinipedii,* Beijing, Delhi/CAS, EAI (L not Euro-American)	320/321
pimB	648856	T	C	*M. africanum, M. bovis, M. canettii, M. caprae, M. microti, M. pinipedii,* Beijing, Delhi/CAS, EAI (L not Euro-American)	107/108
rpoB	763031	T	C	No L4	1075/1076
katG	2154724	C	A	*M. africanum, M. bovis, M. canettii, M. caprae, M. microti, M. pinipedii,* Beijing, Delhi/CAS, EAI (L not Euro-American)	463/464
Rv2952	3304966	G	A	Beijing	526

Rv3908	4393590	C	G	Beijing	48
fgd1	491591	A	T	Haarlem	270/271
gyrB	6112	G	C	L1	330/331
gyrB	6124	C	T	L1.1.2	334/335
gyrA	8452	C	T	L1	384/385
rpoC	763884	CCC	TC A	L1	172/173
rpoC	765171	C	T	L1.1	601/602
embB	4247646	A	C	*M. africanum, M. bovis, M. canettii, M. caprae, M. microti, M. pinipedii*, EAI	378/379
rpoB	762434	T	G	L3	876/877
pncA	2289047	G	A	Dheli/CAS	65/66
oxyR':ahpC	2726105	G	A	Dheli/CAS, L3	-
fbpA	4266647	A	G	CAS	4
ubiA	4269351	G	A	*M. africanum, M. bovis*	161
atpE	1461251	G	T	L6 /West African 2	69
embB	4246864	C	T	*M. africanum, M. bovis*	117
Rv1234	1377185	C	G	MAF1/AFRI2/L5	70
Rv1258c	1406761	-	G	Beijing	193/194
Rv3221c	3597683	G	T	Indo-Oceanic, EAI_MANU	28
*inhA*_pro	1673338	G	A	L5/ West African 1	-
gyrA	9566	C	T	L5	755
inhA	1674434	T	C	L6 /West African 2	78
katG	2155503	G	A	*M. africanum, M. bovis*	203
embB	4249732	C	G	L4.7	1073
mutT2:narG	1287112	T	C	L6	-
rpoC	764995	C	G	L4.3	542
rrs	1472337	C	T	L4.3.2	-

Discusión general y Conclusiones

4. Discusión general y conclusiones

La presente tesis doctoral recoge un conjunto de trabajos en los que se han aplicado técnicas de secuenciación masiva para comprender mejor la epidemiología molecular de la TB en Aragón. Gracias al gran desarrollo de esta nueva tecnología en los últimos años, se ha convertido en una tecnología asequible para los laboratorios clínicos, permitiendo el remplazo de las técnicas de genotipificado tradicionales (174).

Para el estudio recogido en la Publicación 1 se combinaron técnicas de tipado tradicional (IS*6110*-RFLP y *Spoligotyping*) y las RDs del genoma con la tecnología AmpliSeq, una técnica de secuenciación masiva dirigida a los genes de interés para estudiar los *M. africanum* aislados en nuestra comunidad. En este caso, pudimos detectar SNPs para clasificar los aislados en L5 o L6. Además, mediante secuenciación capilar, se pudo confirmar la especificidad para el L6 de la inserción IS*6110* en el gen *moaX*, y que puede servir para identificarlo de forma rápida.

Para el estudio de la variabilidad genética de IS*6110* llevado a cabo en la Publicación 2 se utilizó la secuenciación capilar para obtener la secuencia completa de las distintas copias de IS*6110*, ya que al ser una secuencia repetida, mapeaban indistintamente con las secuencias de las copias de la cepa usada como referencia haciendo imposible su análisis. Una parte del estudio se realizó utilizando los genomas completos disponibles en el NCBI de cepas de referencia, como H37Rv, CDC1551, *M. africanum* GM041182 y *M. bovis* AF2122/97. Este estudio de la variabilidad de IS*6110* abre la puerta al debate sobre si las mutaciones encontradas en la secuencia afectan a la habilidad de trasposición, tal como parece observarse en las cepas de la familia X, cepas con al menos una de sus IS*6110* mutadas y con bajo número de copias a pesar de ser L4. A raíz de este estudio, desarrollamos un algoritmo que permite extraer las lecturas que contienen IS*6110* de entre los millones de lecturas obtenidos con la WGS y poder así descifrar su localización en el genoma mediante BLAST.

El estudio de TB recurrente recogido en la Publicación 3 pretendía explorar la recurrencia como una aproximación al fenómeno de latencia al tratarse de un proceso de reactivación de la enfermedad dentro del mismo paciente, a diferencia de otros trabajos en los que se estudiaba entre pacientes diferentes, suponiendo que uno había infectado al otro (175,176). Este trabajo fue realizado utilizando WGS para identificar linajes, SNPs y la tasa de mutación a lo largo del periodo de latencia, así como las localizaciones de las copias de IS*6110* utilizando el algoritmo de localización desarrollado. Se encontraron movimientos de IS*6110* entre aislados del mismo paciente (identificables por encontrarse en un número bajo de lecturas), así como SNPs presentes en el primer aislado pero que no se encontraron en el segundo, lo que sugiere la existencia de diferentes clones en el paciente. Nuestro trabajo, al contrario que los estudios anteriores, sugiere que la tasa de mutación es constante a lo largo de la latencia, pero estas diferencias de resultados se pueden deber a la distinta metodología utilizada.

