
Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Química Biológica

# **Desarrollo de herramientas bioinformáticas para la priorización de blancos moleculares y análisis genómico de resistencia a antibióticos en *Mycobacterium tuberculosis*.**

Tesis para optar al título de Doctor de la Universidad de Buenos Aires en el Área Química Biológica

**Ezequiel Jorge Sosa**

Director de tesis: Darío Fernández Do Porto

Consejero de Estudios: Nancy Lopez

Lugar de trabajo: IQUIBICEN - FCEN - UBA

Buenos Aires, 2022

# Índice

# Agradecimientos

Primero que nada quiero agradecer a mi director Darío, que con su infinita paciencia me fue acompañando, insistiendo que no era tan fácil escribir la tesis y corrigiendo ( varias rondas durante meses), sin él, esta tesis hubiese ido al cajón de las cosas que jamás voy a terminar.

A Adrián y a Marcelo, que siempre me ayudaron y orientaron, me dieron el lugar que ocupo hoy, donde puedo trabajar de lo que me gusta y seguir aprendiendo de todos.

A Lucas, Lanza, Mode y Radu, que establecieron las bases de esta tesis y me tocó continuar con parte de las líneas de trabajo que dejaron.

A Claudia, de quien también aprendí a trabajar y con el tiempo llegamos a formar una amistad basada en largas horas de trabajo remoto :)

¡Al equipo de BIA! Que ya cambió tantas veces, desde los originales con Dario, Sergio, German y Jony, pasando por Gero, Leo, Andy, Agustin, Guada, Facu, Daiana hasta la forma actual con Miri, Claudio, Fran, Maria, Gus, y Juanma, con personajes importados como Lupe, Karina, Lucas, Yasminn y Rafa. Con todos me sentí/siento muy cómodo y siempre trabajamos con buena onda y cagándonos de risa todos los días, no puedo pedir más. No puedo dejar de hacer una mención especial a Fede y a Flor, que me acompañaron codo a codo este último año y me dieron una mano enorme con la tesis, particularmente Flor con su STAN de figuras y su TOC a la hora de corregir los formatos de este documento.

A mis amigos de toda la vida Ariel, Leo, Charly, Martin, Edic y Pablo, que no me ayudaron en absolutamente nada relacionado con la tesis, pero siento que tenían que ser nombrados de alguna forma :)

A Lucas, con quien arrancamos comiendo pizza en Pin Pun después de corregir finales y ahora es mi amigo y siempre puedo contar con su consejo y apoyo!

A mi familia Marcelo, Paulita, Manu, Rafa, Paula, Cami, Cata, Claudito, Maria Rosa, por estar siempre conmigo y compartir tantas cosas!

¡Obviamente a mis papas! Por enseñarme a que siempre tengo que trabajar de lo que me gusta y me apasiona! Gracias a eso puedo levantarme contento todos los días! No pude seguir su ejemplo de levantarme temprano o ser puntual, pero siempre trate de hacer lo que me gusta, todos los días. En este agradecimiento también están incluidos mis suegros, Hector y Alicia, que siempre están ahí para ayudaros y para compartir buenos momentos!

El agradecimiento final y la dedicación de esta tesis van para mi compañera Mariana, quien me banca en todo y a pesar que no hacemos un buen equipo a la hora de remar en un kayak, en todo lo demás si :) compartimos risas, picos de estrés, y sobre todo buenos memes y dibujitos :) y a mi pequeño Chichini! Pasó tan rápido, desde que nos apapachamos por primera vez cuando pesaba menos de un kilo, hasta hoy que nos pegamos altas siestas, ¡siempre juntos!

# Abreviaturas

| | |
|---|---|
| ADN | Ácido desoxirribonucleico |
| AMS | Alineamiento múltiple de secuencias |
| ARN | Ácido ribonucleico |
| CIM | Concentración inhibitoria mínima |
| CSA | *Catalytic Site Atlas* |
| DB | Base de Datos |
| dNTPs | Desoxinucleótidos trifosfato |
| DR | Repeticiones dirigidas |
| DS | Score de Drogabilidad |
| DS | Score de Drogabilidad |
| DSA | Determinación de susceptibilidad a antibióticos |
| EBI | *European Bioinformatics Institute* |
| EC | *Enzyme Commission* |
| EMB | Etambutol |
| ERON | *Especies reactivas de oxígeno y nitrógeno* |
| FAS | Sintesis de acidos grasos |
| FQL | Fluoroquinolonas |
| GO | Gene Ontology |
| HTS | High throughput screening |
| ID | Identificador |
| INH | Isoniazida |
| LAM | Latinoamérica y Mediterráneo |
| Lm | *Listeria monocytogenes* |
| LSPs | *Large sequence polymorphisms* / polimorfismos de secuencias largas |
| mg | Mili gramo |
| min | Minutos |
| ml | Mili litro |
| mM | Milimolar |
| MRSA | *Staphylococcus aureus* meticilino resistente |

| | |
|---|---|
| *Mtb* | *Mycobacterium tuberculosis* |
| *MTBC* | Complejo *Mycobacterium tuberculosis* |
| NCBI | Centro Nacional de Información biotecnológica de los Estados Unidos |
| NGS | Next-Generation Sequencing (Secuenciación de 2da generación) |
| OMS | Organización Mundial de la Salud |
| PCR | Reacción en cadena de la polimerasa |
| PDB | *Protein Data Bank* |
| PFAM | *Protein Family Database* / Base de dominios proteicos del EBI |
| PG | Peptidoglicano |
| PSD | Prueba de sensibilidad a las drogas |
| PZA | Pirazinamida |
| RIF | Rifampicina |
| RMN | Resonancia magnética nuclear |
| SAR | *Structure-activity relationship* / Relación estructura actividad |
| SF | Scoring function / función de puntuación-rankeo |
| SIDA | Síndrome de inmunodeficiencia adquirida |
| SMILES | Especificación de introducción lineal molecular simplificada |
| SNP | *Single nucleotide polymorphism* / polimorfismo de nucleótido único |
| SP | Swiss-Prot (subconjunto de UniProt) |
| STR | Estreptomicina |
| TB | Tuberculosis |
| TB MDR | Tuberculosis multirresistente |
| TB XDR | Tuberculosis extremadamente resistente |
| TP | Target-Pathogen |
| Tr TrEMBL | TrEMBL Translated EMBL (subconjunto de UniProt) |
| UniProt | *Universal Protein Resource* |
| UniRef | UniProt Reference Clusters Database (subconjunto de UniProt) |
| VS | *Virtual screening* |
| WGS | *Whole genome sequencing* / Secuenciación de genoma completo |
| WT | *wild type* / fenotipo silvestre |

