Universitat Politècnica de València (UPV)

Escola Tècnica Superior D'Enginyeria Agronòmica i del Medi Natural (ETSEAMN)

Grado en Biotecnología

Trabajo de Fin de Grado

'GENOTIPADO A TIEMPO REAL BASADO EN SNPS PARA CARACTERIZAR LA VARIACIÓN DE LAS CEPAS DEL COMPLEJO DE MYCOBACTERIUM TUBERCULOSIS Y SU TRANSMISIÓN EN PAÍSES DE ALTA Y BAJA CARGA DE LA ENFERMEDAD'.

Alumna: Carla Mariner Llicer

Tutor: José Gadea Vacas

Cotutor externo: Iñaki Comas Espadas

Curso Académico: 2016/2017

Valencia, 10 de Julio de 2017.

TÍTULO: 'Genotipado a tiempo real basado en SNPs para caracterizar la variación de las cepas del complejo de *Mycobacterium tuberculosis* y su transmisión en países de alta y baja carga de la enfermedad'.

RESUMEN

El genotipado de aislados del complejo de *Mycobacterium tuberculosis* es una técnica que permite tanto entender la epidemiología y biología del microorganismo como su transmisión. Las micobacterias que pertenecen al complejo *Mycobacterium tuberculosis* se clasifican en distintos linajes y sublinajes que presentan polimorfismos de nucleótidos únicos (SNPs) que son ideales como marcadores genéticos para caracterizar las cepas procedentes de muestras de aislados clínicos.
Existen diversas técnicas de genotipado entre ellas el *SNP-typing* a tiempo real, que consiste en la realización de PCRs (reacción en cadena de la polimerasa) a tiempo real. La técnica emplea oligos que contienen el SNP característico de un determinado linaje. El análisis posterior de las curvas de fusión del producto de la PCR permite clasificar las muestras determinando la ausencia o presencia del SNP según la temperatura de fusión, técnica conocida como High Resolution Melting (HRM) o análisis de las curvas de fusión.
El siguiente trabajo consiste en la determinación del linaje de cepas del complejo de *Mycobacterium tuberculosis* procedentes de países de alta (Liberia) y baja incidencia (España). Para ello se utilizará un protocolo optimizado para detectar varios linajes en una sola PCR multiplex. Posteriormente, también se va a determinar el sublinaje al que pertenecen las muestras. Además, como cabe la posibilidad de que se hayan dado co-infecciones en los pacientes, se va a tratar de optimizar un método para poder detectarlas.

TITLE: Real-time SNP typing to characterize strain variation and transmission of *Mycobacterium Tuberculosis* Complex in high and low-burden countries.

ABSTRACT

Genotyping of *Mycobacterium tuberculosis* complex isolates allows understanding molecular epidemiology and biology of this microorganism but also how it is transmitted. *Mycobacterium tuberculosis* complex microorganisms can be classified into different lineages and sublineages, which have characteristic single nucleotide polymorphisms (SNPs). These SNPs can be used as genetic markers to classify *Mycobacterium tuberculosis* complex strains into phylogenetic lineages.
Nowadays, there are a lot of genotyping methods for *Mycobacterium tuberculosis* complex. One of them is based on SNPs and is called real time SNP-genotyping. This technique consists on real-time PCRs (polymerase chain reaction) using primers containing a specific SNP to recognize the strains that belong to its lineage. Moreover, the High-Resolution Melting (HRM) technique is the one that allows the identification of the SNP and samples classification according to the melting temperature by analyzing the melting curves from the PCR product.
This project consists of the classification of different samples coming from high (Liberia) and low (Spain) burden countries into lineages by using real-time SNP genotyping technique. To do that, an optimized multiplex PCR protocol is going to be used to determine different lineage in the same reaction. After that, the identification of strain sublineages is going to be performed. In addition to

that, we are going to try to optimize a procedure to detect mixed infections to consider the possibility that the patients have been infected by more than one *Mycobacterium tuberculosis* strains belonging to different lineages.

PALABRAS CLAVE:
Genotipado, *Mycobacterium tuberculosis*, SNP, PCR a tiempo real, linajes.

KEY WORDS
Genotyping, *Mycobacterium tuberculosis*, SNP, Real Time PCR, lineages.

Alumna: Carla Mariner Llicer
Tutor Académico: Prof. D. José Gadea.
Cotutor externo: D. Iñaki Comas.

Valencia, Julio de 2017.

AGRADECIMENTOS

En primer lugar, me gustaría agradecer a Iñaki la oportunidad de poder formar parte de su equipo durante estos meses para la realización de este proyecto y a Manoli por ser una gran mentora, por todos los consejos que me ha dado y por todo lo que he podido aprender con ellos. Pero, además, quería daros las gracias en general a todo el equipo, sobre todo, por haberme transmitido vuestra pasión por la ciencia.

Muchas gracias equipo TB, sin vosotros este trabajo no hubiera sido igual.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN.

1.1 Tuberculosis: descripción de la enfermedad.

La tuberculosis humana y animal (TB) es una enfermedad causada por un conjunto de micobacterias Gram positivas que pertenecen al complejo de *Mycobacterium tuberculosis* (MTBC). Entre el grupo de microorganismos que componen el MTBC se encuentra *Mycobacterium tuberculosis* que, junto a *Mycobacterium africanum*, son los patógenos obligados responsables de la enfermedad en humanos (Hershberg *et al.*, 2008). *M. tuberculosis* afecta mayoritariamente a los pulmones y se transmite de persona a persona por el aire pero, para poder transmitirse tiene que provocar la enfermedad, por lo que la transmisión es un punto clave del ciclo de la tuberculosis que aún no se ha conseguido controlar (Brites and Gagneux, 2015).El riesgo de padecer la enfermedad tras ser infectado es del 10%, no obstante, este porcentaje aumenta si el paciente presenta un sistema inmunitario comprometido, sobre todo si padece alguna enfermedad infecciosa como el VIH (Virus de la Inmunodeficiencia Humana) (WHO, 2016).

Esta enfermedad representa un grave problema a nivel mundial. Actualmente, se ha visto una disminución de casos en los países desarrollados, no obstante, sigue siendo una de las 10 principales causas de mortalidad y morbilidad en los países en vía de desarrollo (Zaman, 2010; WHO, 2017). Además, La incidencia de esta enfermedad es muy elevada en todo el mundo. En 2015 se produjeron 10.4 millones de nuevos casos y 1.8 millones de muertes (WHO, 2016). Un 26% de los nuevos casos aparecieron en los países africanos (WHO, 2017). También se tiene que tener en cuenta que, un tercio de los casos, 1.2 millones en 2015, están relacionados con co-infecciones con VIH y en torno a 480.000 personas están infectadas por cepas multiresistentes a fármacos (MDR) (WHO, 2016, 2017).

Por otro lado, existen muchos casos de TB, aproximadamente 1.7 billones en el mundo (en 2014), en los que la enfermedad se encuentra latente, lo que significa que han sido infectados pero aún no han desarrollado la enfermedad ni la pueden transmitir (WHO, 2016, 2017). Por esta razón, aumenta la necesidad de encontrar nuevas técnicas de diagnóstico y de buscar nuevas vacunas y tratamientos para mejorar el control de la enfermedad (Houben and Dodd, 2016).

En cuanto al control de la enfermedad, existen dos tipos de pruebas para diagnosticar tanto la infección como la enfermedad. En el primer caso, para detectar la tuberculosis latente existen pruebas de inmunodiagnóstico como la de la tuberculina (TST) o el interferón gamma (IGRA) que son baratos y accesibles para casi todos los países. El problema que presentan es la generación de falsos positivos ya que son capaces de detectar antígenos específicos que pueden estar presentes en el paciente sin presentar la enfermedad. En el segundo caso, la detección de la tuberculosis activa se basa en la sintomatología de los pacientes, la identificación de lesiones en los pulmones mediante radiografías, en técnicas de confirmación como el cultivo de esputos, que la que más se usa, técnicas moleculares como la PCR a tiempo real y técnicas microscópicas en las que se llevan a cabo tinciones. Para la prevención existe una única vacuna atenuada derivada de *M. bovis*, bacillus Calmette-Guerin o BCG. Esta vacuna es bastante inespecífica y la suelen usar algunos países para vacunar a los recién nacidos y protegerlos de la TB extrapulmonar (Copin et al., 2014) en cambio, no es efectiva en el caso de los adultos (Delogu, Manganelli and Brennan, 2014). Después de tantos años de investigación, aún no se conocen los mecanismos del sistema inmune que pueden proteger contra la TB por lo que no se ha

podido establecer una vacuna efectiva contra la enfermedad (Ernst, 2012). Por otro lado, el tratamiento de la tuberculosis activa dura 6 meses y consiste en una combinación de cuatro fármacos que son la piracinamida y etambutol los primeros dos meses y la isoniacida y rifampicina durante los últimos cuatro meses. Este tratamiento sirve en el caso de que los pacientes no estén infectados por cepas resistentes a alguno de estos antibióticos. En el caso de que las cepas sean multirresistentes (MDR) (se conocen como MDR las cepas que presentan resistencia a isoniacida y rifampicina a la vez) el tratamiento es más largo y se utilizan otro tipo de antibióticos, conocidos como fármacos de segunda línea, como las fuoroquinolonas junto a fármacos inyectables. Si las cepas son resistentes a los fármacos de segunda línea (XDR) el tratamiento presenta mayores dificultades y las probabilidades de que aparezcan efectos secundarios aumentan. Además, los tratamientos para los pacientes con cepas MDR o XDR son muy caros y muchos países no los pueden utilizar. No obstante, en la actualidad se están investigando nuevas técnicas de diagnóstico, sobre todo a nivel molecular, prevención y tratamiento (Dheda, Barry and Maartens, 2016; Internal Clinical Guidelines Team (UK)., 2016).

1.2 Incidencia en África.

La región africana presenta una proporción significativa de casos de pacientes infectados con TB. De todos los casos de tuberculosis existentes, 2.7 millones de personas afectadas viven en países Africanos pero, existe un gran número de casos que aún están por diagnosticar (WHO, 2017). Además, tras varios estudios se cree que el origen de la enfermedad está en esta región y se ha ido expandiendo por el mundo debido a las migraciones de las personas (Gagneux *et al.*, 2006).

1.3 Técnicas de genotipado.

Para poder comprender mejor la biología, epidemiología y evolución del MTBC, es necesario genotipar las cepas, es decir, caracterizarlas y clasificarlas en linajes (Comas and Gagneux, 2009). Las técnicas de genotipado se pueden aplicar tanto para epidemiología como para estudios evolutivos. Por un lado, el uso de las técnicas de genotipado en estudios epidemiológicos, permiten identificar cadenas de transmisión de la enfermedad y diferenciar los casos en los que se produce una recaída o una reinfección con una cepa distinta. Por otro lado, en estudios evolutivos el genotipado de cepas permite determinar su historia evolutiva, su distribución geográfica y la asociación entre el genotipo y el fenotipo, es decir, la relación existente entre la variación entre las cepas y la manifestación clínica de la enfermedad (Gagneux *et al.*, 2006).

A lo largo del tiempo se han empleado distintos métodos para el genotipado de cepas de MTBC. Para ello, se ha tenido en cuenta que las secuencias de ADN de las bacterias monomórficas presenta una tasa de variabilidad baja (Comas *et al.*, 2009).

Los primeros métodos de genotipado que se usaron fueron técnicas basadas en geles, como la Pulsed-Field Gel Electrophoresis (PFGE) y los RFLPs (*Restriction Fragment Length Polymorphisms*). El problema de estas técnicas era que, a pesar de ser muy útiles en estudios epidemiológicos, para usarlas en estudios filogenéticos se requerían grandes cantidades de ADN de buena calidad y eran poco reproducibles entre laboratorios (Achtman, 2008; Coscolla and Gagneux, 2014).

Debido a estas limitaciones, se empezaron a usar las técnicas basadas en la reacción en cadena de la polimerasa (PCR) ya que, una de sus ventajas era que requerían poca cantidad de ADN. Las dos técnicas principales eran el tipado de espoligos, basado en CRISPR o repeticiones palindrómicas cortas regulatorias agrupadas (*Clustered Regulatory Short Palindromic Repeats),* y el tipado de MIRUs (*Mycobacterial Interspersed Repeat Units*) que consiste en el número variable de repeticiones en tándem (VNTRs). En este caso, las limitaciones de estas dos técnicas se encontraban en su tendencia a generar convergencias evolutivas y, por ello, en sus dificultades para clasificar las cepas en linajes filogenéticos de forma robusta (Comas *et al.*, 2009).

Para llevar a cabo estudios filogenéticos, los marcadores que se deben de utilizar tienen que ser únicos e irreversibles (Gagneux and Small, 2007). De este modo, para solventar el problema de las técnicas de tipado basadas en la PCR, se desarrollaron nuevos sistemas basados en marcadores más robustos como los LSPs (deleciones genómicas o de polimorfismos de secuencia larga) o SNPs (polimorfismos de nucleótidos únicos). Los LSPs se consideraron buenos marcadores para la clasificación de las cepas en linajes y sublinajes debido a que en MTBC existe una probabilidad muy baja de que se produzca la transferencia horizontal de genes. Por tanto, al considerarse eventos únicos en la evolución de MTBC, permitían construir filogenias robustas (Gagneux *et al.*, 2006; Comas *et al.*, 2009; Coscolla and Gagneux, 2014).

A pesar de la gran robustez de los LSPs como marcadores filogenéticos, su principal limitación se encontraba cuando se querían comparar cepas ya que, no eran capaces de ofrecer información sobre la diversidad existente entre los linajes a los que pertenecían, es decir, no reflejaban las distancias genéticas, por lo que inferir la evolución molecular de MTBC a partir de ellos resultaba una tarea difícil. Además, eran dependientes de la disponibilidad de material genético de una cepa de referencia (Filliol, Motiwala and Cavatore, 2006; Hershberg *et al.*, 2008; Comas and Gagneux, 2009). Por otro lado, se ha visto que el análisis basado en SNPs tiende menos a ser distorsionado por la presión selectiva, por lo que se propuso como método de genotipado (Filliol, Motiwala and Cavatore, 2006). Los SNPs son marcadores ideales para el genotipado de MTBC ya que representan eventos evolutivos únicos y no muestran homoplasias (Gagneux and Small, 2007; Comas *et al.*, 2009).