En la presente tesis doctoral se ha utilizado la tecnología de WGS para el estudio de brotes de TB de interés en nuestra población, Publicaciones 4-7. A pesar de lo prometedora que es la WGS en epidemiología molecular, su aplicación al estudio de brotes de TB no ha tenido un resultado tan bueno como se esperaba porque presenta algunas dificultades difíciles de soslayar. Su utilidad en determinar el linaje al que pertenece una cepa, así como para su caracterización molecular y perfil de resistencias, es clara. Sin embargo, elucidar las cadenas de transmisión resulta mucho más complicado, aun cuando se tienen todos los casos del brote, como es el caso de la Publicación 4 incluida en esta tesis. En este brote, la variabilidad genética de los aislados era demasiado pequeña para que se pudieran distinguir, por lo que los datos epidemiológicos resultaron decisivos para establecer conexiones (177). En el caso de las Publicaciones 5 y 6, aunque no se disponía de todos los aislados del brote, estos abarcaban un periodo más largo de tiempo y las variaciones entre ellos permitieron identificar subgrupos de transmisión que explicaban cómo había sido la evolución de la transmisión a lo largo del tiempo, pero sin concretar la dirección de la transmisión en muchos de los casos.

Otro problema viene a la hora de fijar el límite de SNPs para considerar que dos aislados son resultado de la transmisión reciente de una misma cepa. Tradicionalmente, se ha establecido como criterio que dos aislados que tienen 12 o menos SNPs entre ellos son la misma cepa y a partir de ahí serían cepas distintas (178). Este umbral funcionó bien para los brotes estudiados en las Publicaciones 4, 5 y 6, y para la mayoría de los estudiados en la Publicación 7. Sin embargo, igual que en algunos casos permitió demostrar que dos aislados que eran considerados la misma cepa por tipado tradicional de IS*6110*-RFLP en realidad no lo eran, en otros casos resultó ambiguo, especialmente cuando el número de SNPs estaba entre 19 y 45. Teniendo en cuenta que las copias de IS*6110* se localizaban exactamente en el mismo sitio y el número de copias era el mismo, y que el número de SNPs está cerca de 12, parece lógico considerar que sí que son la misma cepa, lo que lleva a plantearse cómo de inamovible debe ser el umbral de 12 SNPs cuando el contexto de los aislados analizados no se conoce bien.

Con respecto al Trabajo 8, está centrado en el desarrollo de un protocolo para genotipificar y obtener el perfil de resistencias de las cepas directamente de muestra clínica, mediante una óptima extracción del ADN y la posterior aplicación de la tecnología de secuenciación AmpliSeq. Aunque actualmente existe el panel comercial Deeplex® Myc-TB, este está pensado para la tecnología de secuenciación de Illumina, mientras que el nuestro utiliza la tecnología Ion Torrent, por lo que podría ser útil para los laboratorios que tienen esta última plataforma. Además, hemos demostrado que los resultados obtenidos con nuestro panel son equivalentes e incluso mejores que los obtenidos con el panel comercial Deeplex® Myc-TB. El AmpliSeq presenta algunas ventajas frente a la WGS, como un coste inferior y una cantidad de ADN de partida también inferior.

Conclusiones

Publicación 1

1. Los casos de tuberculosis en Aragón provocados por *M. africanum* son escasos, y la mayoría de ellos se dan en inmigrantes del oeste de África.
2. El Linaje 6 es más frecuente que el Linaje 5 en Aragón.
3. Todas las cepas del L6 presentan una copia de IS*6110* en el gen *moaX*, lo que puede ser útil para su rápida identificación mediante un test basado en PCR.

Publicación 2

1. El análisis de las 158 secuencias de IS*6110*, pertenecientes a 55 cepas diferentes, mostró que el 13.3% de ellas portaban una mutación.
2. La mutación Gly215Ser del *orfB* parece ser característica de las cepas de la familia X. Esta mutación podría afectar de alguna manera a la transposición de IS*6110*, lo que explicaría que esta familia del Linaje 4 tenga un bajo número de copias, a pesar de que este linaje se caracteriza por tener alto número de copias.
3. Se observó un mayor porcentaje de copias mutadas entre las cepas de bajo número de copias (31.2% frente al 6.2% en las cepas de alto número de copias).
4. Este estudio incorpora las mutaciones en la secuencia de IS*6110* como un factor más para estudiar la variabilidad entre las diferentes familias del MTBC.

Publicación 3

1. Los pacientes VIH positivo sufren reactivación de la enfermedad durante los dos primeros años desde el episodio de tuberculosis inicial con más frecuencia que los VIH negativo.
2. Los movimientos de IS*6110* observados entre un aislado y el siguiente parecen ocurrir con más frecuencia en pacientes que han tardado más de dos años en sufrir la reactivación de la enfermedad. Además, se detectaron SNPs presentes en el primer aislado que no se encontraron en el segundo. Ambas observaciones sugieren la presencia de varios clones evolucionados en el paciente, de manera que en cada episodio, el clon dominante puede ser diferente al del episodio anterior.
3. Se encontró una tendencia descendente entre el número de SNPs entre aislados y el tiempo transcurrido entre episodios para periodos de tiempo más largos, si bien no fue significativo, por lo que la tasa de mutación parece ser constante durante todo el tiempo entre episodios.

Publicación 4

1. La WGS fue útil para caracterizar la cepa a nivel molecular.
2. Las características especiales de este brote, en el que había un superdiseminador que permaneció infeccioso durante dos años sin ser diagnosticado e infectó a la mayoría de casos del brote, provocó que la WGS por sí sola no permitiera resolver la cadena de transmisión.

3. La información epidemiológica fue fundamental para establecer las conexiones entre los casos, dada la poca variabilidad genética entre ellos.

Publicación 5

1. La WGS permitió descartar cuatro casos que habían sido incluidos en este brote mediante la técnica de tipado IS*6110*-RFLP al demostrar que se trataba de cepas diferentes a la estudiada.
2. La cepa causante de este brote, que circuló entre la población aragonesa durante al menos 25 años, afectó especialmente a población vulnerable como pacientes VIH positivo, usuarios de drogas intravenosas, personas encarceladas y personas sin hogar.
3. Pudimos observar que la cepa no desarrolló resistencias a antibióticos a pesar del mal cumplimiento del tratamiento por muchos de los casos.
4. La WGS fue útil para identificar tres subgrupos de transmisión dentro del brote. Uno de ellos presentaba un SNP en un factor de virulencia, que podría ser el responsable de la exitosa transmisión de esta variante entre la población aragonesa.