# Resumen

La tuberculosis (TB) es una enfermedad crónica causada por la bacteria intracelular facultativa *Mycobacterium tuberculosis* (*Mtb*), un patógeno altamente exitoso. A pesar de contar con más de 100 años de investigación, la TB actualmente es la enfermedad infecciosa más mortífera por detrás del COVID-19. Se calcula que un cuarto de la población mundial se encuentra infectada latentemente con este patógeno. Los últimos reportes de la organización Mundial de la Salud (OMS) estimaron que alrededor 1.6 millones de personas murieron durante 2021 a causa de esta enfermedad. En particular, las cepas multirresistentes (MDR, resistentes a isoniacida y rifampicina) y extremadamente resistentes (*Mtb XDR*, resistente a isoniacida, rifampicina, una fluoroquinolona y un aminoglucósido inyectable de segunda línea), representan un problema serio para los sistemas de salud. Argentina, en relación a otros países, tiene una carga media de esta enfermedad. Alrededor de 10500 nuevos casos de TB son notificados anualmente, con cerca de 1000 decesos anuales. Aunque las mismas suelen tratarse y resolverse con una terapia estándar con al menos 3 antibióticos diferentes durante 6-9 meses, no es el caso con las infecciones con *Mtb* XDR, que requiere un tratamiento más costoso y extendido en el tiempo y muchas veces es de pronóstico reservado.

En este trabajo se desarrolló un flujo de trabajo bioinformático (*pipeline*) con el objetivo de analizar las bases moleculares de resistencia a antibióticos de *M. tuberculosis* y se aplicó al estudio de los aislamientos XDR circulantes en Argentina entre los años 2008-2016, cuyos genomas fueron obtenidos por técnicas de secuenciación masiva. Las variantes genotípicas obtenidas para cada aislamiento fueron contrastadas contra bases de datos de variantes asociadas a resistencia y se realizó una análisis filogenético sobre las mismas. Con los datos procesados, no se observaron orígenes comunes para las mutaciones de resistencia más frecuentes, pero sí dentro de determinados grupos monofiléticos para algunas drogas. Dentro de las variantes màs frecuentes para las drogas de primera línea se pueden mencionar *katG* Ser315Thr y *fabG1* C-15T para INH, *rpoB* Ser450Trp y Asp435Val para RIF, *embB* Gly406Asp y Met306Ile para EMB y variantes en el gen de *pncA* para PZA.

En el marco de la emergencia provocada por las cepas resistentes, multirresistentes y extremadamente resistentes, se requiere con urgencia el desarrollo de nuevos antimicrobianos. A pesar de esta situación crítica, el diseño y la producción de nuevas drogas ha resultado ineficaz por diferentes razones que incluyen el elevado costo asociado al desarrollo de las mismas y una tasa de éxito relativamente baja. El desarrollo de drogas es un proceso complejo que requiere entre 10 y 20 años para atravesar las distintas etapas de investigación y una inversión aproximada de 1500 millones de dólares. En esta situación resulta esencial la adecuada selección inicial del blanco molecular a ser inhibido por el antibiótico en desarrollo. En este sentido, los métodos de secuenciación masiva han sido una enorme fuente de datos, creando nuevas oportunidades para el diseño y desarrollo de drogas para combatir a los agentes infecciosos, incluyendo a los resistentes y multirresistentes. En la presente tesis se ha desarrollado Target-Pathogen, una aplicación web diseñada y desarrollada específicamente como un recurso online que permite la integración y valoración de diferentes características de todas las proteínas de un genoma (función, rol metabólico, homología con proteínas humanas, drogabilidad estructural, esencialidad, expresión) para facilitar la identificación y priorización de proteínas como blancos adecuados. En la base de datos se incluyen los genomas de 25 de los microorganismos más relevantes para la salud humana en continua expansión y actualización. Utilizando Target-Pathogen se seleccionaron y priorizaron una serie de blancos moleculares prometedores para el desarrollo de drogas para *Mycobacterium tuberculosis* en la fase latente, entre los que se destaca *ino1*, una proteìna que participa de vìa de la síntesis de micotiol.

**Palabras clave**: multirresistencia, genómica, tuberculosis, drogas, blancos.

# Abstract

**Resistance diagnostic and drug development methods from *Mycobacterium tuberculosis* genomics data**

Tuberculosis (TB) is a chronic disease caused by the intracellular bacterium *Mycobacterium tuberculosis* (*Mtb*), a highly successful pathogen. Despite having more than 100 years of research, *Mtb is the second leading infectious killer after COVID-19*. It is estimated that a quarter of the world population is latently infected with this pathogen. The latest reports from the World Health Organization (WHO) estimated that around 1.6 million people died during 2021 from this disease. In particular, multiresistant strains (MDR, resistant to isoniazid and rifampin) and extremely resistant (*Mtb* XDR, resistant to isoniazid, rifampin, a fluoroquinolone, and a second-line injectable aminoglycoside), represent a serious problem for health systems. Argentina, in relation to other countries, has a medium burden of this disease. About 10,500 new TB cases are reported annually, with about 1,000 deaths annually. Although they are usually treated and resolved with standard therapy with at least 3 different antibiotics for 6-9 months, this is not the case with *Mtb* XDR infections, which require more expensive and extended treatment over time and is often guarded prognosis.