Más adelante, se hizo la secuenciación del genoma completo de 21 cepas representativas de los seis linajes principales. En estos estudios, se compararon las secuencias obtenidas con la secuencia de referencia de *M. tuberculosis* y se identificaron unos 9.037 SNPs a partir de los cuales se pudieron encontrar SNPs específicos de cada uno de los linajes (Comas *et al.*, 2010; Pérez-Lago *et al.*, 2015).

A partir de este momento, se han ido diseñando nuevos métodos basados en el uso de SNPs específicos de linaje como marcadores para el genotipado de cepas de MTBC, algunos de ellos basados en PCRs específicas de alelo (Stucki *et al.*, 2012; Pérez-Lago *et al.*, 2015).

1.4 Linajes.

A través de las técnicas de genotipado comentadas anteriormente, junto a las nuevas tecnologías de secuenciación masiva (NGS), se han podido identificar distintas especies y subespecies del MTBC.

Las especies y subespecies que componen el MTBC comparten un 99% de identidad entre sus secuencias de ADN, pero, se diferencian en el fenotipo, el rango de hospedadores que abarcan y en el grado de patogenicidad. Algunas de ellas son específicas de humanos *(Mycobacterium tuberculosis y Mycobacterium africanum),* otras afectan a roedores (*Mycobacterium microti*), mientras que otras presentan un espectro más amplio de hospedadores (*Mycobacterium bovis*). Se ha visto que las cepas más epidémicas, es decir, las más abundantes a nivel mundial, comparten una deleción (TbD1) que no se encuentra en *M. africanum* ni en *M. bovis* (Brosch *et al.*, 2002), mientras que estas dos últimas comparten una deleción distinta (RD9). De este modo se dividen en cepas 'modernas' y en 'antiguas' respectivamente (Brosch *et al.*, 2002; Smith *et al.*, 2006; Brites and Gagneux, 2015).

Posteriormente, a partir de la secuenciación de genomas completos se ha podido establecer una clasificación de las cepas de MTBC en 7 linajes filogenéticos principales que provienen de un ancestro común (**Figura 1**) y se encuentran en distintas regiones geográficas del mundo (Brites and Gagneux, 2015; Barbier and Wirth, 2016). Algunos de los linajes están repartidos por todo el mundo mientras que otros son más específicos de alguna región concreta (**Figura 2**). Por un lado, los linajes clasificados como 'modernos' son: el Linaje 4, también conocido como Euro-Americano, que abarca los países de Europa, América y África; el Linaje 2 que se distribuye mayoritariamente en los países del este de Asia y contiene la familia Beijing, y el Linaje 3 que se encuentra sobre todo en el este de África y en el centro y el sur de Asia. Por el otro lado, junto a los linajes que se han adaptado a los animales, se encuentran los linajes 'antiguos' que son: el Linaje 1, también conocido como Indo-Oceánico, que solo se encuentra en el océano Índico y en Filipinas, los Linajes 5 (*M. africanum* West African 1) y 6 (*M. africanum* West African 2) que están restringidos a los países del oeste de África, y el Linaje 7 que es específico de Etiopía (Gagneux *et al.*, 2006; Hershberg *et al.*, 2008; Comas *et al.*, 2013; Firdessa *et al.*, 2013; Barbier and Wirth, 2016).

Sin embargo, por alguna razón desconocida, el oeste de África es la única región del mundo en la que están presentes los seis linajes mayoritarios de MTBC (Gagneux *et al.*, 2006; Gehre *et al.*, 2016).

Esta distribución geográfica de los linajes podría haber tenido lugar debido al aumento de la población a nivel mundial, las migraciones, la urbanización y a las características genéticas de *M. tuberculosis* (Hershberg *et al.*, 2008).

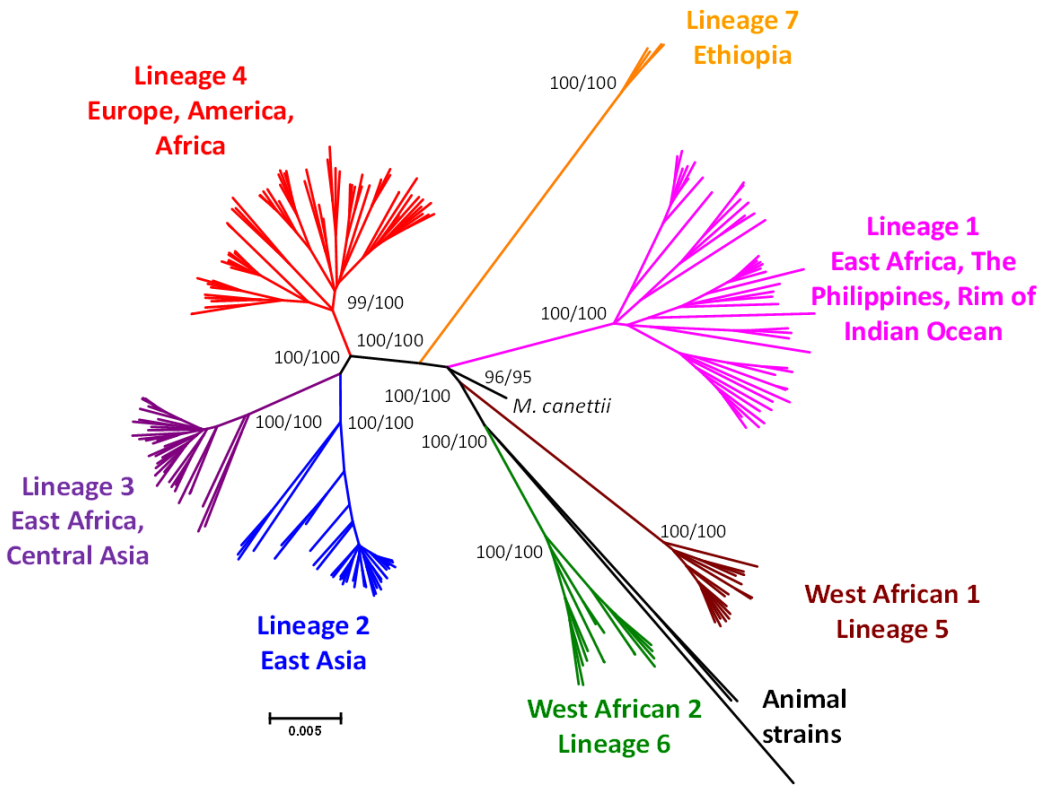


Figura 1: Filogenia del genoma completo de 220 cepas de MTBC. (Comas et al., 2013).

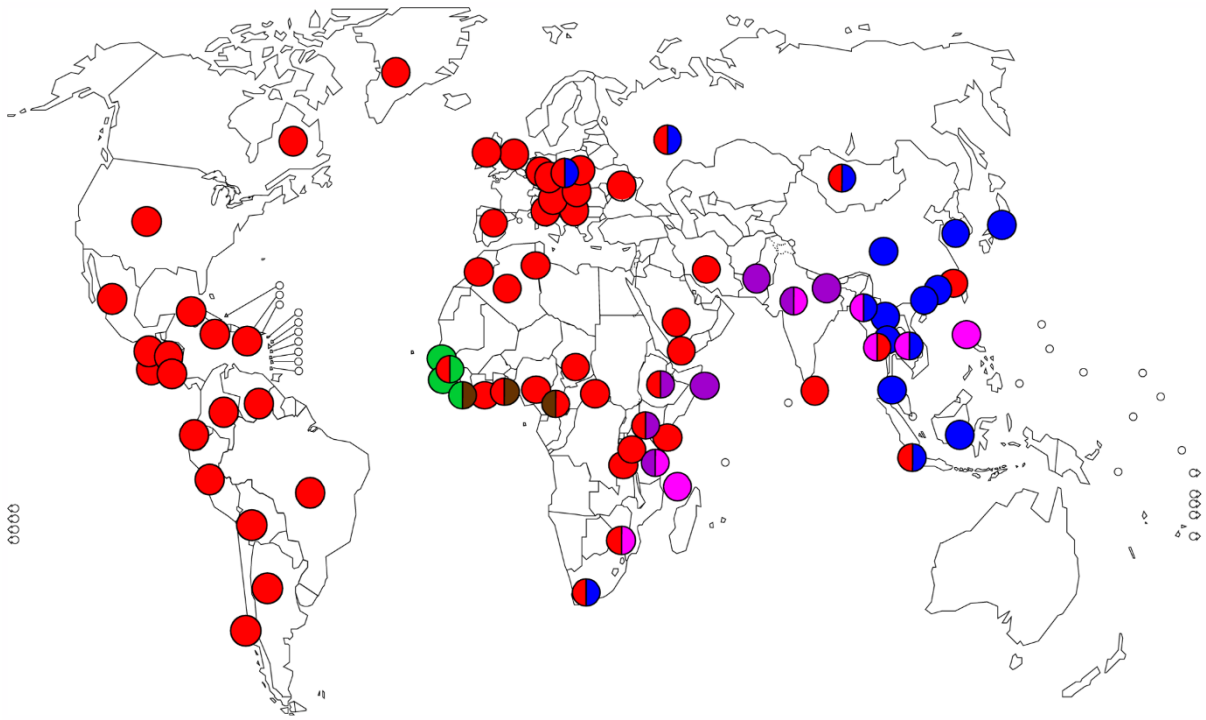


Figura 2: Distribución geográfica de los 6 linajes predominantes. Cada punto corresponde a uno de los ochenta países que se encuentran en la colección global de cepas. Los colores de los puntos corresponden a los de la **Figura 1** e indican el linaje

más abundante de cada país. En esta imagen no se incluye el Linaje 7 ya que cuando se publicó, este linaje aún no había sido descubierto. Imagen adaptada del artículo de Comas et al. 2009.

1.5 Sublinajes del Linaje 4.

De todos los linajes en los que se clasifican las cepas de MTBC, el Linaje 4 es el más abundante a nivel mundial pero, algunos estudios de epidemiología molecular han mostrado una variación considerable en el proceso de transmisión de este linaje (Coscolla & Gagneux, 2014). De este modo, se ha deducido que su diversidad genética y fenotípica puede ser la responsable de la epidemiología de los distintos subtipos del Linaje 4 en distintas partes del mundo (Stucki *et al.*, 2016).

Las cepas del Linaje 4 se dividen en diez sublinajes (**Figura 3**) que se clasifican en tres grupos según su abundancia geográfica. Los sublinajes generales son aquellos que aparecen en un mayor número de países y son L4.1.2/Haarlem, L4.3/LAM y L4.10/PGG3, mientras que los sublinajes específicos se encuentran en una menor cantidad de países y son L4.1.3/Ghana, L4.5/Irán, L4.6.1/Uganda y L4.6.2/Cameroon. Los tres sublinajes restantes son L4.1.1/X, L4.2/Ural y L4.4/Vietnam y se encuentran en una frecuencia intermedia (Stucki *et al.*, 2016).

En cuanto a su distribución geográfica (**Figura 4**), L4.1.3/Ghana, L4.5/Irán, L4.6.1/Uganda y L4.6.2/Cameroon abundan en regiones específicas de África y Asia y están casi completamente ausentes en Europa y América. L4.1.1/X se encuentra con mayor frecuencia en América y en bajas proporciones en algunos países del sur de África, Asia y Europa, mientras que L4.2/Ural y L4.4/Vietnam son muy frecuentes en África y Asia pero no suelen aparecer en América (Stucki *et al.*, 2016).

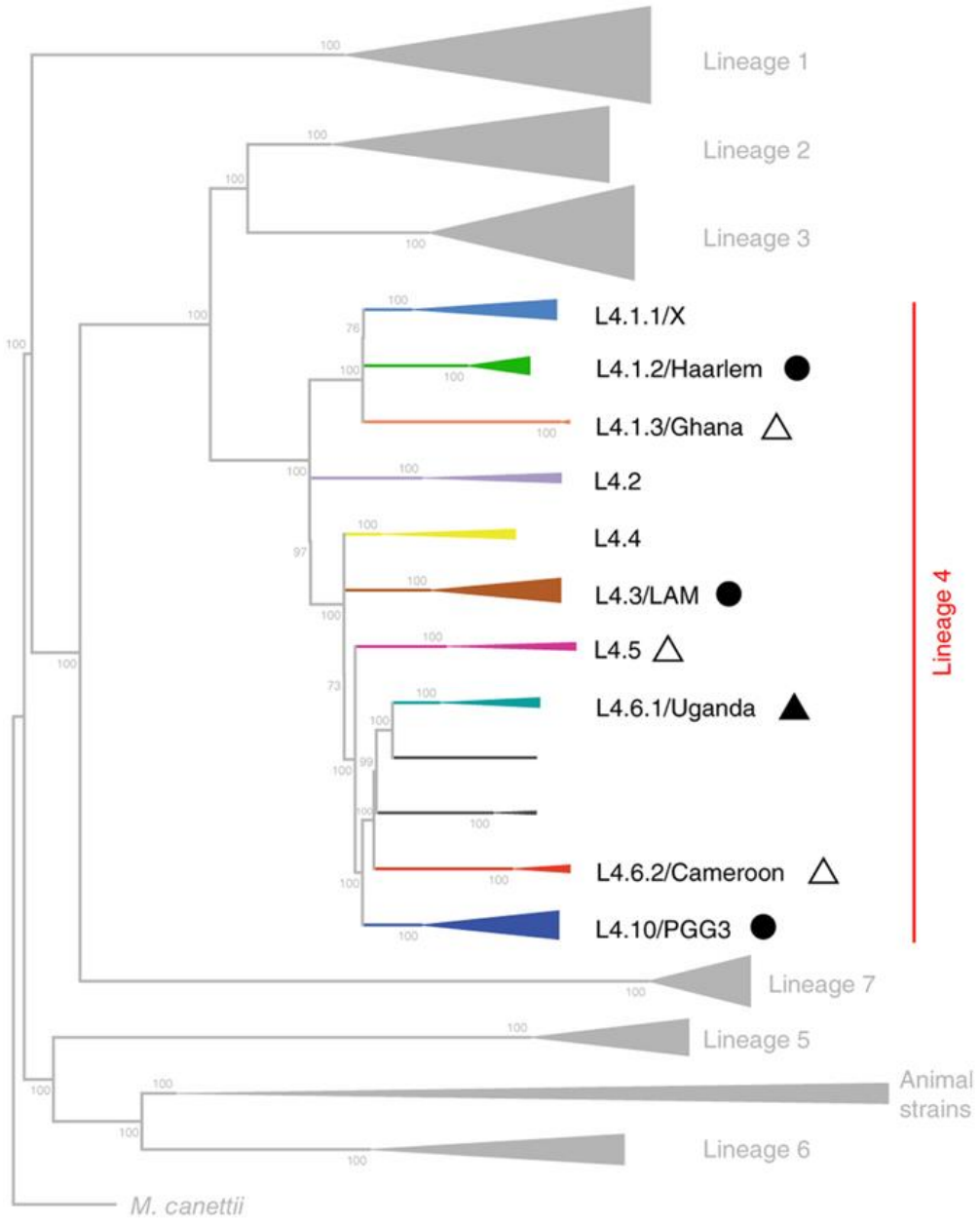


Figura 3: Filogenia de los diez sublinajes del linaje 4. Los círculos representan los sublinajes generales y los triángulos representan los específicos. Figura adaptada de Stucki et al. (Stucki et al., 2016).