Publicación 6

1. La cepa MtZ ha producido el brote más grande que se ha dado en Aragón desde que empezaron los registros, y sigue activo a día de hoy.
2. La WGS permitió estudiar la cepa y el brote a nivel molecular, permitiendo trazar la evolución de la misma a lo largo del tiempo. Además, se pudieron identificar diferentes subgrupos de transmisión que ayudaron a elucidar algunas cadenas de transmisión, si bien no se pudieron establecer para todos los casos.
3. Esta cepa presenta SNPs en genes considerados factores de virulencia, así como en genes relacionados con supervivencia y patogénesis, lo que podría estar detrás de su éxito entre la población aragonesa.

Publicación 7

1. Este trabajo recoge las características moleculares (linaje, presencia o ausencia de RDs, localización de IS*6110* y SNPs en genes considerados factores de virulencia) de las 26 cepas responsables de los mayores brotes que han tenido lugar en Aragón, estudiadas mediante WGS.
2. Gracias a la WGS, se pudo establecer que algunos brotes analizados por IS*6110*-RFLP no estaban provocados por la misma cepa, comprobándose el mayor poder discriminatorio de la secuenciación masiva.
3. Se han caracterizado las cepas más prevalentes en nuestra comunidad, cuyos genomas quedan disponibles para considerarse en futuros protocolos de vigilancia.

Trabajo 8

1. Las principales ventajas del AmpliSeq frente a la WGS son la pequeña cantidad de ADN necesaria para su realización y su menor precio.
2. La interpretación de los resultados es rápida y automática gracias al software que se ha desarrollado en el trabajo.

3. Nuestro panel detectó correctamente el linaje de las cepas en el 100% de las que venían de cultivo, en el 95% de las muestras clínicas con un BK positivo y en el 41.2% de las que tenían BK negativo. El genotipo de resistencia fue identificado correctamente en el 90.1% de las muestras.

Referencias

5. Referencias

1. Daniel TM. The history of tuberculosis. Respir Med. 2006 Nov;100(11):1862–70.
2. Hayman J. Mycobacterium ulcerans: an infection from Jurassic time? Lancet (London, England). 1984 Nov;2(8410):1015–6.
3. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omaïs B, Marmiesse M, et al. Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. PLoS Pathog. 2005 Sep;1(1):e5.
4. Kapur V, Whittam TS, Musser JM. Is Mycobacterium tuberculosis 15,000 years old? Vol. 170, The Journal of infectious diseases. United States; 1994. p. 1348–9.
5. Brosch R, Gordon S V., Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A. 2002;
6. Zimmerman MR. Pulmonary and osseous tuberculosis in an Egyptian mummy. Bull N Y Acad Med. 1979 Jun;55(6):604–8.
7. Barberis I, Bragazzi NL, Galluzzo L, Martini M. The history of tuberculosis: from the first historical records to the isolation of Koch's bacillus. J Prev Med Hyg. 2017 Mar;58(1):E9–12.
8. Hippocrates. Book 1 - Of the epidemics. In: Adams F (Ed.). The genuine works of Hippocrates. Sydenham Soc. 1849;
9. Daniel TM. Pioneers of medicine and their impact on tuberculosis. Rochester: University Rochester Press; 2000.
10. Roberts CA BJ. The bioaerchaeology of tuberculosis. A global view on a reemerging disease. Gainesville: University of Florida Press; 2003.
11. Saeed BW. Malignant tuberculosis. Vol. 18, Journal of Ayub Medical College, Abbottabad : JAMC. Pakistan; 2006. p. 1–2.
12. Frith J. History of tuberculosis Part 1 – Pthisis, consumption and the White Plague. J Mil Veterans' Heal. 2014;22:2.
13. Gradmann C. Robert Koch and the pressures of scientific research: tuberculosis and tuberculin. Med Hist. 2001 Jan;45(1):1–32.
14. Armocida E, Martini M. Tuberculosis: a timeless challenge for medicine. J Prev Med Hyg. 2020 Jun;61(2):E143–7.
15. Sakula A. Carlo Forlanini, inventor of artificial pneumothorax for treatment of pulmonary tuberculosis. Thorax. 1983 May;38(5):326–32.
16. Luca S, Mihaescu T. History of BCG Vaccine. Maedica (Buchar). 2013 Mar;8(1):53–8.
17. Keshavjee S, Farmer PE. Tuberculosis, drug resistance, and the history of modern medicine. N Engl J Med. 2012 Sep;367(10):931–6.
18. Dye C, Williams BG. Eliminating human tuberculosis in the twenty-first century. J R Soc Interface. 2008 Jun;5(23):653–62.
19. World Health Organization. Global tuberculosis report 2021. 2021.
20. World Health Organization. Datos de tuberculosis en España en 2020 [Internet]. Available from: https://worldhealthorg.shinyapps.io/tb_profiles/?_inputs_&entity_type=%22country%22&lan=%22ES%22&iso2=%22ES%22
21. European Centre for Disease Prevention and Control/WHO Regional Office for Europe. Tuberculosis surveillance and monitoring in Europe 2019 – 2017 data. 2019.

22. Salud Pública. Boletín Epidemiológico Semanal de Aragón. 2020.

23. Salud Pública. Boletín Epidemiólogico Semanal de Aragón. 2021.

24. Fedrizzi T, Meehan CJ, Grottola A, Giacobazzi E, Fregni Serpini G, Tagliazucchi S, et al. Genomic characterization of Nontuberculous Mycobacteria. Sci Rep. 2017 Mar;7:45258.