In this work, a bioinformatic *pipeline* was developed with the aim of analyzing the molecular bases of resistance to antibiotics in *M. tuberculosis* and it was applied to the study of XDR strains circulating in Argentina between the years 2008-2016, whose genomes were obtained by techniques of massive sequencing. The genotypic variants obtained for each isolate were compared against databases of variants associated with resistance and on phylogenetic analyses, both locally and globally. With the processed data, no common origins were observed for the most frequent resistance mutations. Among the most frequent variants for first-line drugs, we can mention KatG Ser315Thr and fabG1 C-15T for INH, rpoB Ser450Trp and Asp435Val for RIF, embB Gly406Asp and Met306Ile for EMB, and variants in the pncA gene for PZA.

In the context of the emergency caused by resistant, multiresistant and extremely resistant strains, the development of new antimicrobials is urgently required. Despite this critical situation, the design and production of new drugs has been ineffective for different reasons, including the high cost associated with their development and a relatively low success rate. Drug development is

a complex process that requires between 10 and 20 years to go through the different stages of research and an investment of approximately 1.5 billion dollars. In this workflow, the adequate initial selection of the molecular target to be inhibited by the antibiotic in development is essential. In this sense, massive sequencing methods have been an enormous source of data, creating new opportunities for the design and development of drugs to combat infectious agents, including resistant and multi-resistant ones. In this thesis we have developed Target-Pathogen, a web platform specifically designed and developed as an online resource that allows the integration and evaluation of different characteristics of all the proteins of a genome (function, metabolic role, homology with human proteins, structural druggability, essentiality, expression) to facilitate the identification and prioritization of proteins as suitable targets. The database includes the genomes of 23 of the most relevant microorganisms for human health in continuous expansion and updating. Using Target-Pathogen, promising molecular targets for the development of drugs for *Mycobacterium tuberculosis in the latent phase were selected and prioritized*, among which *ino1* stands out, a protein that participates in the synthesis of mycothiol.

**Keywords**: multiresistance, genomics, tuberculosis, drugs, targets.

# Introducción

## Tuberculosis

La Tuberculosis (TB) es una enfermedad infecciosa, considerada una de las primeras enfermedades humanas, generalmente pulmonar, cuyo agente etiológico es la bacteria *Mycobacterium tuberculosis* (Mtb). A pesar de tratarse de una enfermedad prevenible y tratable –contando con diversos tratamientos antibióticos seguros y asequibles, y una vacuna, la del bacillus Calmette-Guerin (BCG) –, la TB constituye, incluso al día de hoy, una amenaza sanitaria a nivel mundial [1]

La enfermedad se transmite por vía aérea. Un individuo con TB activa exhala gotas microscópicas infecciosas, las cuales permanecen en la atmósfera y pueden ser inhaladas por otros individuos que se encuentren cerca. En general, sólo una pequeña proporción de personas infectadas con *M. tuberculosis* desarrolla enfermedad activa, pero la probabilidad es mayor entre las personas con morbilidades asociadas, en particular aquellas co-infectadas con virus de la inmunodeficiencia humana (VIH).

Sin tratamiento, la mortalidad por TB es alta[2]. Sin embargo, TB suele tratarse -y resolverse- con una terapia estándar con al menos 3 antibióticos (generalmente los primeros dos meses con cuatro drogas de primera línea: isoniacida (INH), rifampicina (RIF), etambutol (EMB) y pirazinamida (PZA); seguido de los siguientes cuatro meses con INH y RIF. Con un buen cumplimiento del tratamiento, se logra la cura del 86% de los casos en un periodo de 6-9 meses. Sin embargo, esta primera línea de terapia falla en un porcentaje significativo de casos debido a una pobre adherencia al tratamiento[3] y/o a la aparición o contagio con cepas resistentes.

TB ha ganado importancia mundial desde hace varias décadas debido a la complejidad de su reemergencia, tratamiento y control. En el contexto de la pandemia de SARS-CoV-2 (coronavirus de tipo 2 causante del síndrome respiratorio agudo severo), la TB pasó a ser la enfermedad infecciosa más mortífera por detrás del mismo (sobrepasando al VIH/SIDA)[4]. Este hecho aceleró aún más, la necesidad apremiante de diagnosticar la TB de forma veloz y eficiente. Más aún, la pandemia resultó en la escasez de recursos necesarios para el diagnóstico y tratamiento de la TB, llevando a un aumento alarmante de casos, tanto en sus variantes sensibles como resistentes al tratamiento de antibióticos estándar. Esto generó un retroceso importante en los avances de la lucha contra la TB, volviendo indispensable el desarrollo de nuevas medidas [5,6] [1].

## Epidemiología

A nivel mundial, la tuberculosis es considerada una de las principales causas de mortalidad en humanos[7] y hasta la pandemia causada por el SARS-CoV-2, la primera provocada por un único agente infeccioso. Se estima que el 25 % de la población mundial se encuentra infectada con *Mtb* de manera latente, en el 2020 hubo aproximadamente 10 millones de nuevos casos y 1,5 millones de muertes provocadas por esta enfermedad (Organización Mundial de la Salud 2021).

La epidemiología varía notablemente entre los países de América Latina[8]. La incidencia de TB en América Central (incluido México), el Caribe y América del Sur fue de 25,9, 46,2 y 61,2 por 100.000 habitantes (Figura 1). La farmacorresistencia es un problema creciente en las Américas, particularmente en Perú, donde la tuberculosis resistente a los medicamentos representa el 9% de los casos[8]. En este marco, solo el 33% de los pacientes recibieron pruebas de susceptibilidad a los medicamentos, lo que da como resultado un estimado de 7000 pacientes con tuberculosis resistente a los medicamentos sin diagnosticar o sin tratar[8]. Alrededor de una cuarta parte de la población latinoamericana está infectada de forma latente con *Mtb*.