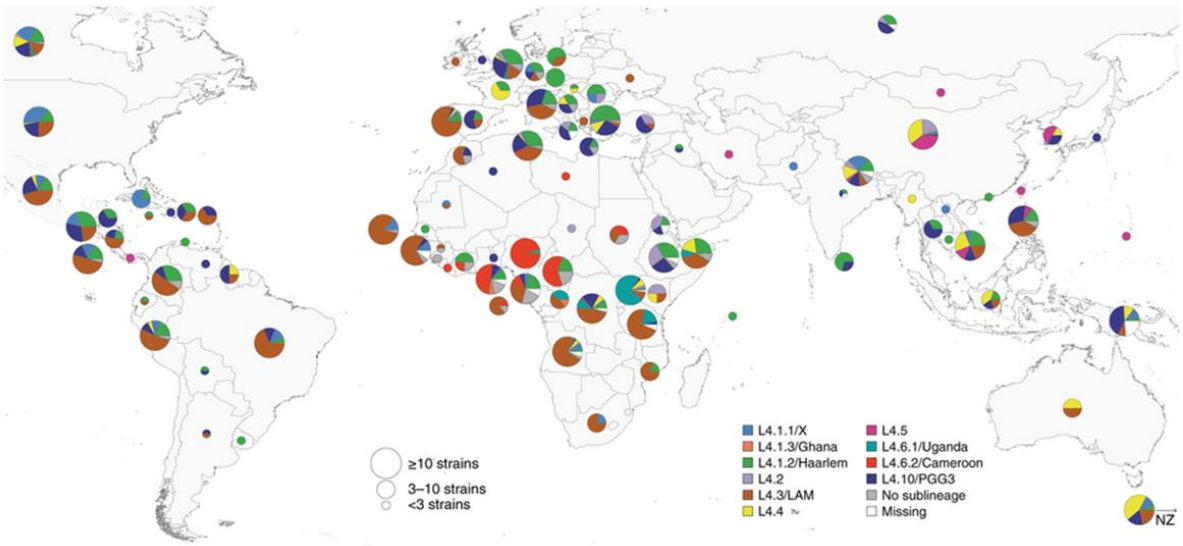


Figura 4: Distribución global de los sublinajes del Linaje 4. Los diagramas de tarta muestran las frecuencias de los sublinajes del linaje 4 siguiendo el código de colores de la **Figura 3**. Figura adaptada de Stucki et al. (Stucki et al., 2016)

1.6 PCR a tiempo real seguida de un ensayo de HRM: Determinación del linaje y sublinaje.

Las estrategias más utilizadas para determinar el linaje de una determinada cepa se realizan a partir de los productos de amplificación de la PCR a tiempo real mediante la técnica molecular del análisis de las curvas de fusión o HRM (*High Resolution Melting*) (Wampande *et al.*, 2015). El ensayo del HRM se caracteriza por ser un método simple y de bajo coste que proporciona resultados de forma rápida y sin la necesidad de realizar un paso de separación por electroforesis (Issa *et al.*, 2014). Además, presenta una resolución tan elevada que es capaz de diferenciar dos moléculas de ADN que se diferencian tan solo en una base dando lugar a dos curvas de fusión distintas. Por tanto, el HRM posibilita la determinación de la ausencia o presencia de un determinado SNP y la discriminación de dos muestras de ADN distintas pertenecientes a distintos linajes basándose en su temperatura de fusión (Tm). Para la monitorización de la Tm se suelen llevar dos estrategias, una es el uso de fluoróforos inespecíficos (SYBR Green® o EvaGreen®) que se intercalan entre las cadenas del ADN, y la otra consiste en el uso de sondas marcadas con fluorescencia complementarias a la región de interés.

Después de realizar un primer paso de amplificación mediante qPCR, se realiza el análisis de las curvas de fusión de alta resolución a partir de los productos de amplificación de la qPCR. El análisis de las curvas de fusión es el que permite determinar la Tm a partir de las medidas de fluorescencia teniendo en cuenta que el pico más alto de fluorescencia se produce en el momento en el que todo el ADN está en forma de doble cadena (dsADN), cuando la temperatura empieza a aumentar, las dsADN se separan liberando el fluoróforo y provocando una disminución de la fluorescencia. De este modo, la Tm del

ADN de la muestra se define como la temperatura a la cual se encuentra en la muestra el 50% del ADN en forma de dsADN y la otra mitad en forma de cadena simple (ssADN) (KAPABiosystems, 2017).

1.7 Co-infecciones.

Se tiende a pensar que la TB se produce por la infección de una sola cepa de *MTBC* pero, en realidad, se ha demostrado que un solo paciente puede haberse infectado por varias cepas genéticamente diferentes. Este suceso se conoce como co-infección (Mallard *et al.*, 2010; Gan *et al.*, 2016). Además, las co-infecciones se suelen dar en más casos de los esperados y se suelen producir con mayor frecuencia por cepas de sublinajes distintos (Warren *et al.*, 2004). Para poder detectar las distintas cepas que pueden estar presentes en el huésped, se suelen utilizar técnicas de genotipado basadas en la PCR (Mallard *et al.*, 2010). Hasta ahora, en estudios epidemiológicos, se han utilizado marcadores como polimorfismos de inserción específicos de cepa (como el IS6110) (Fomukong *et al.*, 1998), deleciones de cromosomas (Ho *et al.*, 2000), VNTRs y tipado de espoligos (Warren *et al.*, 2004).

Por otro lado, debido a la baja sensibilidad y resolución de algunas de estas técnicas, también se han desarrollado otros métodos como la secuenciación de genomas completos basada en la tecnología de secuenciación de última generación (NGS), que ha resultado ser una técnica discriminatoria y de alta sensibilidad para este tipo de estudios (Gan *et al.*, 2016).

A nivel epidemiológico, las co-infecciones pueden causar problemas clínicos y de salud pública dando lugar a una discordancia en los perfiles de las pruebas de susceptibilidad a fármacos ya que, en algunos casos, pueden coexistir cepas resistentes junto a otras sensibles y esto dificulta el tratamiento. Las frecuencias de las co-infecciones pueden diferir dependiendo del nivel de transmisión de MTBC que presenta cada región (Gan *et al.*, 2016).

1.8 Secuenciación masiva.

A pesar de que la técnica de genotipado mediante SNPs sea bastante precisa y permita tener una primera visión del genotipo de las cepas de MTBC de una determinada región, en diversas ocasiones genera resultados que no están del todo claros. En estos casos, para comprobar si el tipado se ha realizado de forma correcta, se pueden usar las técnicas de secuenciación masiva. Actualmente, el desarrollo de las NGS ha permitido poder analizar el genoma completo de las células a una resolución bastante elevada, no obstante, estas nuevas tecnologías presentan algunas limitaciones como el coste, el almacenamiento de los datos y su posterior análisis bioinformático (WJ, 2015).

1.9 Liberia.

Las técnicas mencionadas anteriormente han permitido establecer una distribución de los linajes y sublinajes en una gran parte del mundo. A pesar de ello, aún existen países en los que no se ha realizado ningún estudio. Liberia es uno de los países del oeste de África de los que no se tiene ningún tipo de información acerca de la distribución de cepas de MTBC (Gehre *et al.*, 2016). Por ello, uno de los objetivos de este trabajo es caracterizar las cepas de tuberculosis procedentes de aislados clínicos de Liberia utilizando la técnica de genotipado mediante SNPs.

2. OBJETIVOS.

Los objetivos del siguiente trabajo son identificar los linajes a los que pertenecen las cepas procedentes de muestras de aislados clínicos de Liberia utilizando la técnica del tipado con SNPs y, además, diseñar un experimento para la detección de co-infecciones. Para ambos ensayos se utilizará la técnica de la PCR a tiempo real seguida del análisis de las curvas de fusión. Además, en el caso de que se encuentren resultados anómalos o inesperados, se realizará la secuenciación genómica de las muestras mediante técnicas de secuenciación masiva de última generación (NGS).

3. MATERIALES Y MÉTODOS.

3.1 Muestras biológicas, ADNs:

Se analizaron un total de 81 muestras procedentes de inactivados de cultivo de aislados clínicos de pacientes de Liberia con TB. A pesar de que las muestras ya habían sido inactivadas en Liberia, se decidió volverlas a inactivar para asegurar que no hubiera ningún microorganismo vivo que pudiera provocar la enfermedad. La inactivación se hizo por calor sometiendo a las muestras a 95ºC durante 10 minutos y luego centrifugándolas y recogiendo el sobrenadante. Todo este proceso se llevó a cabo en el laboratorio de riesgo biológico de nivel 3. El hecho de que los ADNs fueran de inactivados clínicos podría implicar la presencia de impurezas que posteriormente interfirieran en las qPCRs. No obstante, se decidió no realizar una purificación de los mismos ya que se corría el riesgo de que al partir de bajas concentraciones de ADN este se perdiera durante el proceso de purificación.

Los ADNs utilizados como control provenían de una colección de muestras clínicas de MTBC de hospitales de la Comunitat Valenciana y del Instituto de Salud Pública y Tropical de Suiza. Las muestras de referencia fueron caracterizadas en estudios anteriores mediante la técnica de secuenciación de genomas completos, por lo que se conocía el linaje y sublinaje al que pertenecían.

3.2 Cuantificación de las muestras:

La concentración de las muestras fue determinada por fluorescencia utilizando el Qubit®. Qubit® es una tecnología basada en un fluoróforo que se intercala entre las hebras del ADN de doble cadena (dsADN) y permite determinar la concentración de ADN de la muestra a una elevada sensibilidad (Thermo Fisher Scientific, 2017). Para llevar a cabo la cuantificación, se siguió el protocolo propio del Qubit® que consiste en la preparación de un mix compuesto por 199µL de la disolución tampón y 1µL del reactivo (fluoróforo), por muestra. Luego se añaden 2µL de la muestra a 198µL del mix del tampón con el fluoróforo, se incuba durante 2 minutos, y se mide la concentración de dsADN en el fluorímetro Qubit®.

Los ADNs que se usaron como control también se cuantificaron y luego se normalizaron a una concentración de 2ng/µL con Tris 10mM a pH=8 para que fuera similar a las de las muestras de Liberia y de este modo la amplificación por PCR resultara homogénea.

3.3 Genotipado basado en SNPs mediante PCR-HRM a tiempo real:

La determinación del linaje de las muestras se realizó mediante una amplificación de las regiones del genoma de MTBC que contienen el SNP característico de cada linaje y para ello se llevó a cabo una PCR a tiempo real seguida de un ensayo HRM. Para las PCRs a tiempo real se usó el kit KAPA HRM FAST qPCR® de Kapa biosystems que incluye el cloruro de magnesio ($MgCl_2$) 25mM y el KAPA HRM FAST qPCR Master Mix 2X ® (KAPA mix) que contiene la ADN polimerasa KAPA HRM FAST®, el tampón de reacción, los nucleótidos (dNTPs) y el fluoróforo EvaGreen®. Además, para cada PCR se usaron los cebadores para la amplificación de la región del genoma que presentaba el SNP característico dependiendo del linaje o sublinaje que se iba a testar.

Para facilitar el análisis y reducir el número de ensayos por muestra, se realizaron dos qPCRs multiplex, utilizando los parámetros que habían sido optimizados en estudios anteriores (resultados no publicados). Los linajes multiplexados fueron, por una parte, los linajes 2, 3 y 4, y por otra, los linajes 1 y 6. Los linajes 5 y el *M. Bovis* fueron analizados individualmente. La qPCR se realizó en un volumen total de 10µL que contenía 5µL del KAPA mix 2X, 1µL de $MgCl_2$, 0.2µL del cebador 10mM (reverso y directo) o 0.6µL en el caso del cebador para el L4 y 1µL de ADN molde. La cantidad de agua miliQ fue ajustada hasta llegar al volumen final de 10µL según la PCR. Los parámetros usados para la PCR fueron los siguientes, una primera fase de incubación a 95ºC durante 5 minutos, 50 ciclos de amplificación con un paso de desnaturalización de 10 segundos a 95ºC, un paso de anillado de cebadores de 20 segundos a 57ºC y una extensión de 20 segundos a 72ºC.