25. Dulberger CL, Rubin EJ, Boutte CC. The mycobacterial cell envelope - a moving target. Nat Rev Microbiol. 2020 Jan;18(1):47–59.

26. Rogall T, Wolters J, Flohr T, Böttger EC. Towards a phylogeny and definition of species at the molecular level within the genus Mycobacterium. Int J Syst Bacteriol. 1990 Oct;40(4):323–30.

27. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. Nat Genet. 2013 Feb;45(2):172–9.

28. Koeck J-L, Fabre M, Simon F, Daffé M, Garnotel E, Matan AB, et al. Clinical characteristics of the smooth tubercle bacilli "Mycobacterium canettii" infection suggest the existence of an environmental reservoir. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis. 2011 Jul;17(7):1013–9.

29. Brites D, Gagneux S. Co-evolution of Mycobacterium tuberculosis and Homo sapiens. Immunol Rev. 2015 Mar;264(1):6–24.

30. Smith NH, Kremer K, Inwald J, Dale J, Driscoll JR, Gordon S V, et al. Ecotypes of the Mycobacterium tuberculosis complex. J Theor Biol. 2006 Mar;239(2):220–5.

31. Mostowy S, Cousins D, Brinkman J, Aranaz A, Behr MA. Genomic deletions suggest a phylogeny for the Mycobacterium tuberculosis complex. J Infect Dis. 2002 Jul;186(1):74–80.

32. Smith NH, Gordon S V, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG. Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Microbiol. 2006 Sep;4(9):670–81.

33. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM. Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proc Natl Acad Sci U S A. 2004 Apr;101(14):4871–6.

34. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, De Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2006;

35. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, et al. Functional and evolutionary genomics of Mycobacterium tuberculosis: Insights from genomic deletions in 100 strains. Proc Natl Acad Sci. 2004;

36. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS Biol. 2008 Dec;6(12):e311.

37. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet. 2010 Jun;42(6):498–503.

38. Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, et al. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. Emerg Infect Dis. 2013 Mar;19(3):460–3.

39. Coscolla M, Gagneux S. Does M. tuberculosis genomic diversity explain disease diversity? Drug Discov Today Dis Mech. 2010;7(1):e43–59.

40. Coscolla M, Gagneux S. Consequences of genomic diversity in Mycobacterium tuberculosis. Semin Immunol. 2014 Dec;26(6):431–44.

41. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, et al. Origin, spread and demography of the Mycobacterium tuberculosis complex. PLoS Pathog. 2008 Sep;4(9):e1000160.

42. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. 2013;45(10):1176–82.

43. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol. 2018 Apr;16(4):202–13.

44. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet. 2016;48(12):1535–43.

45. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 2014;5:4–8.

46. Gehre F, Antonio M, Otu JK, Sallah N, Secka O, Faal T, et al. Immunogenic Mycobacterium africanum strains associated with ongoing transmission in the Gambia. Emerg Infect Dis. 2013;

47. Bold TD, Davis DC, Penberthy KK, Cox LM, Ernst JD, De Jong BC. Impaired fitness of Mycobacterium africanum despite secretion of ESAT-6. J Infect Dis. 2012;

48. Asante-Poku A, Otchere ID, Osei-Wusu S, Sarpong E, Baddoo A, Forson A, et al. Molecular epidemiology of Mycobacterium africanum in Ghana. BMC Infect Dis. 2016;

49. Asante-Poku A, Yeboah-Manu D, Otchere ID, Aboagye SY, Stucki D, Hattendorf J, et al. Mycobacterium africanum Is Associated with Patient Ethnicity in Ghana. PLoS Negl Trop Dis. 2015;

50. Coscolla M, Lewin A, Metzger S, Maetz-Rennsing K, Calvignac-Spencer S, Nitsche A, et al. Novel Mycobacterium tuberculosis complex isolate from a wild chimpanzee. Emerg Infect Dis. 2013;

51. Rahim Z, Möllers M, Te Koppele-Vije A, De Beer J, Zaman K, Matin MA, et al. Characterization of Mycobacterium africanum subtype i among cows in a dairy farm in Bangladesh using spoligotyping. Southeast Asian J Trop Med Public Health. 2007;

52. Alfredsen S, Saxegaard F. An outbreak of tuberculosis in pigs and cattle caused by Mycobacterium africanum. Vet Rec. 1992;

53. Gudan A, Artuković B, Cvetnić Ž, Špičić S, Beck A, Hohšteter M, et al. Disseminated tuberculosis in hyrax (Procavia capensis) caused by Mycobacterium africanum. J Zoo Wildl Med. 2008;

54. Castets M, Boisvert H, Grumbach F, Brunel M, Rist N. Les bacilles tuberculeux de type Africain: note préliminaire. Rev Tuberc Pneumol (Paris). 1968;

55. Meyer L, David HL. EVALUATION DE L'ACTIVITE UREASE ET DE L'ACTIVITE β-GLUCOSIDASE POUR L'IDENTIFICATION PRATIQUE DES MYCOBACTERIES. Ann Microbiol (Paris). 1979;

56. Thorel MF. Isolation of mycobacterium africanum from monkeys. Tubercle. 1980;

57. Winglee K, Manson McGuire A, Maiga M, Abeel T, Shea T, Desjardins CA, et al. Whole Genome Sequencing of Mycobacterium africanum Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction. PLoS Negl Trop Dis. 2016 Jan;10(1):e0004332.

58. Otchere ID, Coscollá M, Sánchez-Busó L, Asante-Poku A, Brites D, Loiseau C, et al. Comparative genomics of Mycobacterium africanum Lineage 5 and Lineage 6 from Ghana suggests distinct ecological niches. Sci Rep. 2018;

59. Isea-Peña MC, Brezmes-Valdivieso MF, González-Velasco MC, Lezcano-Carrera MA, López-Urrutia-Lorente L, Martín-Casabona N, et al. Mycobacterium africanum, an emerging disease in high-income countries? Int J Tuberc Lung Dis. 2012;

60. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998 Jun;393(6685):537–44.

61. Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList--10 years after. Tuberculosis (Edinb). 2011 Jan;91(1):1–7.

62. McEvoy CRE, Falmer AA, van Pittius NCG, Victor TC, van Helden PD, Warren RM. The role of IS6110 in the evolution of Mycobacterium tuberculosis. Tuberculosis. 2007;87(5):393–404.

63. Van Embden JDA, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: Recommendations for a standardized methodology. Journal of Clinical Microbiology. 1993.

64. Thierry D, Brisson-Noel A, Vincent-Levy-Frebault V, Nguyen S, Guesdon JL, Gicquel B. Characterization of a Mycobacterium tuberculosis insertion sequence, IS6110, and its application in diagnosis. J Clin Microbiol. 1990;28(12):2668–73.

65. Otal I, Martin C, Vincent-Levy-Frebault V, Thierry D, Gicquel B. Restriction fragment length polymorphism analysis using IS6110 as an epidemiological marker in tuberculosis. J Clin Microbiol. 1991;29(6):1252–4.

66. Brisson-Noel A, Aznar C, Chureau C, Nguyen S, Pierre C, Bartoli M, et al. Diagnosis of tuberculosis by DNA amplification in clinical practice evaluation. Lancet (London, England). 1991 Aug;338(8763):364–6.

67. Gonzalo-Asensio J, Pérez I, Aguiló N, Uranga S, Picó A, Lampreave C, et al. New insights into the transposition mechanisms of IS6110 and its dynamic distribution between Mycobacterium tuberculosis Complex lineages. PLoS Genet. 2018;

68. Sekine Y, Izumi KI, Mizuno T, Ohtsubo E, Ishihama A. Inhibition of transpositional recombination by OrfA and OrfB proteins encoded by insertion sequence IS3. Genes to Cells. 1997;2(9):547–57.

69. Wall S, Ghanekar K, McFadden J, Dale JW. Context-sensitive transposition of IS6110 in mycobacteria. Microbiology. 1999;145(11):3169–76.

70. Dale JW, Tang TH, Wall S, Zainuddin ZF, Plikaytis B. Conservation of IS6110 sequence in strains of Mycobacterium tuberculosis with single and multiple copies. Tuber Lung Dis. 1997;78(5–6):225–7.

71. Hermans PWM, Van Soolingen D, Bik EM, De Haas PEW, Dale JW, Van Embden JDA. Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. Infect Immun. 1991;59(8):2695–705.

72. Vera-Cabrera L, Hernández-Vera MA, Welsh O, Johnson WM, Castro-Garza J. Phospholipase region of Mycobacterium tuberculosis is a preferential locus for IS6110 transposition. J Clin Microbiol. 2001;

73. Beggs ML, Eisenach KD, Cave MD. Mapping of IS6110 insertion sites in two epidemic strains of Mycobacterium tuberculosis. J Clin Microbiol. 2000 Aug;38(8):2923–8.

74. Sampson SL, Warren RM, Richardson M, van der Spuy GD, van Helden PD. Disruption of coding regions by IS6110 insertion in Mycobacterium tuberculosis. Tuber lung Dis Off J Int Union against Tuberc Lung Dis. 1999;79(6):349–59.

75. Kurepina NE, Sreevatsan S, Plikaytis BB, Bifani PJ, Connell ND, Donnelly RJ, et al. Characterization of the phylogenetic distribution and chromosomal insertion sites of five IS6110 elements in Mycobacterium tuberculosis: non-random integration in the dnaA-dnaN region. Tuber lung Dis Off J Int Union against Tuberc Lung Dis. 1998;79(1):31–42.

76. Fang Z, Forbes KJ. A Mycobacterium tuberculosis IS6110 preferential locus (ipl) for insertion into the genome. J Clin Microbiol. 1997;35(2):479–81.

77. Fang Z, Doig C, Morrison N, Watt B, Forbes KJ. Characterization of IS1547, a new member of the IS900 family in the Mycobacterium tuberculosis complex, and its association with IS6110. J Bacteriol. 1999;181(3):1021–4.

78. Reyes A, Sandoval A, Cubillos-Ruiz A, Varley KE, Hernández-Neuta I, Samper S, et al. IS-seq: a novel high throughput survey of in vivo IS6110 transposition in multiple Mycobacterium tuberculosis

genomes. BMC Genomics. 2012;

79. Alonso H, Aguilo JI, Samper S, Caminero JA, Campos-Herrero MI, Gicquel B, et al. Deciphering the role of IS6110 in a highly transmissible Mycobacterium tuberculosis Beijing strain, GC1237. Tuberculosis [Internet]. 2011;91(2):117–26. Available from: http://dx.doi.org/10.1016/j.tube.2010.12.007

80. Warren RM, Sampson SL, Richardson M, Van Der Spuy GD, Lombard CJ, Victor TC, et al. Mapping of IS6110 flanking regions in clinical isolates of Mycobacterium tuberculosis demonstrates genome plasticity. Mol Microbiol. 2000;37(6):1405–16.

81. Yesilkaya H, Dale JW, Strachan NJC, Forbes KJ. Natural transposon mutagenesis of clinical isolates of Mycobacterium tuberculosis: how many genes does a pathogen need? J Bacteriol. 2005 Oct;187(19):6726–32.

82. Ho TBL, Robertson BD, Taylor GM, Shaw RJ, Young DB. Comparison of *Mycobacterium tuberculosis* Genomes Reveals Frequent Deletions in a 20 kb Variable Region in Clinical Isolates. Yeast [Internet]. 2000;1(4):272–82. Available from: https://www.hindawi.com/journals/ijg/2000/147574/abs/

83. Safi H, Barnes PF, Lakey DL, Shams H, Samten B, Vankayalapati R, et al. IS6110 functions as a mobile, monocyte-activated promoter in Mycobacterium tuberculosis. Mol Microbiol. 2004 May;52(4):999–1012.