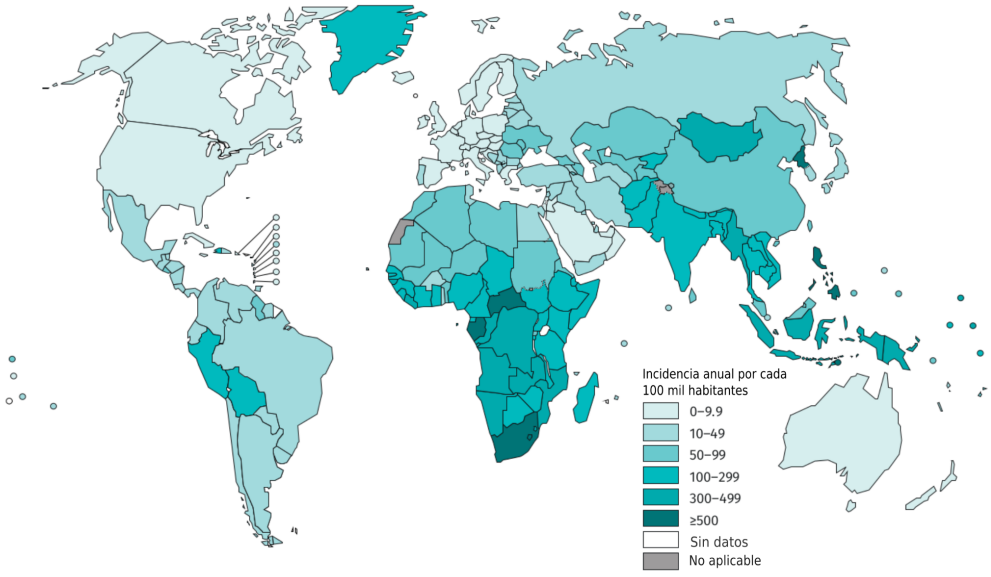



Figura 1: Incidencia mundial de la tuberculosis en el año 2020 (Figura adaptada. Organización Mundial de la Salud, 2021).

Los datos obtenidos por el Ministerio de Salud de la Nación arrojaron cerca de 10.900 nuevos casos en Argentina durante 2020 (Figura 2) con una tasa de mortalidad del 1.4 por cada 100 mil habitantes. Se estima que la cifra del 2021 fue de alrededor de 11.690 casos.

De los 6.857 casos de tuberculosis pulmonar confirmados bacteriológicamente en 2020, en 2.176 (31,7%) se registró la realización de la prueba de sensibilidad a las drogas (PSD): 1.966 (90,3%) fueron sensibles y 210 (9,7%) presentaron algún tipo de resistencia.

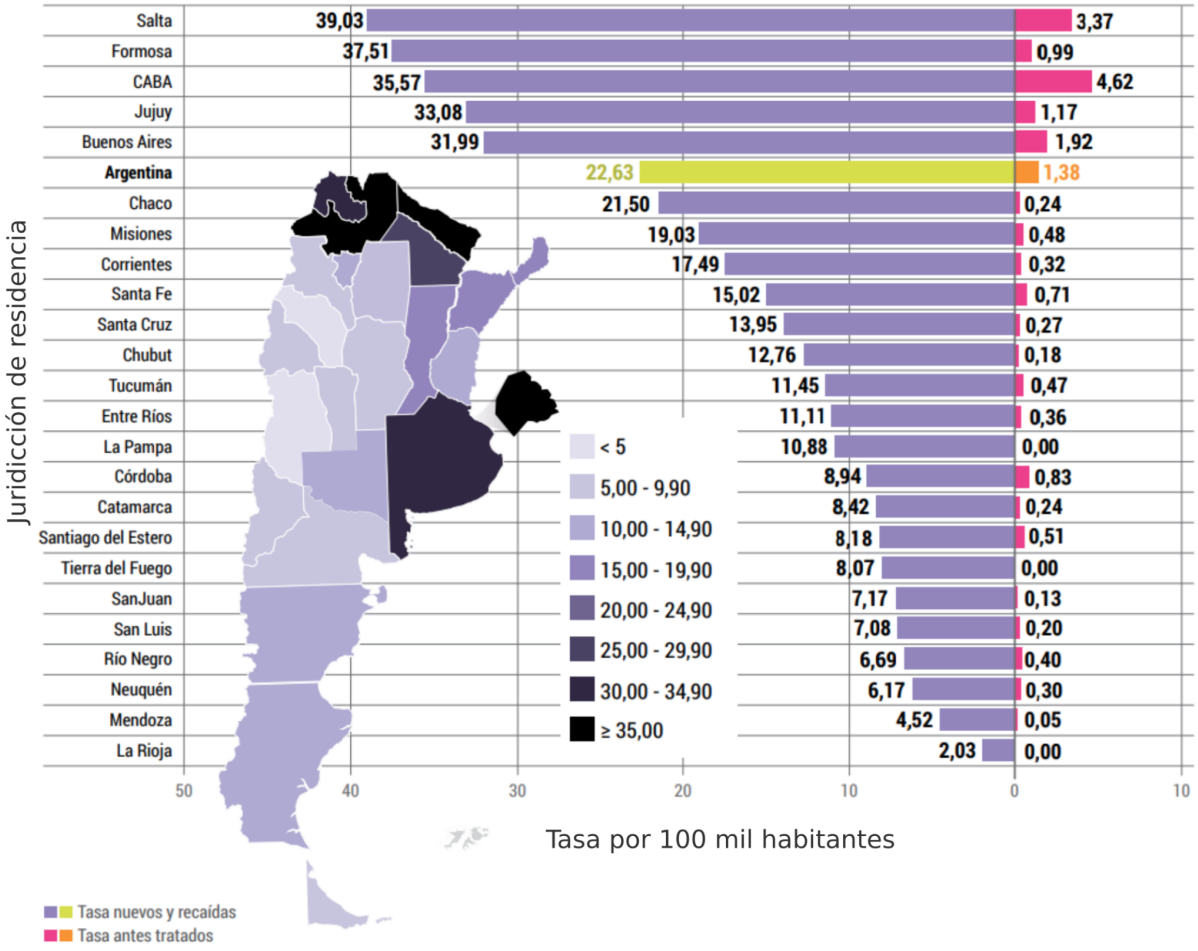



Figura 2: Incidencia de la tuberculosis en el año 2020. Tasas de notificación de casos de tuberculosis cada 100 mil habitantes según jurisdicción de residencia y tipo de paciente (Ministerio de Salud de la Nación, Argentina, 2022).