A partir de los productos de la amplificación se realizó el paso de fusión (HRM). Se parte del producto de amplificación ya que, en ese punto, la fluorescencia es muy elevada debido a que, al estar el todo el ADN en forma de doble cadena (dsADN), el fluoróforo EvaGreen® se encuentra intercalado entre las dos hebras emitiendo la máxima fluorescencia. Los parámetros utilizados para el HRM constaron de un primer paso de desnaturalización a 97ºC durante 5 minutos seguido de una re-hibridación al bajar la temperatura a 70ºC durante 1 minuto y por último una subida a 97ºC durante 1 segundo. De este modo, al aumentar la temperatura se favorece la separación de las dos cadenas del ADN y se produce una reducción de la fluorescencia debida a la liberación del fluoróforo. Por tanto, en el momento en el que la diferencia de fluorescencia es mayor se determina la Tm del ADN. La forma de las curvas de fusión que se originan en el paso de HRM depende de la presencia o ausencia de un determinado SNP. Por tanto, el linaje al que pertenecen las muestras clínicas se determina al comparar sus curvas de fusión con las de los controles de linaje conocido.

En cada una de las PCRs se incluyeron el control de PCR, sin ADN, controles negativos y controles positivos de muestras de linaje conocido pertenecientes al linaje que se iba a testar y se hicieron dos réplicas de cada muestra. Los controles negativos eran muestras con ADN en el que estaban ausentes los SNPs de los linajes que se iban a testar.

La detección de la fluorescencia se llevó a cabo en el termociclador LightCycler 96® de Roche en una placa de 96 pocillos.

El análisis de las curvas de fusión se llevó a cabo mediante el software LightCycler 96® de Roche considerando que las muestras eran positivas para un determinado linaje si seguían la misma curva de fusión que las muestras control del determinado linaje. Este programa genera las curvas de fusión y clasifica las muestras automáticamente según la forma de la curva realizando un análisis que consta de cuatro pasos. Para ello, utiliza una serie de algoritmos que intensifican las diferencias de la caída de la fluorescencia entre las curvas y genera un conjunto de gráficos. En primer lugar, se identifican los negativos de PCR que, al no presentar ADN, no emiten fluorescencia. Luego, se realiza una normalización de las curvas de fusión seleccionando manualmente una región de la curva que comprende la Tm del ADN del linaje que se está testando, es decir, se selecciona una región que incluya la curva de caída de la fluorescencia durante la fusión determinando unos valores de temperatura anteriores y posteriores a la temperatura de fusión. Además, se selecciona como umbral de fluorescencia un punto en cual el ADN se encuentra totalmente desnaturalizado. Después de

modificar todos estos parámetros y de seleccionar como referencia las muestras control que no presentan el SNP del linaje que se está testando, el programa genera el '*difference plot*' o gráfico normalizado de la diferencia de fluorescencia, en el que separa los distintos grupos de muestras según la similitud de sus curvas y su Tm permitiendo de este modo asignar las cepas a su linaje correspondiente.

3.4 Identificación de MTBC mediante qPCR:

La PCR cuantitativa (qPCR) a tiempo real se llevó a cabo en un conjunto de muestras que no dieron positivas para ninguno de los linajes para poder detectar si realmente contenían ADN perteneciente a MTBC. Este paso se debería de haber realizado antes de empezar a genotipar las muestras para confirmar que todas ellas pertenecían al MTBC, pero no se hizo debido a que fueron recibidas desde Liberia asegurando que eran de pacientes que padecían tuberculosis.

Para llevar a cabo la qPCR se utilizó el kit KAPA PROBE FAST qPCR MAster Mix (2X)® de Kapa biosystems que contiene una ADN polimerasa hot start, lo cual significa que la actividad del enzima está mediada por la acción de un anticuerpo que va unido a la propia ADN polimerasa de forma que, la mantiene inactiva hasta el primer ciclo de desnaturalización. De este modo, se evita la formación de productos de PCR inespecíficos que podrían darse en los pasos de preparación de la PCR y, además, aumenta la eficiencia de la reacción. La qPCR se realizó en un volumen total de 20µL en el que se añadió, por cada muestra, 10µL del Kapa Fast Probe master mix 2X® que contiene la ADN Taq polimerasa Kapa hot start®, dNTPs, $MgCl_2$ y estabilizadores, 2µL de una disolución con los cebadores reverso y directo a 10µM cada uno, 0.6µL de la sonda fluorescente a 10µM, 3.4µL de agua miliQ y 4µL de ADN de las muestras. Además, se añadieron 5 estándares, muestras de concentración de ADN conocida (1ng/µL, $1 \cdot E^{-1}$ng/µL, $1 \cdot E^{-2}$ng/µL, $1 \cdot E^{-3}$ng/µL y $1 \cdot E^{-4}$ng/µL para hacer la recta patrón. Los cebadores utilizados en este caso permitían amplificar una región característica del genoma de MTBC.

Para la detección de la fluorescencia se usó el termociclador LightCycler 96® de Roche. Se llevó a cabo una amplificación de 3 pasos en la que se realizó una desnaturalización previa a 95ºC durante 3 minutos para activar el enzima, otra desnaturalización a 95ºC durante 10 segundos, 20 segundos a 60ºC para que tenga lugar la hibridación de los cebadores y finalmente 1 segundo a 72ºC para la extensión y obtención de los resultados.

Los resultados fueron analizados mediante el software LightCycler 96® seleccionando la opción de PCR cuantitativa. El programa se encarga de generar unas curvas de amplificado tanto de los controles estándar como de las muestras y luego, comparando las curvas de las muestras con las curvas de los estándares de concentración conocida, interpola la concentración de ADN de MTBC que hay en las muestras. En este experimento, la qPCR se usó con la finalidad de determinar si el conjunto de muestras que no se habían podido asignar a uno de los 6 linajes, pertenecían al MTBC. Para ello, se consideró que las muestras eran positivas para MTBC si amplificaban, mientras que se tomaron como negativas aquellas que no amplificaban ya que, al no ser MTBC, no presentaban la región reconocida por los cebadores.

3.5 Prueba de detección de Co-infecciones:

Para poder identificar la presencia de co-infecciones en una misma muestra se diseñó un experimento combinando controles de ADNs de cepas de distinto linaje en proporciones conocidas. Se hicieron dos pruebas, una para los linajes 3 y 4, y otra para los linajes 2 y 4 utilizando muestras control de linaje conocido procedentes de la colección de muestras clínicas de hospitales de la Comunitat Valenciana.

Para ello, en primer lugar, se prepararon alícuotas de una concentración de 1ng/µL de cada uno de los controles con tris 10mM. A partir de las alícuotas, se realizaron disoluciones con combinaciones de distintas proporciones de cada uno de los ADNs. Las proporciones que se probaron se muestran en la **Tabla 1**. Se probaron las mismas proporciones para la prueba con los linajes 2 y 4.

Tabla 1: Proporciones de ADNs de los controles de los linajes 3 y 4 probadas en la prueba de detección de co-infecciones.

Control L3	90%	75%	50%	25%	10%
Control L4	10%	25%	50%	75%	90%

A continuación, se realizaron tres PCRs a tiempo real con un paso final de HRM, utilizando las mismas cantidades de reactivos y parámetros descritos anteriormente. Para cada PCR a tiempo real, se utilizaron unos cebadores diferentes. Se realizó una PCR de multiplexado para los linajes 3 y 4, en la que se juntaron los cebadores de los dos linajes que se estaban testando, para poder determinar a partir de qué concentración era posible detectar dos ADNs de cepas de linaje distinto presentes en una misma muestra. Por otro lado, en las dos PCRs restantes se testaron los linajes 3 y 4 por separado. Se realizaron los mismos pasos para los linajes 2 y 4.

Las reacciones de PCR se llevaron a cabo en una placa de 96 pocillos en el termociclador LightCycler 96® de Roche y el análisis de las curvas de fusión se hizo con el software LightCycler 96® del mismo modo que se ha explicado anteriormente.

3.6 Secuenciación genómica mediante NGS:

Para la secuenciación genómica se utilizó la plataforma MiSeq de Illumina. En este ensayo, en primer lugar, se tienen que preparar las librerías de ADNs genómicos primer paso consiste en la generación de librerías. Para ello se siguió el protocolo Nextera® XT de Illumina. La preparación de librerías consta de 5 pasos.

El primero es la fragmentación del ADN mediante la acción de una transposasa, que a la vez que corta, añade parte de los adaptadores necesarios para la secuenciación. Antes de la fragmentación, se requiere un paso previo de cuantificación y normalización del ADN genómico a 0,5ng/µL. Posteriormente, se mezclan 5µL del ADN genómico (0.5ng/µL por muestra) con 10µL de tampón de fragmentación de ADN (TD), a esta mezcla se le añaden 5µL de la transposasa (amplicon tagment mix® o ATM), se centrifuga 1 minuto a 280g y se introduce en el termociclador a 55ºC durante 5 minutos

para que se produzca la reacción. Luego se añaden 5µL del tampón de neutralización y se incuba a temperatura ambiente durante 5 minutos para detener la actividad del enzima.

El segundo paso consiste en la amplificación del ADN fragmentado. En este paso se terminan de añadir los adaptadores para la obtención de 'clusters'. Para ello, se realizan distintas combinaciones de cebadores para cada una de las librerías. Se añaden 15µL de Nextera PCR Master Mix® y 5µL de cada índice a las muestras con el ADN fragmentado, se centrifuga y se introduce en el termociclador para llevar a cabo la PCR de indexado. Los parámetros utilizados para la PCR de indexado son:

- 72ºC 3'
- 95ºC 30''
- 12 ciclos de:
    - 95ºC 10''
    - 55ºC 30''
    - 72ºC 30''
    - 72ºC 5'
- 10ºC Enfriamiento

En el tercer paso, se lleva a cabo la limpieza de las librerías amplificadas para seleccionar el tamaño de los fragmentos de ADN y eliminar los fragmentos cortos. Para ello, se usan AMPure XP beads ®, 'beads' magnéticas, se añaden 30µL de '*beads'* (ratio 0.6X) por muestra en cada pocillo de una placa, y tras centrifugar la placa que contiene las librerías, se añaden 50µL de cada librería a cada pocillo, se agita la placa a 1800rpm durante 2 minutos, se deja a temperatura ambiente 5 minutos más y luego se coloca la placa sobre una placa magnética para que las '*beads'* a las que se habrá unido el ADN del tamaño deseado se separen del sobrenadante. Luego se realizan dos lavados con etanol al 80% descartando el sobrenadante. Tras retirar el etanol residual y dejar secar, se añaden 27,5µL del tampón de resuspensión (tris 10mM, pH=8), se agita 2 minutos a 1800rpm, se incuba a temperatura ambiente otros 2 minutos, se coloca en la placa magnética y en este caso, como el ADN se habrá separado de las 'beads', se recoge el sobrenadante con los fragmentos de ADN de tamaño deseado y se transfiere a una nueva placa. Para comprobar que el tamaño de los fragmentos de ADN era el deseado (500-1000bp), se corrieron las librerías en el Bioanalyzer® de Agilent Technologies en un Chip de ADN de alta sensibilidad (Agilent DNA High Sensitivity Kit) y se cuantificaron mediante el Qubit®. El Bioanalyzer es un tipo de electroforesis automatizada rápida y sencilla que permite cuantificar y obtener el tamaño de los fragmentos de ADN utilizando solo 1µL de muestra.

En el cuarto paso se normalizan las librerías ajustándolas todas a la misma concentración de ADN, a 4nM, a partir del tamaño de los fragmentos y la concentración de ADN.

Finalmente, se prepara el PAL (*pool of amplified libraries*) juntando las librerías normalizadas para la secuenciación. Para ello, se centrifugan las librerías, se transfieren 5µL de cada una de ellas a un mismo eppendorf y se mide la concentración de ADN del PAL para comprobar que está a 4nM.

Una vez preparado el PAL, se procede a la preparación de los ADNs para la secuenciación. Para ello, se realiza una desnaturalización de los ADNs a pH básico añadiendo 5µL de NaOH 0,2N a 5µL del PAL a 4nM durante 5 minutos a temperatura ambiente. Luego se añaden 990µL del tampón de hibridación

HT1 para ajustar la concentración del PAL a 20pM. Finalmente, se diluye la librería a 12,5pM añadiendo 225µL de HT1 a 375µL del PAL, se retiran 18µL de la librería a 12,5pM y se añaden 18µL de phiX para generar diversidad en la secuenciación.

Las lecturas obtenidas fueron analizadas mediante Kraken, un programa bioinformático que utiliza un algoritmo basado en k-meros para determinar la taxonomía de las secuencias de ADN procedentes, normalmente, de estudios metagenómicos (Wood *et al.*, 2014).

4. RESULTADOS Y DISCUSIÓN

4.1 Cuantificación y extracción:

Como se ha comentado anteriormente, se recibieron 81 muestras de aislados clínicos procedentes de Liberia. En primer lugar, se llevó a cabo la cuantificación de dichas muestras y se obtuvieron unas concentraciones muy bajas (**Tabla 2)**.

Al comparar las concentraciones obtenidas con las de los controles, que estaban sobre 10ng/µL, y ver que las muestras estaban 10 veces más diluidas, se decidió ajustar la concentración de los ADNs control a 2ng/µL para que al realizar las PCRs a tiempo real, las amplificaciones fueran homogéneas.

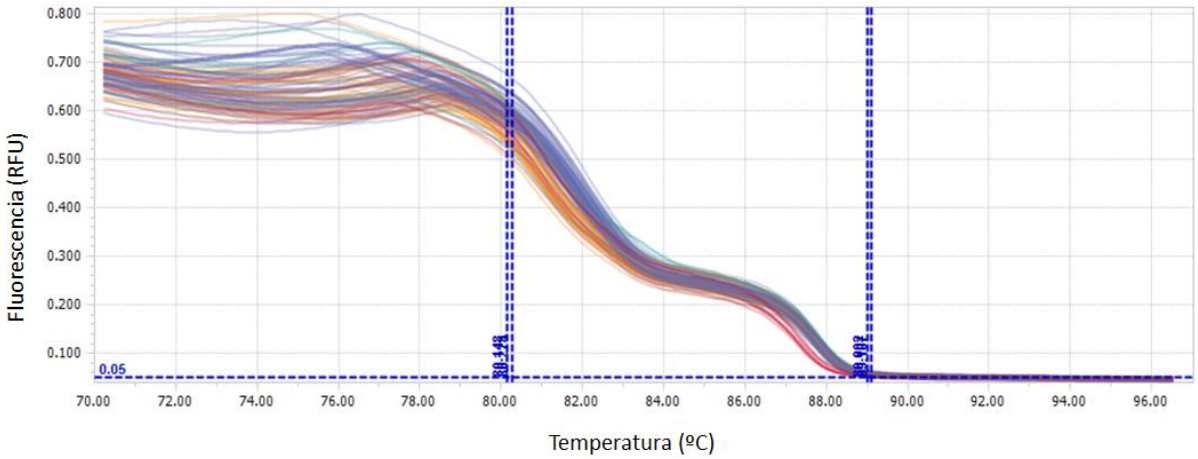
Tabla 2: Concentraciones de DNA de las muestras de Liberia obtenidas en Qubit®.