84. Soto CY, Menéndez MC, Pérez E, Samper S, Gómez AB, García MJ, et al. IS6110 Mediates Increased Transcription of the phoP Virulence Gene in a Multidrug-Resistant Clinical Isolate Responsible for Tuberculosis Outbreaks. J Clin Microbiol. 2004;42(1):212–9.

85. Escombe AR, Oeser C, Gilman RH, Navincopa M, Ticona E, Martínez C, et al. The detection of airborne transmission of tuberculosis from HIV-infected patients, using an in vivo air sampling model. Clin Infect Dis an Off Publ Infect Dis Soc Am. 2007 May;44(10):1349–57.

86. Storla DG, Yimer S, Bjune GA. A systematic review of delay in the diagnosis and treatment of tuberculosis. BMC Public Health. 2008 Jan;8:15.

87. Cardona P-J. The Progress of Therapeutic Vaccination with Regard to Tuberculosis. Front Microbiol. 2016;7:1536.

88. Mitchell G, Chen C, Portnoy DA. Strategies Used by Bacteria to Grow in Macrophages. Microbiol Spectr. 2016 Jun;4(3).

89. Lee J, Remold HG, Ieong MH, Kornfeld H. Macrophage apoptosis in response to high intracellular burden of Mycobacterium tuberculosis is mediated by a novel caspase-independent pathway. J Immunol. 2006 Apr;176(7):4267–74.

90. Cardona P-J. Pathogenesis of tuberculosis and other mycobacteriosis. Enfermedades Infecc y Microbiol Clin (English ed). 2018 Jan;36(1):38–46.

91. Kaufmann SHE, Evans TG, Hanekom WA. Tuberculosis vaccines: time for a global strategy. Sci Transl Med. 2015 Feb;7(276):276fs8.

92. Bañuls A-L, Sanou A, Van Anh NT, Godreuil S. Mycobacterium tuberculosis: ecology and evolution of a human bacterium. J Med Microbiol. 2015 Nov;64(11):1261–9.

93. Gibson SER, Harrison J, Cox JAG. Modelling a Silent Epidemic: A Review of the In Vitro Models of Latent Tuberculosis. Pathog (Basel, Switzerland). 2018 Nov;7(4).

94. Veatch A V, Kaushal D. Opening Pandora's Box: Mechanisms of Mycobacterium tuberculosis Resuscitation. Trends Microbiol. 2018 Feb;26(2):145–57.

95. Kang DD, Lin Y, Moreno J-R, Randall TD, Khader SA. Profiling early lung immune responses in the mouse model of tuberculosis. PLoS One. 2011 Jan;6(1):e16161.

96. Wolf AJ, Linas B, Trevejo-Nuñez GJ, Kincaid E, Tamura T, Takatsu K, et al. Mycobacterium tuberculosis infects dendritic cells with high frequency and impairs their function in vivo. J Immunol.

2007 Aug;179(4):2509–19.

97. Soto-Ramirez MD, Aguilar-Ayala DA, Garcia-Morales L, Rodriguez-Peredo SM, Badillo-Lopez C, Rios-Muñiz DE, et al. Cholesterol plays a larger role during Mycobacterium tuberculosis in vitro dormancy and reactivation than previously suspected. Tuberculosis (Edinb). 2017 Mar;103:1–9.

98. Voskuil MI, Visconti KC, Schoolnik GK. Mycobacterium tuberculosis gene expression during adaptation to stationary phase and low-oxygen dormancy. Tuberculosis (Edinb). 2004;84(3–4):218–27.

99. Brzostek A, Pawelczyk J, Rumijowska-Galewicz A, Dziadek B, Dziadek J. Mycobacterium tuberculosis is able to accumulate and utilize cholesterol. J Bacteriol. 2009 Nov;191(21):6584–91.

100. Loebel RO, Shorr E, Richardson HB. The Influence of Foodstuffs upon the Respiratory Metabolism and Growth of Human Tubercle Bacilli. J Bacteriol. 1933 Aug;26(2):139–66.

101. Kumar A, Toledo JC, Patel RP, Lancaster JRJ, Steyn AJC. Mycobacterium tuberculosis DosS is a redox sensor and DosT is a hypoxia sensor. Proc Natl Acad Sci U S A. 2007 Jul;104(28):11568–73.

102. Ernst JD. The immunological life cycle of tuberculosis. Nat Rev Immunol. 2012 Jul;12(8):581–91.

103. Furin J, Cox H, Pai M. Tuberculosis. Lancet (London, England). 2019 Apr;393(10181):1642–56.

104. Suárez I, Fünger SM, Kröger S, Rademacher J, Fätkenheuer G, Rybniker J. The Diagnosis and Treatment of Tuberculosis. Dtsch Arztebl Int. 2019 Oct;116(43):729–35.

105. Pande T, Pai M, Khan FA, Denkinger CM. Use of chest radiography in the 22 highest tuberculosis burden countries. Vol. 46, The European respiratory journal. England; 2015. p. 1816–9.

106. Steingart KR, Henry M, Ng V, Hopewell PC, Ramsay A, Cunningham J, et al. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. Lancet Infect Dis. 2006 Sep;6(9):570–81.

107. No Title. Geneva; 2011.

108. World Health Organization. The use of liquid medium for culture and DST [Internet]. 2007. Available from: http://www.who.int/tb /laboratory/policy_liquid_medium_for_culture_dst/en/index.html

109. World Health Organization. TB Diagnostics and Laboratory Services. Information Note. [Internet]. 2014. Available from: http://www.who.int/tb/dots/lab.pdf

110. Lawn SD, Nicol MP. Xpert® MTB/RIF assay: development, evaluation and implementation of a new rapid molecular diagnostic for tuberculosis and rifampicin resistance. Future Microbiol. 2011 Sep;6(9):1067–82.

111. Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC, Dendukuri N. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. Cochrane database Syst Rev. 2014 Jan;2014(1):CD009593.

112. Maynard-Smith L, Larke N, Peters JA, Lawn SD. Diagnostic accuracy of the Xpert MTB/RIF assay for extrapulmonary and pulmonary tuberculosis when testing non-respiratory samples: a systematic review. BMC Infect Dis. 2014 Dec;14:709.

113. Pai M, Nicol MP, Boehme CC. Tuberculosis Diagnostics: State of the Art and Future Directions. Microbiol Spectr. 2016 Oct;4(5).

114. Ling DI, Zwerling AA, Pai M. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. Eur Respir J. 2008 Nov;32(5):1165–74.

115. Theron G, Peter J, Richardson M, Barnard M, Donegan S, Warren R, et al. The diagnostic accuracy of the GenoType(®) MTBDRsl assay for the detection of resistance to second-line anti-tuberculosis drugs. Cochrane database Syst Rev. 2014 Oct;(10):CD010705.

116. Pai M, Denkinger CM, Kik S V, Rangaka MX, Zwerling A, Oxlade O, et al. Gamma interferon release

assays for detection of Mycobacterium tuberculosis infection. Clin Microbiol Rev. 2014 Jan;27(1):3–20.

117. CDC. Hojas informativas. Prueba cutánea de la tuberculina [Internet]. Available from: https://www.cdc.gov/tb/esp/publications/factsheets/testing/skintesting_es.htm

118. World Health Organization. WHO: Treatment of tuberculosis: guidelines—4th edition. World Health Organization Press; 2010.

119. Migliori GB, Tiberi S, Zumla A, Petersen E, Chakaya JM, Wejse C, et al. MDR/XDR-TB management of patients and contacts: Challenges facing the new decade. The 2020 clinical update by the Global Tuberculosis Network. Int J Infect Dis IJID Off Publ Int Soc Infect Dis. 2020 Mar;92S:S15–25.

120. Coll P, García de Viedma D. Molecular epidemiology of tuberculosis. Enfermedades Infecc y Microbiol Clin (English ed). 2018 Apr;36(4):233–40.

121. van Soolingen D, de Haas PE, Kremer K. Restriction fragment length polymorphism typing of mycobacteria. Methods Mol Med. 2001;54:165–203.

122. Warren RM, van der Spuy GD, Richardson M, Beyers N, Booysen C, Behr MA, et al. Evolution of the IS6110-based restriction fragment length polymorphism pattern during the transmission of Mycobacterium tuberculosis. J Clin Microbiol. 2002 Apr;40(4):1277–82.

123. García De Viedma D, Pérez-Lago L. The Evolution of Genotyping Strategies To Detect, Analyze, and Control Transmission of Tuberculosis. Microbiol Spectr. 2018 Oct;6(5).

124. Kamerbeek J, Schouls L, Kolk A, Van Agterveld M, Van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J Clin Microbiol. 1997;

125. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, et al. Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 2006 Mar;6:23.

126. Frothingham R, Meeker-O'Connell WA. Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats. Microbiology. 1998 May;144 (Pt 5:1189–96.

127. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C. Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome. Mol Microbiol. 2000 May;36(3):762–71.

128. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. J Clin Microbiol. 2006;

129. Oelemann MC, Diel R, Vatin V, Haas W, Rüsch-Gerdes S, Locht C, et al. Assessment of an optimized mycobacterial interspersed repetitive- unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. J Clin Microbiol. 2007 Mar;45(3):691–7.

130. Cowan LS, Diem L, Monson T, Wand P, Temporado D, Oemig T V, et al. Evaluation of a two-step approach for large-scale, prospective genotyping of Mycobacterium tuberculosis isolates in the United States. J Clin Microbiol. 2005 Feb;43(2):688–95.

131. van Deutekom H, Supply P, de Haas PEW, Willery E, Hoijng SP, Locht C, et al. Molecular typing of Mycobacterium tuberculosis by mycobacterial interspersed repetitive unit-variable-number tandem repeat analysis, a more accurate method for identifying epidemiological links between patients with tuberculosis. J Clin Microbiol. 2005 Sep;43(9):4473–9.

132. de Beer JL, van Ingen J, de Vries G, Erkens C, Sebek M, Mulder A, et al. Comparative study of IS6110 restriction fragment length polymorphism and variable-number tandem-repeat typing of Mycobacterium tuberculosis isolates in the Netherlands, based on a 5-year nationwide survey. J Clin Microbiol. 2013 Apr;51(4):1193–8.

133. Barnes PF, Cave MD. Molecular epidemiology of tuberculosis. N Engl J Med. 2003 Sep;349(12):1149–

56.

134. Couvin D, David A, Zozio T, Rastogi N. Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the Mycobacterium tuberculosis genotyping database. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2019 Aug;72:31–43.

135. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463–7.

136. Chidgeavadze ZG, Beabealashvilli RS, Atrazhev AM, Kukhanova MK, Azhayev A V, Krayevsky AA. 2′,3′-Dideoxy-3′ aminonucleoside 5′-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. Nucleic Acids Res. 1984 Feb;12(3):1671–86.

137. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1–8.

138. Nyrén P, Lundin A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. Anal Biochem. 1985 Dec;151(2):504–9.

139. Hyman ED. A new method of sequencing DNA. Anal Biochem. 1988 Nov;174(2):423–36.

140. Voelkerding K V, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009 Apr;55(4):641–58.

141. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 2006 Feb;34(3):e22.

142. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008 Nov;456(7218):53–9.

143. Balasubramanian S. Sequencing nucleic acids: from chemistry to medicine. Chem Commun (Camb). 2011 Jul;47(26):7281–6.

144. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012 Jul;13:341.

145. Chaitankar V, Karakülah G, Ratnapriya R, Giuste FO, Brooks MJ, Swaroop A. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. Prog Retin Eye Res. 2016 Nov;55:1–31.

146. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011 Jul;475(7356):348–52.

147. Golan D, Medvedev P. Using state machines to model the Ion Torrent sequencing process and to improve read error rates. Bioinformatics. 2013 Jul;29(13):i344-51.

148. Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, et al. Virtual terminator nucleotides for next-generation DNA sequencing. Nat Methods. 2009 Aug;6(8):593–5.

149. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014 Sep;30(9):418–26.

150. Haque F, Li J, Wu H-C, Liang X-J, Guo P. Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA. Nano Today. 2013 Feb;8(1):56–74.

151. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009 Apr;4(4):265–70.

152. Li J, Stein D, McMullan C, Branton D, Aziz MJ, Golovchenko JA. Ion-beam sculpting at nanometre length scales. Nature. 2001 Jul;412(6843):166–9.

153. Dekker C. Solid-state nanopores. Nat Nanotechnol. 2007 Apr;2(4):209–15.

154. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative Genome Viewer. Nat Biotechnol. 2011;29(1):24–6.

155. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015 Aug;31(16):2745–7.

156. Freese NH, Norris DC, Loraine AE. Integrated genome browser: visual analytics platform for genomics. Bioinformatics. 2016 Jul;32(14):2089–95.

157. Hatherell HA, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: A systematic review. BMC Med. 2016;

158. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. 2013 Feb;13(2):137–46.

159. Mehaffy C, Guthrie JL, Alexander DC, Stuart R, Rea E, Jamieson FB. Marked microevolution of a unique Mycobacterium tuberculosis strain in 17 years of ongoing transmission in a high risk population. PLoS One. 2014;9(11):e112928.

160. Kato-Maeda M, Ho C, Passarelli B, Banaei N, Grinsdale J, Flores L, et al. Use of whole genome sequencing to determine the microevolution of Mycobacterium tuberculosis during an outbreak. PLoS One. 2013;8(3):e58235.

161. Smit PW, Vasankari T, Aaltonen H, Haanperä M, Casali N, Marttila H, et al. Enhanced tuberculosis outbreak investigation using whole genome sequencing and IGRA. Vol. 45, The European respiratory journal. England; 2015. p. 276–9.

162. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics. 2012;13 Suppl 8(Suppl 8):S8.

163. Guerra-Assunção JA, Houben RMGJ, Crampin AC, Mzembe T, Mallard K, Coll F, et al. Recurrence due to relapse or reinfection with Mycobacterium tuberculosis: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. J Infect Dis. 2015 Apr;211(7):1154–63.

164. Lambert M-L, Hasker E, Van Deun A, Roberfroid D, Boelaert M, Van der Stuyft P. Recurrence in tuberculosis: relapse or reinfection? Lancet Infect Dis. 2003 May;3(5):282–7.

165. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, et al. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. Lancet Respir Med. 2013 Dec;1(10):786–92.

166. Guerra-Assunção JA, Crampin AC, Houben RMGJ, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife. 2015 Mar;4.

167. Witney AA, Gould KA, Arnold A, Coleman D, Delgado R, Dhillon J, et al. Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. J Clin Microbiol. 2015 May;53(5):1473–83.

168. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: An observational study. Lancet Respir Med. 2014;

169. Stucki D, Ballif M, Bodmer T, Coscolla M, Maurer A-M, Droz S, et al. Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. J Infect Dis. 2015 Apr;211(8):1306–16.

170. Luo T, Yang C, Peng Y, Lu L, Sun G, Wu J, et al. Whole-genome sequencing to detect recent transmission of Mycobacterium tuberculosis in settings with a high burden of tuberculosis. Tuberculosis (Edinb). 2014 Jul;94(4):434–40.

171. Pérez-Lago L, Comas I, Navarro Y, González-Candelas F, Herranz M, Bouza E, et al. Whole genome sequencing analysis of intrapatient microevolution in Mycobacterium tuberculosis: potential impact on the inference of tuberculosis transmission. J Infect Dis. 2014 Jan;209(1):98–108.

172. Dookie N, Rambaran S, Padayatchi N, Mahomed S, Naidoo K. Evolution of drug resistance in Mycobacterium tuberculosis: a review on the molecular determinants of resistance and implications for personalized care. J Antimicrob Chemother. 2018 May;73(5):1138–51.

173. World Health Organization. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance. 2021.

174. Cirillo DM, Cabibbe AM, De Filippo MR, Trovato A, Simonetti T, Rossolini GM, et al. Use of WGS in Mycobacterium tuberculosis routine diagnosis. Int J Mycobacteriology. 2016;

175. Colangeli R, Arcus VL, Cursons RT, Ruthe A, Karalus N, Coley K, et al. Whole genome sequencing of Mycobacterium tuberculosis reveals slow growth and low mutation rates during latent infections in humans. PLoS One. 2014;9(3):e91024.

176. Colangeli R, Gupta A, Vinhas SA, Chippada Venkata UD, Kim S, Grady C, et al. Mycobacterium tuberculosis progresses through two phases of latent infection in humans. Nat Commun. 2020 Sep;11(1):4870.

177. Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F. Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. PLoS Med [Internet]. 2016;13(10):1–18. Available from: http://dx.doi.org/10.1371/journal.pmed.1002137

178. Lalor MK, Casali N, Walker TM, Anderson LF, Davidson JA, Ratna N, et al. The use of whole-genome sequencing in cluster investigation of a multidrug-resistant tuberculosis outbreak. Eur Respir J [Internet]. 2018;51(6):1702313. Available from: http://erj.ersjournals.com/lookup/doi/10.1183/13993003.02313-2017