## *Mycobacterium tuberculosis*

El Complejo *Mycobacterium tuberculosis* (o MTBC) comprende a un grupo de especies micobacterianas de crecimiento lento, y estrechamente relacionadas genéticamente: Mtb (Figura 3) *Mycobacterium canettii*, *Mycobacterium africanum*, *Mycobacterium microti*, *Mycobacterium bovis*, *Mycobacterium caprae* y *Mycobacterium pinnipedii*– capaces de causar TB tanto en humanos como en animales.

Las bacterias pertenecientes al género *Mycobacterium* son bacilos inmóviles y no esporulados pertenecientes al orden Actinomycetales. Estos bacilos contienen un alto porcentaje (61-71%) de guanina y citosina (G+C) en su genoma, una pared celular con alto contenido de lípidos y varios ácidos micólicos característicos de la envoltura de las micobacterias. Estos lípidos poco convencionales actúan como reservas de carbono y energía. También están involucrados en la estructura y funcionamiento de membranas y de orgánulos membranosos dentro de la célula. Los lípidos constituyen más de la mitad del peso seco de la célula. La composición lipídica de los bacilos puede variar durante su ciclo de vida en cultivo, dependiendo de la disponibilidad de nutrientes. La envoltura celular confiere las principales características del género: resistencia al ácido alcohol, extrema hidrofobicidad, persistencia en las lesiones, resistencia a antibióticos y sus distintivas características inmunológicas. Esta resistencia se cree que depende de la integridad de la cubierta de cera y su envoltura altamente hidrofóbica [9]. También es probable que contribuya al lento crecimiento de algunas especies ya que habría una restricción en el ingreso de nutrientes [10].

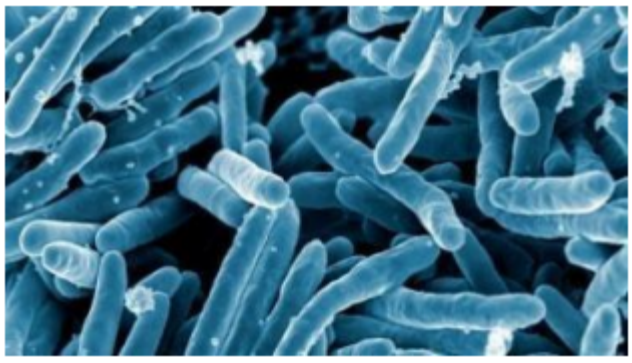


Figura 3: *Mycobacterium tuberculosis*. Imagen obtenida mediante microscopía electrónica de barrido (SEM). National Institute of Allergy and Infectious Diseases.

*M. tuberculosis* es considerado genéticamente monomórfico, porque tiene bajos niveles de diversidad genética y homoplasia [11] (eventos de mutaciones independientes que resultan en el mismo genotipo entre cepas con diferentes ancestros) y muy raros eventos de recombinación homóloga. Sin embargo, en comparación con otras bacterias monomórficas, presenta una sustancial variación en su ADN (ácido desoxirribonucleico). Los polimorfismos identificados como polimorfismos largos (LSPs, del inglés *large sequence polymorphisms*) y polimorfismos de nucleótido único (SNPs, del inglés *single nucleotide polymorphisms*) son filogenéticamente informativos y útiles para los estudios poblacionales y epidemiológicos.

Además, el cromosoma posee otras regiones repetitivas que son fuentes de variabilidad: la secuencia de inserción IS6110, las secuencias polimórficas repetitivas ricas en GC (*polymorphic GC-rich repetitive sequences*, PGRSs), así como también polimorfismos en las repeticiones cortas palindrómicas agrupadas regularmente espaciadas (*clustered regularly interspaced short palindromic repeats*, CRISPRs) y repeticiones en tándem de número variable (*variable number tandem repeats*, VNTR). Todos ellos han sido aplicados a estudios de epidemiología molecular [12,13]

## Resistencia antimicrobiana

El uso exitoso de antibióticos ha enfrentado desafíos debido a que los patógenos microbianos están desarrollando diversas formas de resistencia en las últimas décadas (Figura 4) y *Mycobacterium tuberculosis* no está exenta de este fenómeno [14].

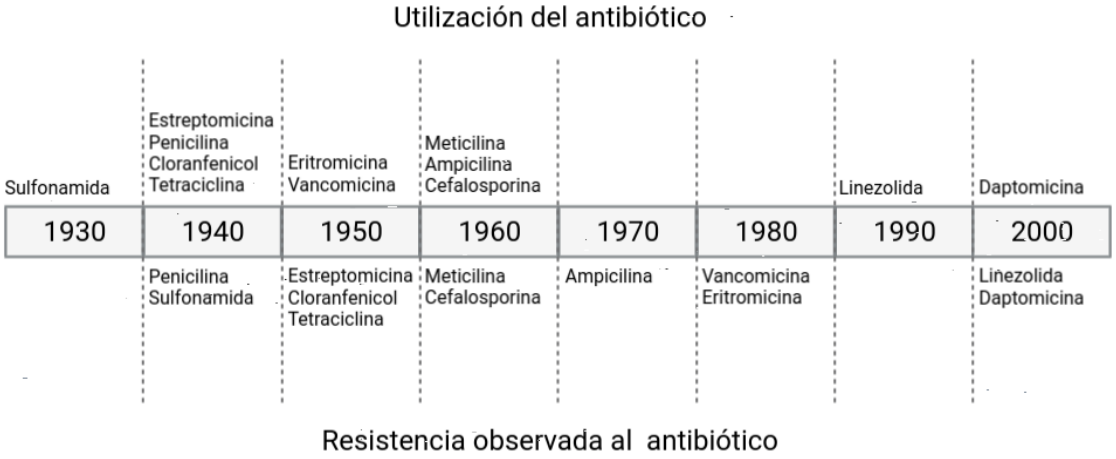



Figura 4: Línea de tiempo comparativa de desarrollo de antibiótico y posterior aparición de resistencia microbiana. Modificado de Clatworthy et al. 2007 [15]

Países de todo el mundo, independientemente del nivel de ingresos, están elevando los niveles de resistencia a los antibióticos [16,17–19]. El centro de prevención y control de enfermedades de Estados Unidos declaró en 2013 que la raza humana se encuentra ahora en la "era post-antibiótica" y recientemente, la Organización Mundial de la Salud (OMS) advirtió que la crisis de resistencia a los antibióticos es cada vez más grave [20,21]. Las bacterias resistentes detectadas con mayor frecuencia han sido *Escherichia coli*, *Klebsiella pneumoniae*, *Staphylococcus aureus, Mycobacterium tuberculosis* y *Streptococcus pneumoniae*, seguidas de la *Salmonella spp* [22].