Muestra	[DNA] ng/uL	Muestra	[DNA] ng/uL	Muestra	[DNA] ng/uL
10D	0,053	**11D**	0,242	**12D**	0,754
32D	0,054	**69D**	0,245	**26D**	0,81
71D	0,054	**78D**	0,253	**76D**	0,819
1D	0,059	**38D**	0,289	**62D**	0,879
40D	0,073	**67D**	0,289	**53D**	1,18
31D	0,074	**54D**	0,340	**79D**	1,22
24D	0,078	**2D**	0,354	**47D**	1,37
81D	0,078	**66D**	0,374	**57D**	1,39
33D	0,086	**49D**	0,387	**43D**	1,51
13D	0,090	**5D**	0,432	**51D**	1,51
28D	0,104	**3D**	0,437	**6D**	1,61
36D	0,105	**73D**	0,438	**56D**	2,28
34D	0,131	**72D**	0,447	**8D**	2,31
22D	0,138	**9D**	0,452	**35D**	2,67
42D	0,146	**80D**	0,454	**64D**	3,27
7D	0,147	**19D**	0,484	**65D**	3,29
77D	0,149	**50D**	0,507	**60D**	4,16
4D	0,167	**45D**	0,514	**29D**	4,24
23D	0,167	**48D**	0,533	**63D**	6,26
17D	0,171	**37D**	0,55	**21D**	7,99
25D	0,172	**61D**	0,612	**58D**	8,08
74D	0,193	**30D**	0,629	**20D**	10,3
44D	0,196	**55D**	0,646	**14D**	< 0.005
27D	0,203	**18D**	0,673	**15D**	< 0.005
75D	0,207	**16D**	0,697	**41D**	< 0.005
52D	0,211	**68D**	0,709	**59D**	< 0.005
39D	0,228	**46D**	0,716	**70D**	< 0.005

4.2 Análisis de las curvas de fusión del HRM:

El análisis de las curvas de fusión genera tres tipos de gráficos que son las curvas de fusión (**Figura 5 A**), los picos de fusión normalizados (**Figura 5 B**) y las curvas normalizadas de diferencia de fluorescencia (**Figura 6**).

A



B

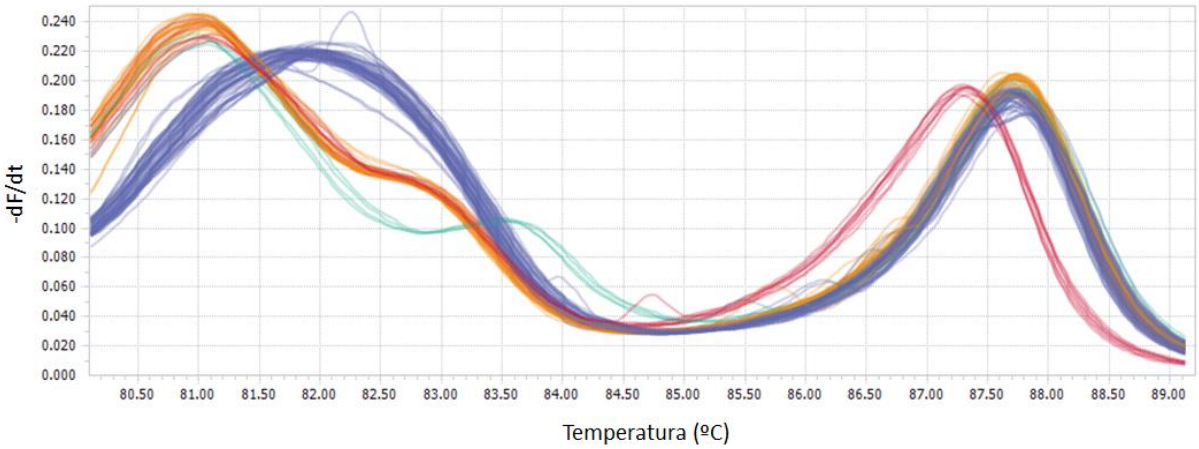


Figura 5: Tipos de gráficos generados en el análisis del HRM. **A**) Representación de las curvas de fusión para los linajes 2, 3 y 4 multiplexados. Las líneas verticales discontinuas simbolizan las temperaturas previas y posteriores a la temperatura de fusión que se seleccionan manualmente para que el software clasifique los ADNs en los linajes correspondientes según la presencia o ausencia del SNP de dicho linaje teniendo en cuenta su Tm. **B)** Representación de los picos de las curvas de fusión normalizadas en el que el eje de las abscisas representa la temperatura y el de las ordenadas muestra la diferencia de la fluorescencia con el tiempo. Se muestran las Tm para los linajes 2, 3 y 4 multiplexados en una PCR a tiempo real. Las curvas moradas representan el linaje 4 y su Tm es de 81ºC, las curvas verdes representan el linaje 3 y su Tm es de 83ºC, las curvas amarillas pertenecen al linaje 2 cuya Tm está alrededor de 87ºC y las curvas rojas corresponden al control negativo.

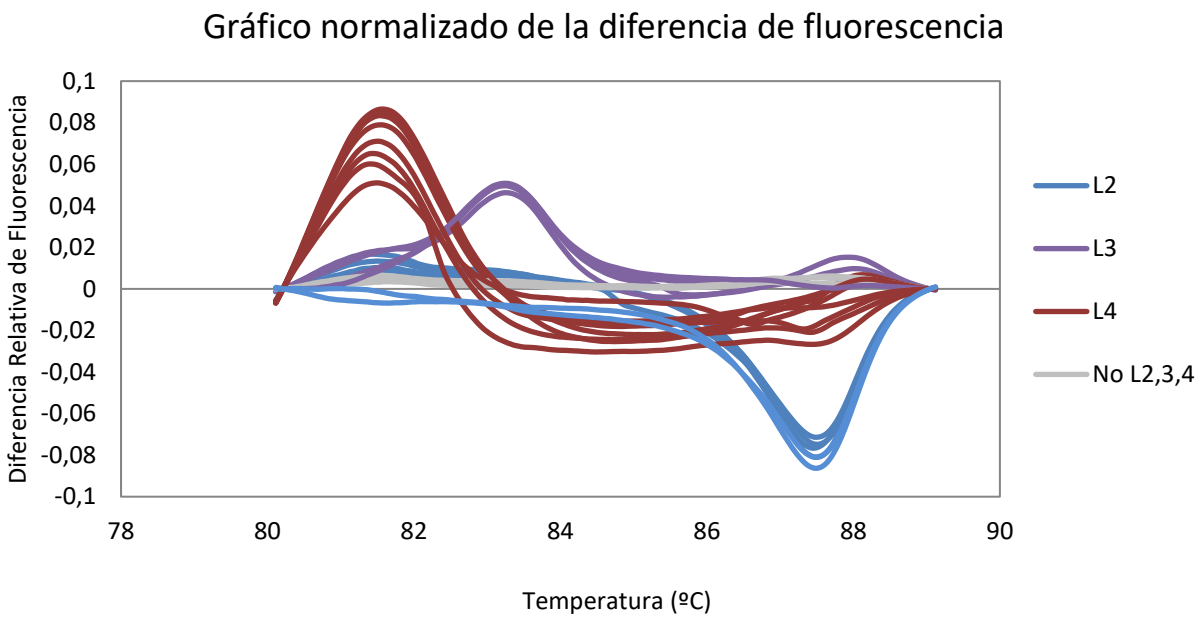
La asignación de cada una de las muestras a su linaje correspondiente se hizo analizando las curvas de diferencia de fluorescencia. En la **Figura 6** se muestran tres ejemplos del resultado de las curvas de diferencia de fluorescencia normalizada cada una de las muestras. Después de seleccionar manualmente la región que engloba la curva de caída de fluorescencia ajustando una temperatura anterior a la Tm más baja de los linajes que se testan en la PCR multiplex y otra posterior a la Tm más alta, se genera una agrupación de las muestras según la forma de las curvas.

De este modo, en la **Figura 6.A** se observa que la mayoría de las muestras pertenecen al linaje 4 ya que sus curvas de fusión siguen el patrón de las curvas de las muestras control (curvas rojas). Por otro lado, solo un grupo pequeño de muestras pertenecen al linaje 2 (curvas azules), y ninguna pertenece al linaje 3 (curvas moradas) ya que solo los controles dieron un pico de fluorescencia en la Tm correspondiente al linaje 3.

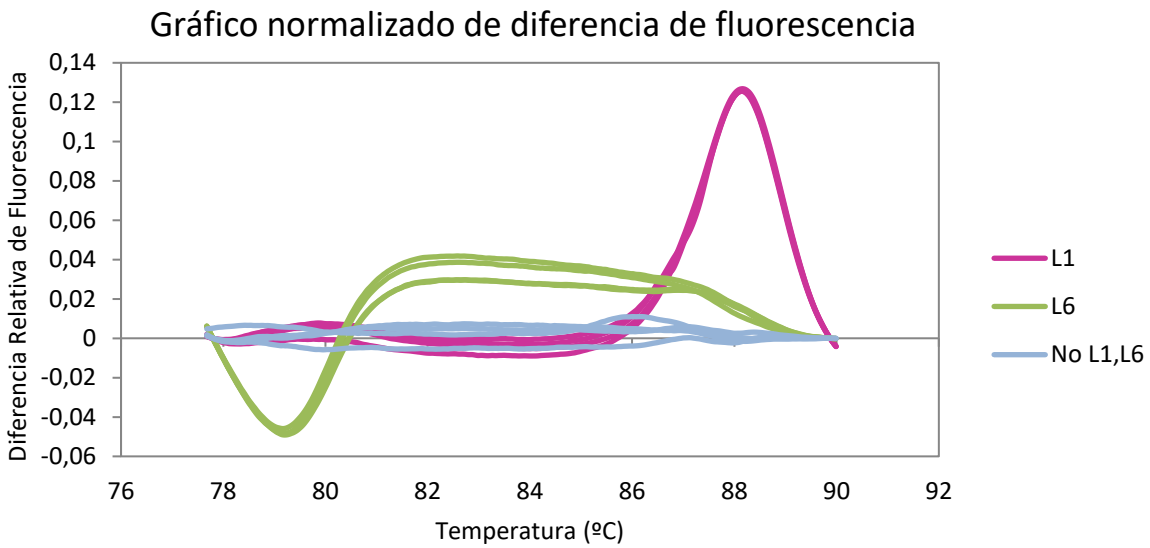
En el caso del ejemplo de PCR de multiplexado de los linajes 1 y 6 (**Figura 6.B**), se observa que algunas de las muestras siguen el patrón de los controles del linaje 6 (curvas verdes) y otras pocas se agrupan con los controles del linaje 1 (curvas rosas).

Por otro lado, las PCRs uniplex, en las que solo se testa un linaje, presentan un análisis mucho más sencillo. Para la normalización de las curvas se ajusta la región seleccionando una temperatura un grado por debajo otra un grado por encima de la Tm. En la **Figura 6.C** se muestra el resultado obtenido para el análisis de las curvas de fusión de la PCR uniplex para el linaje 5. En ella se observa que las dos réplicas de una determinada muestra siguen el patrón de las curvas de los controles del linaje 5, por lo que se consideran positivas para dicho linaje.

A



B



C

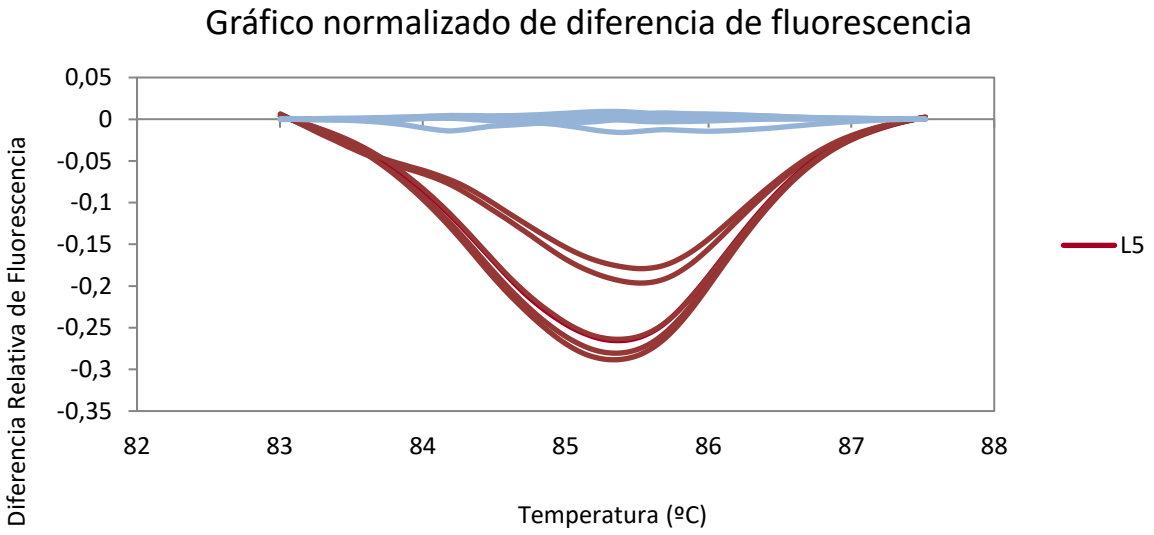


Figura 6: Gráfico normalizado de la diferencia de fluorescencia. **A)** PCR de multiplexado para los linajes 2 (curvas de color azul), 3 (curvas de color morado) y 4 (curvas rojas). **B)** PCR de multiplexado para los linajes 1(curvas rosas) y 6 (curvas verdes). **C)** PCR uniplex para el linaje 5 (curvas rojas).