En el manejo del paciente y la epidemiología de la TB existen los conceptos de monorresistencia y polirresistencia. La OMS define la monorresistencia como resistencia de *Mtb* a uno de los fármacos de primera línea, mientras que la polirresistencia se refiere a la resistencia a dos o más fármacos de primera línea que no sean INH junto con RIF. Las cepas resistentes RIF e INH, MDR-TB (*multidrug resistant*) –y las cepas XDR-TB (*extensively drug resistant*) resistentes a RIF, INH, un antibiótico inyectable de segunda línea (capreomicina, kanamicina o amikacina) y una fluoroquinolona [23,24], son un problema serio para el sistema de salud global. Finalmente, aunque la organización mundial de la salud no reconoce el término de forma oficial aún, existen casos de TDR-TB (Totally drug resistant) los cuales, como su nombre lo indica, poseen resistencias a todas las drogas actualmente utilizadas contra TB [25].

La infección por una cepa resistente de un paciente que nunca ha recibido tratamiento es conocida como resistencia primaria. Ésta incluye infección por cepas silvestres que nunca han estado en contacto con fármacos, es decir que poseen una resistencia natural, así como la resistencia que se desarrolla como consecuencia de la exposición de una cepa determinada a un fármaco, pero en otro paciente. Se define, por otro lado, como resistencia secundaria o adquirida, aquella que se desarrolla en pacientes que han recibido tratamiento con drogas antituberculosas, debido a la selección de cepas mutantes resistentes espontáneas (Figura 5), en la mayoría de los casos, como consecuencia a un tratamiento inadecuado o incumplimiento de la terapia [17–19].

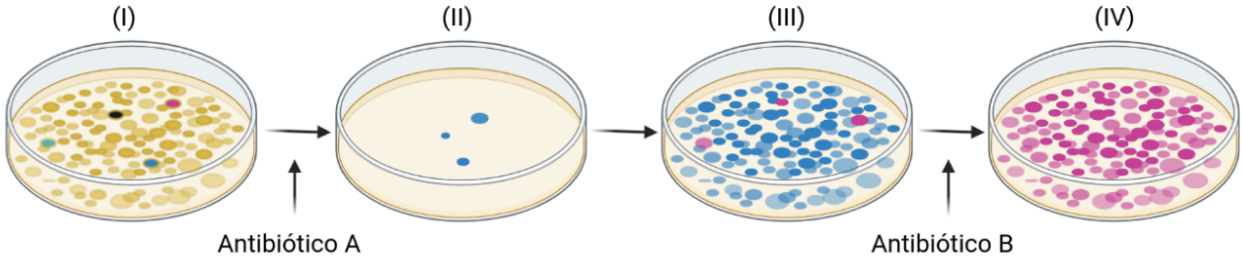



Figura 5: Obtención de polirresistencias adquiridas por selección de mutantes.

## Desarrollo de nuevas drogas

En este contexto resulta apremiante el desarrollo de nuevas drogas. Sin embargo, las inversiones en el área de investigación y desarrollo de nuevos antibióticos se ha desalentado, en las últimas décadas, debido a las restricciones de los organismos de control y al elevado costo monetario que estos desarrollos involucran, aun en el ámbito de las grandes industrias farmacéuticas. Por otra parte, el desarrollo de antibióticos ya no se considera una inversión rentable para la industria debido a que los antibióticos se usan durante períodos relativamente cortos y con frecuencia son curativos [26]. Cabe destacar también, que, la elección aleatoria de compuestos o las modificaciones estructurales en compuestos conocidos ya no representan metodologías atractivas ni eficaces para acompañar la veloz aparición de bacterias resistentes [27]. En consecuencia, la incorporación de antibióticos al mercado ha tenido una curva marcadamente descendente en las últimas décadas (Figura 6), por lo que antibióticos disponibles para uso clínico cada vez son menos en relación a la aparición de cepas bacterianas multiresistentes, capaces de sobrevivir a la acción de múltiples fármacos [28].

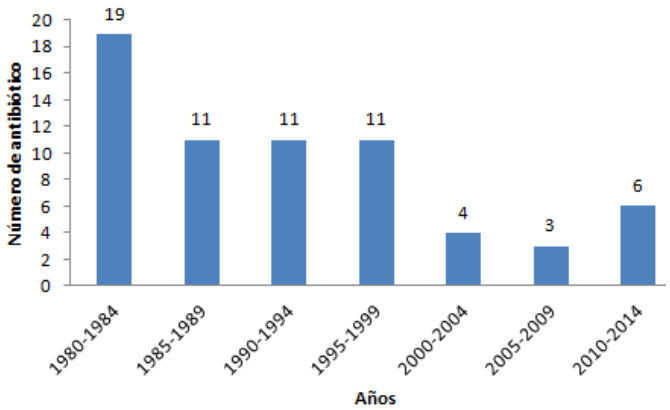



Figura 6: Número de aprobaciones de aplicación de nuevos fármacos antibacterianos por intervalo de años. Tomado de Ventola et al. 2015 [28].