4.3 Determinación del linaje:

El genotipado de las muestras se empezó realizando una PCR multiplex para los linajes 2, 3 y 4. Para ello, se disponía de unos parámetros puestos a punto en experimentos anteriores (resultados no publicados) para poder detectar las cepas pertenecientes a estos tres linajes en una sola reacción. Se decidió testar estos tres linajes en primer lugar ya que son los más extendidos alrededor del mundo (Stucki *et al.*, 2016) por tanto, el hecho de juntarlos en una misma reacción de PCR permitió determinar el genotipo de la mayoría de las muestras incrementando la rapidez del proceso. Para la PCR de multiplexado, se disponía de unos cebadores para el linaje 4 modificados con una cola poli-AT para aumentar su temperatura de fusión y favorecer, de este modo, la separación de los picos de fusión ya que, en estudios anteriores, se vio que si no se incrementaba la Tm del cebador, no se podían discriminar los SNPs en base a su Tm. Por otro lado, como anteriormente se vio que había una competencia entre los cebadores, se utilizó una cantidad tres veces mayor del cebador para el linaje 4 que para los de los linajes 2 y 3. El siguiente paso fue realizar una PCR multiplex a tiempo real para los linajes 1 y 6 de todas las muestras, para ver si en algún caso había alguna co-infección. En cada una de las PCR multiplex se agruparon los linajes de la forma mencionada anteriormente debido a las Tm características de cada linaje (**Tabla 4**) ya que, cuanto mayor es la diferencia entre las Tm, mejor se diferencian los picos de fusión y por tanto, se facilita la discriminación entre linajes. Las secuencias de los cebadores utilizados se encuentran en el **Anexo I.**

Con estas dos PCR, repitiendo aquellas muestras cuyo resultado no estaba demasiado claro, se obtuvieron 14 cepas pertenecientes al linaje 1, 7 cepas pertenecientes al linaje 2, 54 cepas del linaje 4 y 1 cepa del 6. De este modo, faltaban 5 muestras sin identificar, por lo que se decidió realizar una PCR uniplex para el linaje 5 y otra para el linaje animal *M. Bovis* obteniendo como resultado solo 1 muestra positiva para el linaje 5. Los resultados se muestran en la **Tabla 3**. El linaje 7 no se probó en este experimento debido a que es específico de Etiopía.

Tabla 3: Distribución de la cantidad de cepas pertenecientes a cada linaje.

Linaje	Nº de muestras
L1	14
L2	7
L3	0
L4	54
L5	1
L6	1
BOVIS	0
No identificada	1

4.4 Determinación de los sublinajes para el Linaje 4:

A continuación, se pasó a la determinación de los sublinajes para las muestras del linaje 4 realizando PCRs a tiempo real de cada sublinaje por separado debido a que no había mucha diferencia entre las Tm de los sublinajes (**Tabla 4**). Para ello, se probaron los cebadores para los SNPs de los sublinajes más comunes, sus secuencias se encuentran en el **Anexo II**. De las 54 muestras se obtuvieron 14 muestras LAM (L4.3), 9 muestras Cameroon (L4.6.2), 8 X (L4.1.1), 7 Haarlem (L4.1.2) 3 PGG3 (L4.10) y 3 Ghana (L4.1.3). Las 10 cepas restantes no pudieron ser asignadas a ningún sublinaje. Los resultados se muestran en la **Figura 7.**

Hasta el momento se desconoce la información acerca de la distribución de los linajes y sublinajes en Liberia por lo que se decidió comparar los resultados obtenidos con la distribución de los sublinajes en los países cercanos a Liberia y se vio que la distribución de los sublinajes era diferente. En los países que se encentran en los alrededores de Liberia, como Sierra Leona, Costa de Ivory y Guinea, los dos sublinajes más abundantes suelen ser L4.1.2/Haarlem y L4.3/LAM y L4.1.1/X es uno de los menos predominantes, mientras que en Liberia se observa que algunos sublinajes específicos de otras regiones, como el Cameroon (L4.6.2), son más abundantes que otros sublinajes que normalmente se suelen dar con más frecuencia, como el Haarlem (L4.1.2) o el PGG3 (L4.10), y que el sublinaje L4.1.1/X está entre los tres primeros, cuando normalmente suele ser de los menos abundantes en el sur de África (Gehre *et al.*, 2016; Stucki *et al.*, 2016). Al obtener unas proporciones diferentes a las habituales se ha decidido que se van a secuenciar y de esta forma se podrán caracterizar las cepas cuyo linaje no se ha podido determinar mediante el genotipado con SNPs.

Tabla 4: Temperaturas de fusión de los amplificados de cada linaje y sublinaje.

Linaje	Tm (ºC)
L1	88
L2	87
L3	83
L4	81
L5	85
L6	78
BOVIS	89

Sublinaje	Familia Espoligo	Tm (ºC)
L4.1.1	X	86,5
L4.1.2	Haarlem	84
L4.1.3	Ghana	81,5
L4.2	Ural	86,3
L4.3	LAM	84,5
L4.4	Vietnam	87
L4.5	Iran	84
L4.6.1	Uganda	84,7
L4.6.2	Cameroon	82,3
L4.10	PGG3	79,5

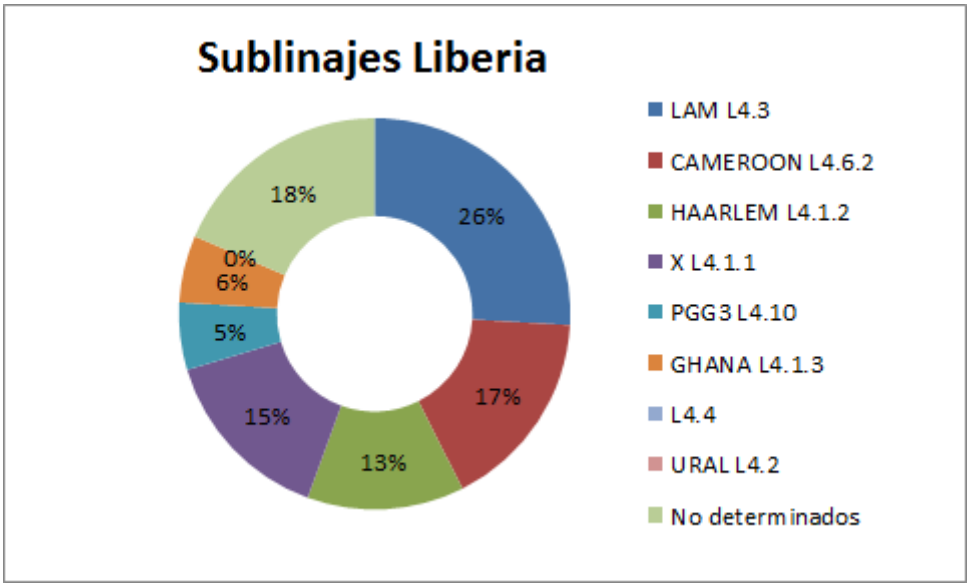


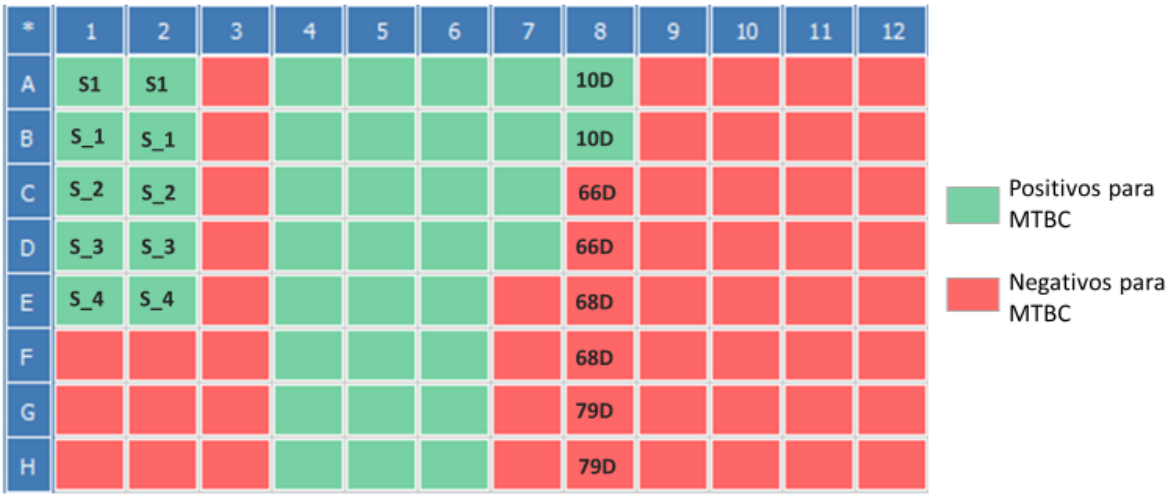
Figura 7: Porcentajes obtenidos de cada uno de los sublinajes.

4.5 Prueba de detección de MTBC:

Para las cuatro muestras restantes que no pudieron asignarse a ningún linaje, se decidió realizar una PCR cuantitativa a tiempo real para comprobar si realmente eran cepas pertenecientes al MTBC o se trataba de otros microorganismos. En este ensayo se usó una sonda fluorescente marcada con FAM en lugar de un fluoróforo como EvaGreen®, por lo que la especificidad del ensayo era mayor.

Como se puede observar en la **Figura 8**, solo una de las muestras (10D) da señal de amplificación por qPCR por lo que se puede decir que es la única que realmente contiene ADN de MTBC. En cambio, en las otras tres muestras no se puede confirmar la presencia del MTBC ya que, el hecho de que no hayan dado ninguna señal de amplificación en la PCR indica que no contienen la región reconocida por los cebadores.

A



B

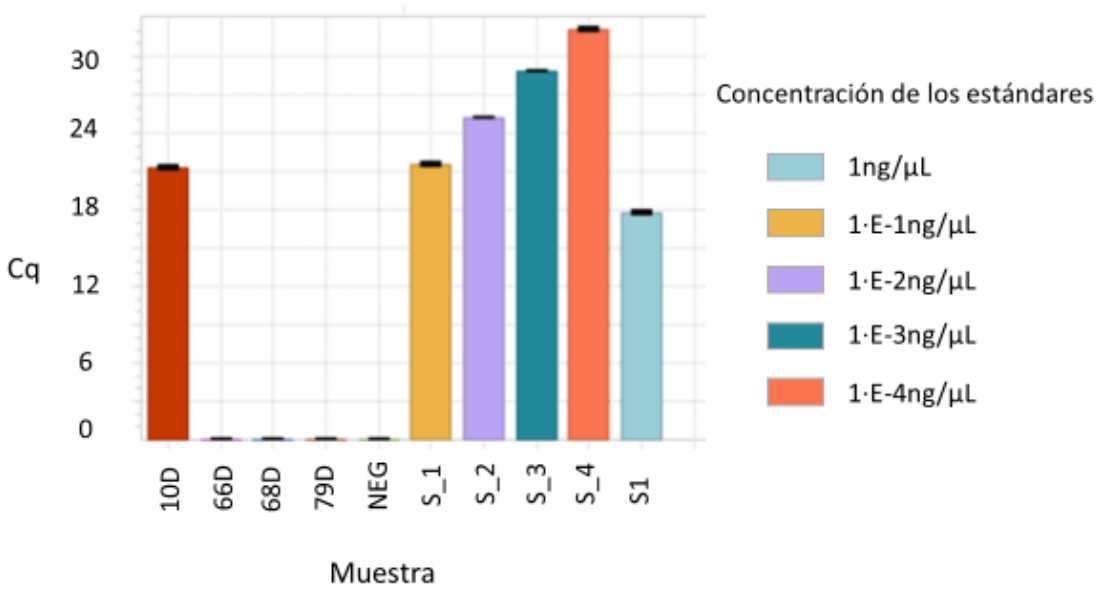


Figura 8: Resultados de la qPCR para identificar MTBC. **A)** 'Heat map': Las filas A-E de las columnas 1 y 2, contienen los controles estándar para la recta patrón con las concentraciones 1ng/µL, $1{\cdot}E^{-1}$ng/µL, $1{\cdot}E^{-2}$ng/µL, $1{\cdot}E^{-3}$ng/µL y $1{\cdot}E^{-4}$ng/µL, de la fila A a la E respectivamente. Las columnas 4, 5, 6 y 7 (A-D) presentan muestras positivas para TB. La columna 8 contiene las muestras cuyo linaje no pudo ser determinado (10D, 66D, 68D y 79D), el resto de celdas corresponden a muestras que no son relevantes en este ensayo. **B)** Representación de los resultados de la qPCR. En el gráfico se representan las muestras en el eje de las abscisas y en el de las ordenadas, el número de ciclos a partir del cual empieza a amplificar cada una de ellas (Cq) que depende de la concentración de ADN de MTBC que presenta la muestra. Las muestras que no han amplificado son las que no presentan MTBC. A la derecha se encuentran los estándares de concentración conocida (1ng/µL, $1{\cdot}E^{-1}$ng/µL, $1{\cdot}E^{-2}$ng/µL, $1{\cdot}E^{-3}$ng/µL y $1{\cdot}E^{-4}$ng/µL) en los cuales se observa que, a mayor concentración de la muestra, menor es la Cq ya que empieza a amplificar a un menor número de ciclos de PCR.