El descubrimiento de fármacos es un proceso complejo y que requiere una importante inversión de recursos y tiempo. El paradigma actual para el desarrollo de una nueva droga comprende dos grandes etapas: el descubrimiento de la droga y posteriormente su desarrollo (Figura 7*)*. La etapa de descubrimiento comprende tres fases: i) Identificación del blanco molecular que se desea atacar; ii) Búsqueda de compuestos líderes capaces de modular (o inhibir, en el caso de antibióticos) la actividad del blanco; iii) Validación del blanco y optimización del fármaco.

Asimismo, la etapa de desarrollo involucra la fase pre-clínica y las cuatro fases clínicas (I, II, III y IV). En la fase pre-clínica se llevan a cabo distintos ensayos para la identificación de un fármaco seguro, potente y eficaz. Durante esta etapa que requiere pruebas exhaustivas, se evalúan aspectos de la farmacodinamia, la farmacocinética y la toxicología en entornos *in vitro* e *in vivo* (en modelos no humanos). En la fase clínica I se llevan a cabo evaluaciones de seguridad del fármaco en pacientes. En las dos fases posteriores, a la vez que se continúa analizando la seguridad del fármaco, se evalúa la dosis y su eficacia, aumentando el tamaño de la población analizada entre la fase clínica II y III. La última etapa es la fase de farmacovigilancia (Fase IV), que permite, durante la etapa de comercialización, detectar efectos adversos no previstos en las etapas previas de control y evaluación del medicamento.

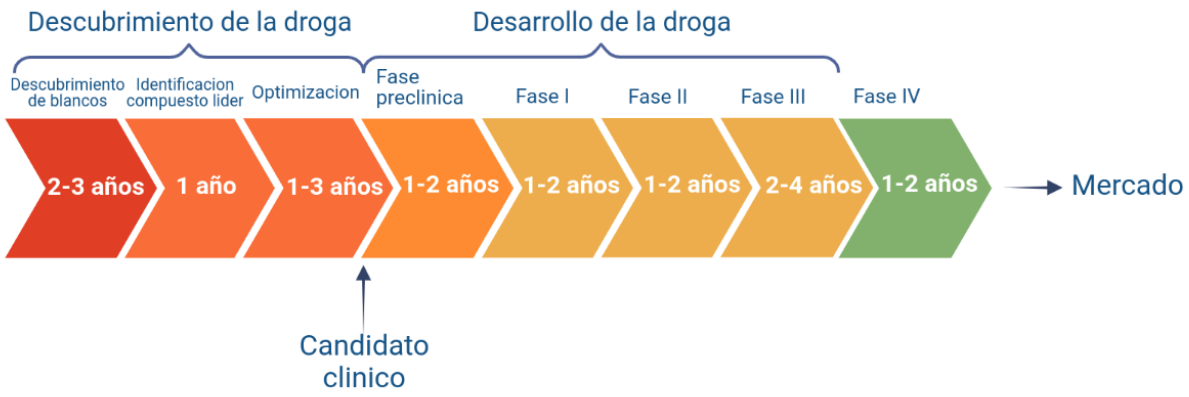



Figura 7: Procesos involucrados en las fases de descubrimiento y desarrollo de una nueva droga.

En la etapa de descubrimiento de blancos es donde la mayoría de los fármacos falla y su inadecuada selección hace al fracaso asegurado de cualquiera de las etapas posteriores de desarrollo, perdiendo una gran cantidad de tiempo y dinero. En este sentido si el mecanismo molecular resulta desconocido o no se comprende qué sucede al inhibir la función de una enzima a nivel sistémico, se podrían presentar efectos adversos inesperados [29,30].

## Genómica

La genómica es una rama de la ciencia que se aboca al estudio de genomas completos, abarcando todos los genes que se encuentran en un organismo [31]. En este sentido, es rol de la genómica dilucidar la composición de genes de un organismo, tanto a nivel de estructura como a nivel de sus funciones moleculares, celulares y fisiológicas y, en última instancia, desvelar el comportamiento de los sistemas biológicos.

La genómica no implica el estudio de genes individuales sino que su propósito es entender los efectos de cada gen en el contexto de todo el genoma. Dentro de la genómica en general, podemos distinguir varias sub-áreas:

- Genómica estructural: es la rama de la genómica orientada a la localización (predicción de genes) y caracterización (asignación de función) de las secuencias que conforman los genes [32], tanto codificantes de proteínas como de ARN (ácido ribonucleico) no codificantes, así como de elementos reguladores.
- Genómica funcional: se aboca a la recolección sistemática de información sobre la función de los genes intentando entender las interacciones entre ellos y la regulación de su expresión. A diferencia de la genómica estructural, la genómica funcional se centra en los aspectos dinámicos de los genes, como su transcripción, la traducción, las interacciones proteína-proteína, en oposición a los aspectos estáticos de la información genómica como la secuencia del ADN o su estructura.
- Genómica comparativa: estudia las semejanzas y diferencias entre genomas de distintos organismos, comparando todos los genes, regiones, rearreglos, etc. entre individuos o especies.

A fin de comprender un organismo real, quien desee caracterizar un sistema biológico no puede limitarse sólo al estudio del genoma estático. El nivel de expresión de los genes de un organismo varía notablemente de un tejido a otro y en diferentes condiciones, ya sea ambientales, fisiológicas, de desarrollo, o circunstancias especiales como una infección. De esta necesidad de estudiar la expresión global de los genomas surgen otras áreas como la transcriptómica, que estudia globalmente los ARNs que se expresan en determinados tejidos o circunstancias de interés, o la proteómica, que estudia el contenido expresado a nivel de proteínas. En todos los casos, la caracterización puede hacerse apoyándose en referencia a un genoma conocido de interés o realizando estrategias denominadas “*de novo*” que no requieren de conocimiento previo.

En sistemas en los que no se conoce la secuencia nucleotídica del genoma completo, si se tiene un set de genes o regiones conocidas, la transcriptómica puede partir de la secuenciación de ADNc de porciones de ARN mensajero (EST) y su caracterización por análisis de expresión mediante micromatrices de ADN (microarrays o chips de ADN). Los desarrollos de NGS han permitido en estas últimas décadas obtener tanto un transcriptoma completo como sus niveles de expresión sin necesidad de contar con un genoma de referencia (RNA-seq). La proteómica suele basarse fuertemente en el análisis mediante espectrometría de masas y otras técnicas de alta sensibilidad.