4.6 Detección de co-infecciones:

Para la detección de co-infecciones se diseñó un ensayo en el que se prepararon disoluciones mezclando dos ADNs de linaje conocido en distintas proporciones para intentar estimar la mínima concentración de ADN que podía ser detectada en la técnica de PCR a tiempo real seguida de un análisis de las curvas de fusión. Para ello se realizaron dos pruebas distintas, en una se combinaron los linajes 2 y 4 y en la otra los linajes 3 y 4. Además, en cada una de las pruebas se testaron tanto los linajes juntos combinando los cebadores para ambos en una PCR de multiplexado, como los linajes por separado. Los resultados de las PCRs multiplex a tiempo real seguidas de HRM se muestran en la **Figura 9**.

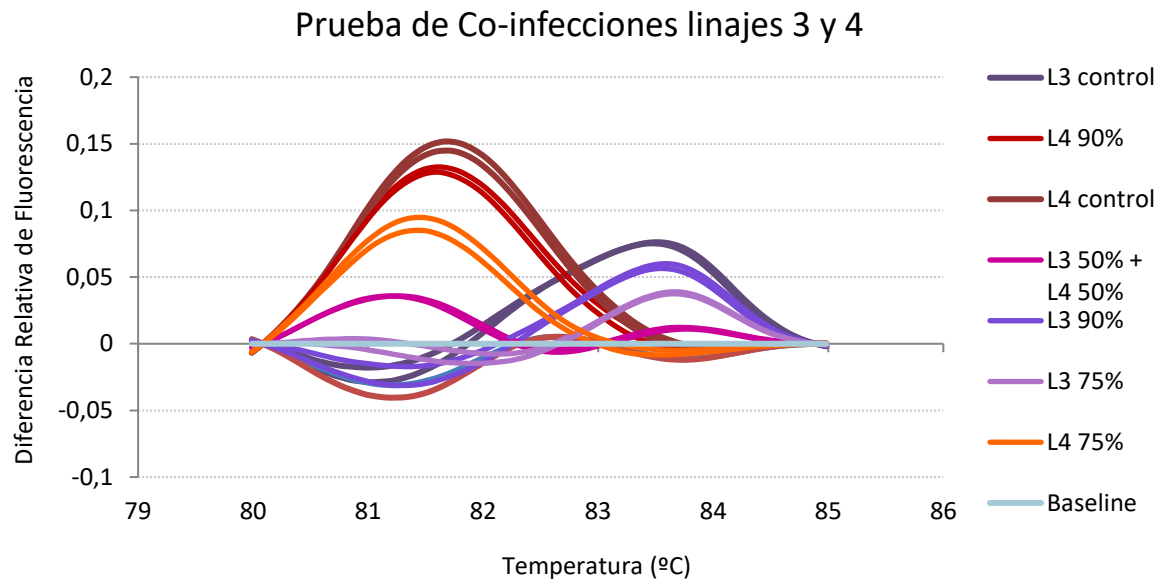
En la **Figura 9.A** se observa que el límite de detección de ADN es 50% en ambos linajes, es decir, cuando la cantidad de un ADN de linaje 4 supera el 50% del total de una muestra coinfectada con otro ADN de linaje 3, se detecta el pico para la Tm del linaje 4 mientras que en el caso contrario, solo aparece el pico correspondiente a la Tm del linaje 3. Cuando ambos ADNs se encuentran al 50%, se observan los picos correspondientes a las Tm de ambos linajes pero a una intensidad muy baja.

Por otro lado, en la **Figura 9.B** se observa que en el caso del linaje 4 se detecta señal de fluorescencia cuando se encuentra a partir del 10%, mientras que en el caso del linaje 2, se observa una señal clara en las muestras que presentan una proporción mayor al 50% de ADN con el SNP para el linaje 2.

Comparando ambos gráficos (**Figura 9**), en la **Figura 9.B** se observa una mejor detección de las co-infecciones ya que se detectan los picos de fluorescencia en las temperaturas correspondientes a las Tm de los linajes 2 y 4 incluso cuando sus proporciones son del 90% y 10% respectivamente, mientras que en el **Figura 9.A** solo aparecen ambos picos cuando los ADNs se encuentran a partes iguales. De este modo, se podría concluir que la sensibilidad de la detección de co-infecciones depende del linaje al que pertenecen las cepas que dan lugar a la co-infección, así como de las proporciones en las que se encuentra cada una de ellas en la muestra. Tal y como se observa en la **Figura 9**, se puede decir que mediante la técnica de la qPCR-HRM se podrían llegar a detectar dos poblaciones de MTBC que se encuentran en una proporción similar (50%-50%) y solo si ambas contienen los SNPs que se testan en dicha PCR multiplex. Además, en las PCR uniplex de ambas pruebas se obtuvo este mismo resultado, es decir, solo las muestras que presentaban más del 50% del ADN del linaje que se estaba testando, daban un pico de fluorescencia. Sin embargo, para poder sacar conclusiones más robustas se deberían de realizar más pruebas con más combinaciones de distintos linajes.

En las muestras de Liberia se intentó detectar co-infecciones, pero esto no pudo ser posible. Relacionando estos resultados con los obtenidos del ensayo de detección de co-infecciones, se podría decir que una posible razón podría ser la baja concentración de partida de ADN de las muestras. De este modo, si alguna muestra estuviera co-infectada con ADNs pertenecientes a dos cepas de distinto linaje y uno de ellos estuviera a una proporción mucho mayor que el otro, el de baja proporción podría haber dado una señal de fluorescencia tan baja que podría haberse confundido con el ruido de fondo de la técnica, haciendo imposible su detección.

A



B

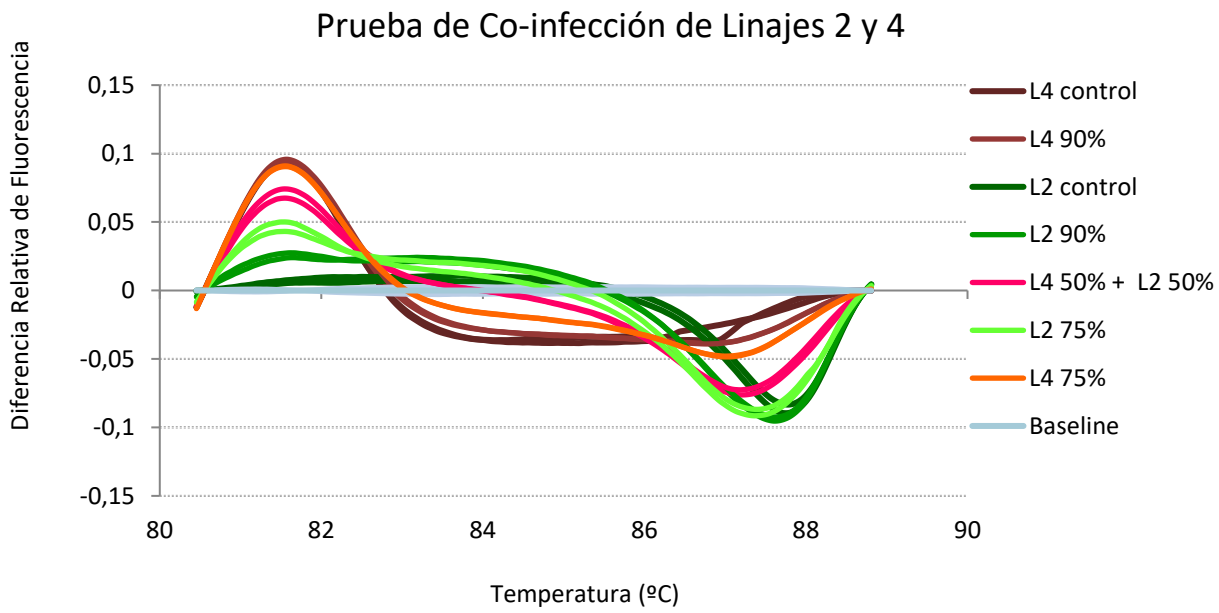


Figura 9: Curvas de fusión de la PCR de multiplexado para detección de co-infecciones. **A)** PCR de multiplexado de los linajes 3 y 4. Las curvas rojas gruesas representan los controles del linaje 4, las rojas claras, las muestras con un 90% de ADN con el SNP para el linaje 4, las naranjas representan las muestras con un 75% de ADN con el SNP para el linaje 4 y las rosas representan las muestras con un 50% de ADN con el SNP para cada linaje. Por otro lado, las curvas moradas oscuras representan los controles con el SNP del linaje 3, las moradas claras representan las muestras con un 90% de ADN con el SNP del linaje 3 y las más claras representan las muestras con un 75% de ADN con el SNP del linaje 3. **B)** PCR de multiplexado para los linajes 2 y 4. Las curvas rojas oscuras representan los controles con el SNP característico del linaje 4 y las verdes oscuras son los controles para el linaje 2. En cuanto a las muestras, las curvas rojas más claras y naranjas representan las muestras con un 90% y 75% de ADN con el SNP del linaje 4, respectivamente. Las curvas verdes representan las muestras con ADN que presenta el SNP característico del Linaje 2 y las curvas rosas representan las muestras con un 50% de la muestra con el ADN que tiene el SNP del linaje 4 y 50% del Linaje 2.

4.7 Secuenciación genómica mediante NGS:

Se llevó a cabo la secuenciación de una de las muestras que resultaron negativas para MTBC en la prueba de la PCR cuantitativa a tiempo real, para poder comprobar si tanto el hecho de partir de inactivados clínicos como la baja concentración de ADN de las muestras afectaban de algún modo a la obtención de las lecturas. La razón era que los inactivados cínicos podían provocar la inhibición de la PCR de indexado en la construcción de las librerías de secuenciación, no obstante, se decidió no realizar un paso previo de purificación del ADN porque se partía de concentraciones bajas de ADN y por tanto, se podría haber perdido el ADN durante el proceso de purificación.

La secuenciación se llevó a cabo mediante el MiSeq de Illumina®. Tras obtener las lecturas se hizo un análisis de las secuencias usando el programa bioinformático Kraken. Kraken es un programa que permite llevar a cabo clasificaciones taxonómicas de secuencias de ADN de los organismos de forma rápida y precisa. Para ello, utiliza como base de datos la RefSeq que contiene genomas bacterianos y, a partir de ellos, genera una biblioteca de k-meros de 31 bases. Al introducir las lecturas de una secuencia, el programa las separa en k-meros de la misma longitud y los compara con los k-meros de la biblioteca. De este modo, es capaz de determinar a qué organismo pertenece el genoma de la muestra desconocida según el porcentaje de K-meros que coinciden entre la secuencia problema y las de los organismos de la base de datos (Wood *et al.*, 2014).

Al analizar la muestra, se obtuvo un porcentaje elevado de coincidencia con el genoma de *Mycobacterium avium*, que es una micobacteria no tuberculosa, y otro pequeño porcentaje de coincidencia con algún microorganismo del género *Bacillus*. Por un lado, *M. avium* es una de las especies más comunes que afecta sobre todo a pacientes con VIH avanzado o fibrosis quística (Uthman Muhammed Mubashir, Uthman Olalekan and Yahaya, 2013; Skolnik, Kirkpatrick and Quon, 2016). Por otro lado, la presencia de ADN de bacterias del género *Bacillus* podría ser debida a restos de algún alimento que se hubiese tomado el paciente ya que las muestras clínicas se obtienen a partir de esputos que luego se cultivan en cultivos para diagnóstico y se inactivan.

El hecho de haber obtenido este resultado ha permitido poder determinar que uno de los pacientes del estudio estaba infectado con una micobacteria no tuberculosa, pero, además, ha permitido poder tener una primera visión de la posibilidad de secuenciar muestras que se encuentran a una concentración más baja de la que recomiendan los protocolos de secuenciación. Por tanto, teniendo en cuenta este resultado se procederá a secuenciar el resto de las muestras en estudios posteriores.

5. CONCLUSIÓN.

Este trabajo ha permitido realizar una clasificación de un conjunto de muestras de MTBC procedentes de pacientes de Liberia, utilizando la técnica del análisis de las curvas de fusión de los productos de la PCR a tiempo real para el genotipado de las cepas mediante SNPs. De este modo, se han podido comprobar algunas de las características principales de esta técnica.

La PCR a tiempo real seguida de un análisis de HRM es una técnica simple, rápida y barata que presenta una resolución elevada a la hora de discriminar variantes de secuencia. Además, el hecho de tener puestos a punto los parámetros para llevar a cabo el multiplexado de distintos linajes incrementa aún más la rapidez del genotipado ya que permite clasificar cepas de diversos linajes en una sola reacción.

El uso de esta técnica ha permitido tener una primera visión de la distribución de los linajes de MTBC en Liberia siendo el Linaje 4 seguido del Linaje 1 los más abundantes. Además, se ha podido identificar la presencia de sublinajes específicos de otras regiones como el Cameroon/L4.6.2 en mayor abundancia a otros más comunes como Haarlem/L4.1.2 o X/L4.1.1. Lo más relevante de los resultados obtenidos es que esta distribución se aleja bastante a la de otros países africanos cercanos a Liberia, por lo que sería de gran interés poder obtener las secuencias de estas cepas.

Además, también ha permitido determinar que las posibles co-infecciones que pudiera presentar alguna de las muestras solo se podrían haber identificado si ambos ADNs con los SNPs de distinto linaje estuviesen en la misma proporción, 50-50%. Por esta razón, junto con el hecho de que las muestras provenían de cultivos de aislados clínicos y este paso podría haber implicado una selección de las cepas más adaptadas y una pérdida de las menos adaptadas, la detección de muestras con co-infecciones no ha sido posible en este ensayo.

No obstante, a pesar de ser una técnica muy robusta también presenta algunas limitaciones ya que solo permite identificar las cepas que contienen los SNPs característicos de los linajes que se testan y, debido a su gran resolución, la probabilidad de obtener falsos positivos y datos anómalos aumenta.

Por tanto, para completar el genotipado se puede realizar la secuenciación masiva de los ADNs. Las técnicas de NGS presentan algunos inconvenientes como el coste o la complejidad del análisis de los datos. Sin embargo, permiten obtener unos resultados mucho más robustos ya que, al obtener las secuencias de cada muestra se pueden llevar a cabo estudios comparativos y de este modo se puede llegar a identificar si el paciente estaba infectado solo con MTBC o si presentaba una co-infección con otro microorganismo. Así mismo, la secuenciación por NGS también permite identificar si una misma muestra presenta una co-infección de cepas de distintos linajes e incluso sublinajes, las cuales son más difíciles de detectar mediante qPCR-HRM.

Finalmente, debido a la interesante distribución de los linajes de las cepas de Liberia obtenida, en estudios posteriores se procederá a la secuenciación de las muestras de Liberia para confirmar el genotipado llevado a cabo en este proyecto y, además, clasificar aquellas cepas que no han podido ser caracterizadas por la técnica de genotipado mediante SNPs y detectar posibles casos de co-infección.

6. BIBLIOGRAFÍA.

ACHTMAN, M. (2008) 'Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens', *Annual Review of Microbiology*, 62(1), pp. 53–70. doi: 10.1146/annurev.micro.62.081307.162832.

BARBIER, M. AND WIRTH, T. (2016) 'The Evolutionary History, Demography, and Spread of the Mycobacterium tuberculosis Complex', *Microbiology Spectrum*, 4(4), pp. 1–21. doi: 10.1128/microbiolspec.TBTB2-0008-2016.

KAPA BIOSYSTEMS. (2017) *Introduction to High Resolution Melt Analysis*. Available at: www.kapabiosystems.com (Accessed: 1 June 2017).

BRITES, D. AND GAGNEUX, S. (2015) 'Co-evolution of Mycobacterium tuberculosis and Homo sapiens', *Immunological Reviews*, 264(1), pp. 6–24. doi: 10.1111/imr.12264.

BROSCH, R. *et al.* (2002) 'A new evolutionary scenario for the Mycobacterium tuberculosis complex', *Proceedings of the National Academy of Sciences*, 99(6), pp. 3684–3689. doi: 10.1073/pnas.052548299.

COMAS, I. *et al.* (2009) 'Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies', *PLoS ONE*, 4(11). doi: 10.1371/journal.pone.0007815.

COMAS, I. *et al.* (2010) 'Human T cell epitopes of mycobacterium tubercolsis are evolutionarily hyperconserved', *Nature genetics*, 42(6), pp. 498–503. doi: 10.1038/ng.590.Human.

COMAS, I. *et al.* (2013) 'Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans', *Nature Genetics*. Nature Publishing Group, 45(10), pp. 1176–1182. doi: 10.1038/ng.2744.

COMAS, I. AND GAGNEUX, S. (2009) 'The past and future of tuberculosis research', *PLoS Pathogens*, 5(10), pp. 1–7. doi: 10.1371/journal.ppat.1000600.

COPIN, R. *et al.* (2014) 'Impact of in vitro evolution on antigenic diversity of Mycobacterium bovis bacillus Calmette-Guerin (BCG)', *Vaccine*, 32(45), pp. 5998–6004. doi: 10.1016/j.vaccine.2014.07.113.

COSCOLLA, M. AND GAGNEUX, S. (2014) 'Consequences of genomic diversity in Mycobacterium tuberculosis', *Seminars in Immunology*, 26(6), pp. 431–444. doi: 10.1016/j.smim.2014.09.012.

DELOGU, G., MANGANELLI, R. AND BRENNAN, M. J. (2014) 'Critical research concepts in tuberculosis vaccine development', *Clinical Microbiology and Infection*, 20, pp. 59–65. doi: 10.1111/1469-0691.12460.

DHEDA, K., BARRY, C. E. AND MAARTENS, G. (2016) 'Tuberculosis', *The Lancet*, 387(10024), pp. 1211–1226. doi: 10.1016/S0140-6736(15)00151-8.

ERNST, J. D. (2012) 'The immunological life cycle of tuberculosis', *Nature Reviews Immunology*. Nature Publishing Group, 12(8), pp. 581–591. doi: 10.1038/nri3259.

FILLIOL, I., MOTIWALA, A. AND CAVATORE, M. (2006) 'Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy', *Journal of …*, 188(2), pp. 759–772. doi: 10.1128/JB.188.2.759.

FIRDESSA, R. *et al.* (2013) 'Mycobacterial lineages causing pulmonary and extrapulmonary Tuberculosis, Ethiopia', *Emerging Infectious Diseases*, 19(3), pp. 460–463. doi: 10.3201/eid1903.120256.

FOMUKONG, N. *et al.* (1998) 'Differences in the prevalence of IS6110 insertion sites in M. tuberculosis strains:low and high copy number of IS6110', *Tub Lung Dis*, 78(1998), pp. 109–116. doi: 10.1016/S0962-8479(98)80003-8.

GAGNEUX, S. *et al.* (2006) 'Variable host-pathogen compatibility in Mycobacterium tuberculosis', *Proceedings of the National Academy of Sciences*, 103(8), pp. 2869–2873. doi: 10.1073/pnas.0511240103.

GAGNEUX, S. AND SMALL, P. M. (2007) 'Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development', *Lancet Infectious Diseases*, 7(5), pp. 328–337. doi: 10.1016/S1473-3099(07)70108-1.

GAN, M. *et al.* (2016) 'Deep whole-genome sequencing to detect mixed infection of mycobacterium tuberculosis', *PLoS ONE*, 11(7), pp. 1–14. doi: 10.1371/journal.pone.0159029.

GEHRE, F. *et al.* (2016) 'A Mycobacterial Perspective on Tuberculosis in West Africa: Significant Geographical Variation of M. africanum and Other M. tuberculosis Complex Lineages', *PLoS Neglected Tropical Diseases*, 10(3), pp. 1–11. doi: 10.1371/journal.pntd.0004408.

HERSHBERG, R. *et al.* (2008) 'High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography', *PLoS Biology*, 6(12), pp. 2658–2671. doi: 10.1371/journal.pbio.0060311.

HO, T. B. *et al.* (2000) 'Comparison of Mycobacterium tuberculosis genomes reveals frequent deletions in a 20 kb variable region in clinical isolates.', *Yeast (Chichester, England)*, 17(4), pp. 272–282. doi: 10.1002/1097-0061(200012)17:4<272::AID-YEA48>3.0.CO;2-2.

HOUBEN, R. M. G. J. AND DODD, P. J. (2016) 'The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling', *PLoS Medicine*, 13(10), pp. 1–13. doi: 10.1371/journal.pmed.1002152.

INTERNAL CLINICAL GUIDELINES TEAM (UK). (2016) 'Tuberculosis: Prevention, Diagnosis, Management and Service Organization.', in. Available at: https://www.ncbi.nlm.nih.gov/books/NBK361233/.

ISSA, R. *et al.* (2014) 'High resolution melting analysis for the differentiation of Mycobacterium species', *Journal of Medical Microbiology*, 63(2014), pp. 1284–1287. doi: 10.1099/jmm.0.072611-0.

MALLARD, K. *et al.* (2010) 'Molecular detection of mixed infections of Mycobacterium tuberculosis strains in sputum samples from patients in Karonga District, Malawi', *Journal of Clinical Microbiology*, 48(12), pp. 4512–4518. doi: 10.1128/JCM.01683-10.

PÉREZ-LAGO, L. *et al.* (2015) 'Fast and low-cost decentralized surveillance of transmission of tuberculosis based on strain-specific PCRs tailored from whole genome sequencing data: A pilot study', *Clinical Microbiology and Infection*. Elsevier, 21(3), p. 249.e1-249.e9. doi: 10.1016/j.cmi.2014.10.003.

SKOLNIK, K., KIRKPATRICK, G. AND QUON, B. S. (2016) 'Nontuberculous Mycobacteria in Cystic Fibrosis', *Current Treatment Options in Infectious Diseases*, 8(4), pp. 259–274. doi: 10.1007/s40506-016-0092-6.

SMITH, N. H. *et al.* (2006) 'Ecotypes of the Mycobacterium tuberculosis complex', *Journal of Theoretical Biology*, 239(2), pp. 220–225. doi: 10.1016/j.jtbi.2005.08.036.

STUCKI, D. *et al.* (2012) 'Two new rapid SNP-typing methods for classifying mycobacterium tuberculosis complex into the main phylogenetic lineages', *PLoS ONE*, 7(7). doi: 10.1371/journal.pone.0041253.

STUCKI, D. *et al.* (2016) 'Mycobacterium tuberculosis lineage 4 comprises globally distributed and

geographically restricted sublineages', *Nature Genetics*, 48(12), pp. 1535–1543. doi: 10.1038/ng.3704.

THERMO FISHER SCIENTIFIC (2017) *Qubit 3.0 Fluorometer*. Available at: www.thermofisher.com/es/es/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit/qubit-fluorometer.html (Accessed: 15 May 2017).

UTHMAN MUHAMMED MUBASHIR, B., UTHMAN OLALEKAN, A. AND YAHAYA, I. (2013) 'Interventions for the prevention of mycobacterium avium complex in adults and children with HIV', *Cochrane Database of Systematic Reviews*, (4). doi: 10.1002/14651858.CD007191.pub2.

WAMPANDE, E. M. *et al.* (2015) 'A single-nucleotide-polymorphism real-time PCR assay for genotyping of Mycobacterium tuberculosis complex in peri-urban Kampala', *BMC Infectious Diseases*. BMC Infectious Diseases, 15(1), p. 396. doi: 10.1186/s12879-015-1121-7.

WARREN, R. M. *et al.* (2004) 'Patients with Active Tuberculosis often Have Different Strains in the Same Sputum Specimen', *American Journal of Respiratory and Critical Care Medicine*, 169(5), pp. 610–614. doi: 10.1164/rccm.200305-714OC.

WHO (2016) 'Global Tuberculosis Report 2016', *Cdc 2016*, (Global TB Report 2016), p. 214. doi: ISBN 978 92 4 156539 4.

WHO (2017) *Tuberculosis (TB).* Available at: http://www.who.int/mediacentre/factsheets/fs104/en/ (Accessed: 29 May 2017).

WJ, A. (2015) 'Next Generation DNA Sequencing (II): Techniques, Applications', *Journal of Next Generation Sequencing & Applications*, 1(S1), pp. 1–10. doi: 10.4172/2469-9853.S1-005.

WOOD, D. E. *et al.* (2014) 'Kraken: ultrafast metagenomic sequence classification using exact alignments.', *Genome biology*, 15(3), p. R46. doi: 10.1186/gb-2014-15-3-r46.

ZAMAN, K. (2010) 'Tuberculosis: A global health problem', *Journal of Health, Population and Nutrition*, 28(2), pp. 111–113. doi: 10.1186/1471-2334-13-122.

ANEXO I: Tabla de los oligos de los linajes utilizados en este estudio.

MTBC Linaje	Posición del SNP	Cambio de nucleótido	Gen	Secuencia de los cebadores 5'-3'	Tamaño del producto	Referencia
1	115499	T/G	nrp	F-ATAATATTGCGTCGGTGTTGG R-ttatatattaATGGGCAGGCC	81bp	No publicados.
2	3304966	G/A	Rv2952	F-TGTTACCCGCACTTTCGGCGTTT R-AGGTCGGCGTATGGGAGGTA	80bp	Fenner, et al. 2011
3	4266647	A/G	fbpA	F-CGTTGAGATGAGGATGAGGG R-GCGACATACCCGTGACGGC	92bp	Fenner, et al. 2011
4	2154724	A/C	KatG	F-CCGAGATTGCCAGCCTTAAG R-ACTGGTACCGCAATACCGTC	64bp	Gagneux, et al. 2006
5	456731	C/T	Rv0380c	F-GCATCGTGTCCGAAGTTCTC R-ATCATCGCCGACATCGATAC	68bp	No publicados.
6	378404	G/A	Rv0309	F-CCGACAGCGAGAACCTGC R-CCATCACGACCGAATGCTT	54bp	Stucki, et al. 2012
M.bovis	2831482	T/G	Rv2515c	F-GTGTTGCTGTCGATGACGC R-ACTGGTACCGCAATACCGTC	91bp	No publicados.

ANEXO II: Tabla con los oligos de los sublinajes utilizados en este estudio.

Sublinaje MTBC	Familia espoligo	Posición del SNP	Cambio de nucleótido	Gen	Secuencia de los cebadores 5'-3'	Tamaño del producto	Referencia
L4.1.1	X	3798451	C/G	Rv3383c	5'-ATCGACTCAATGGCCCGATG 3'-TGACTCTGGATGCGGTTTT	112bp	No publicados.
L4.1.2	Haarlem	4323348	C/T	Rv3848/3849	5'-AAATCCGTTCGTCGTGTGGA 3'-CTGACGTTGTGAGGGGTCAA	82bp	No publicados.
L4.1.3	Ghana	4409231	T/G	Rv3921c	5'-GACCGCCTCCTGCTTTTTG 3'- ACGTCTTCGGCATGATCGAA	53bp	No publicados.
L4.2	Ural	2942377	C/T	Rv2614c	5'-GAGTAGTCCTCCAGTTCGCG 3'-TCAGCTTCCCCGACGAAATC	85bp	No publicados.
L4.3	LAM	1480024	G/T	Rv1318c	5'-CAGGCCAGGATCCACATCAG 3'-TGCTGCTCAATCTCACTCGG	100bp	No publicados.
L4.4	Vietnam	4307886	G/A	Rv3834c	5'-AAGGTGGTGCAGTTCGAC 3'-ACTGCGAGGCGTGGATTC	69bp	No publicados.
L4.5	Iran	2789341	A/C	Rv2483c	5'-GGAGGCCTCACCATCCTTG 3'-ACGAAGGCGGCTACAAAGAA	81bp	No publicados.
L4.6.1	Uganda	435708	G/A	Rv0357c	5'-CAAAGATCCCGCTGGGTCAT 3'-GATATGAGATCGACGGCCGG	58bp	No publicados.
L4.6.2	Cameroon	3191099	C/A	Rv2881c	5'-CATCATGCAGAACACCCATC 3'-CCCATTGTTCTGCTCTTTCG	72bp	No publicados.
L4.10	PGG3	1692141	C/A	Rv1501	5'-GCTCGGTGTTCTTCGACTCA 3'-TGGCCGTTTCAGATAGCACA	107bp	No publicados.