Con el correr de los años estas disciplinas se diversificaron y dieron lugar a la revolución de las ciencias "ómicas"[33], donde el sufijo pasa a ser un neologismo que se refiere al estudio de la totalidad o del conjunto de entidades biológicas como genes, organismos de un ecosistema, proteínas, o incluso las relaciones entre ellos. Surgen así, además de las disciplinas mencionadas, la interacómica, la metabolómica, la metagenómica, la epigenómica, la lipidómica, la fenómica, la secretómica y la glicómica, entre otras[34].

Se hace una mención especial, por ser de particular relevancia para esta tesis, la estructurómica, que estudia el conjunto de estructuras tridimensionales de las proteínas codificadas por un genoma, que en los últimos años realizó un salto cualitativo y cuantitativo enorme con el desarrollo de Alphafold [35] y la base de datos generada a partir del método [36]

## Secuenciación de ADN

Podemos encontrar los orígenes de la secuenciación del ADN en la técnica de Sanger, diseñada en 1975. El método inicial lograba secuenciar fragmentos de pocos cientos de nucleótidos, y era un proceso lento y laborioso. Con el progreso en las tecnologías ópticas, el refinamiento de las técnicas de PCR (reacción en cadena de la polimerasa) [37] y la disponibilidad de moléculas fluorescentes a bajo costo, esta técnica fue evolucionando hasta lo que en la actualidad se conoce como secuenciación de Sanger por electroforesis capilar. Los equipos actuales son totalmente automatizados, logrando bajar considerablemente los tiempos y los costos, llegando a leer fragmentos de ADN de hasta 1500 pb. Sigue utilizándose de manera rutinaria y su calidad permite en muchas ocasiones validar mutaciones puntuales obtenidas por métodos más modernos. A pesar que esta metodología fue preponderante durante muchos años, su costo y rendimiento comenzaron a convertirse en un factor limitante en lo que respecta a secuenciación masiva y rutinaria.

A comienzos del siglo XXI hacen su aparición diversas tecnologías y equipamientos que se conocen en conjunto como métodos de alto rendimiento, de secuenciación masiva o de nueva generación (del inglés "*Next-Generation Sequencing*", o NGS). Estos métodos lograron generar un salto cualitativo en base a la miniaturización y la paralelización. Con esta técnica fueron secuenciadas las muestras de este trabajo, en particular con equipos de la marca Illumina. El funcionamiento de esta tecnología se describe a continuación.

El procedimiento de esta técnica comienza con la fragmentación del ADN, generando pequeños fragmentos de distinto tamaño. Posteriormente se unen a sus extremos unas moléculas llamadas adaptadores (secuencias conocidas de nucleótidos que son complementarias a los oligonucleótidos fijados en la base de las celdas de secuenciación). El conjunto de todos los fragmentos de ADN unidos a sus adaptadores se conoce como biblioteca.

Dentro del secuenciador Illumina, en compartimentos llamados celdas, se bombean polimerasas, dNTPs ( desoxinucleótidos trifosfato ) y soluciones *buffer*. La base de cada celda tiene adosados oligonucleótidos cortos complementarios a las secuencias del adaptador, por lo tanto, los adaptadores de cada biblioteca hibridarán con estos oligonucleótidos, inmovilizando temporalmente las hebras de ADN monocatenarias en la celda de flujo (Figura 8b). Las cadenas hibridadas de ADN se amplifican usando una estrategia de "PCR en puente" que emplea ciclos de extensión del cebador seguidos de desnaturalización química. A través del proceso de amplificación in situ, cada hebra se amplifica por miles. Las bibliotecas de ADN se hibridan con la celda de flujo en cantidades molares bajas (6-20 pM), lo que da como resultado una gran separación física entre las cadenas de ADN. Al final de la amplificación "PCR en puente", quedan grupos de ADN idénticos (clusters) inmovilizados en una superficie 2D, que pueden secuenciarse en masa. Entonces, comienza el procedimiento de secuenciación en sí:

- Una polimerasa incorpora una única base que contiene un fluoróforo y un bloqueo reversible en la posición 3'.
- Se toma una imagen de la celda de flujo usando microscopía fluorescente.
- Los restos fluorescentes y terminadores se cortan enzimáticamente, permitiendo que se incorpore una nueva base en el próximo ciclo.

Esto se repite una cantidad de ciclos dada dependiendo del cartucho de secuenciación utilizado, con una pérdida mínima de señal o precisión (Figura 8c). La cantidad de ciclos de secuenciación determina el largo de las lecturas. Existen dos tipos de lecturas: *paired-end* y

*single-end*. La secuenciación *single-end* implica secuenciar el ADN desde un solo extremo y la secuenciación *paired-end* permite secuenciar ambos extremos de un fragmento de ADN. La tecnología *paired-end* nos brinda información no solo de la secuencia si no que se conoce cual es la distancia entre ambos fragmentos. Esta información resulta de gran utilidad para el ensamblado "*de novo*" de genomas. Después de la secuenciación, el software del secuenciador identifica los nucleótidos (un proceso denominado *base calling* o llamado de bases) y la precisión o calidad prevista para cada una de esas bases durante el llamado de bases.

Los datos que se obtienen del secuenciador se denominan archivos FASTQ, un formato basado en texto que almacena tanto las secuencias de nucleótidos (de cada uno de los fragmentos) como sus puntajes de calidad correspondientes en código ASCII (Figura 8c). Actualmente, la tecnología Illumina ofrece FASTQ con lecturas de máximo 600nt de longitud, dependiendo del kit de secuenciación utilizado. Es importante considerar que ante el uso del tipo de secuenciación *paired-end* las lecturas en realidad tendrán como máximo 300 nt de longitud (300 nt leídos directo y 300 nt leídos reverso, conocido como 2x300).

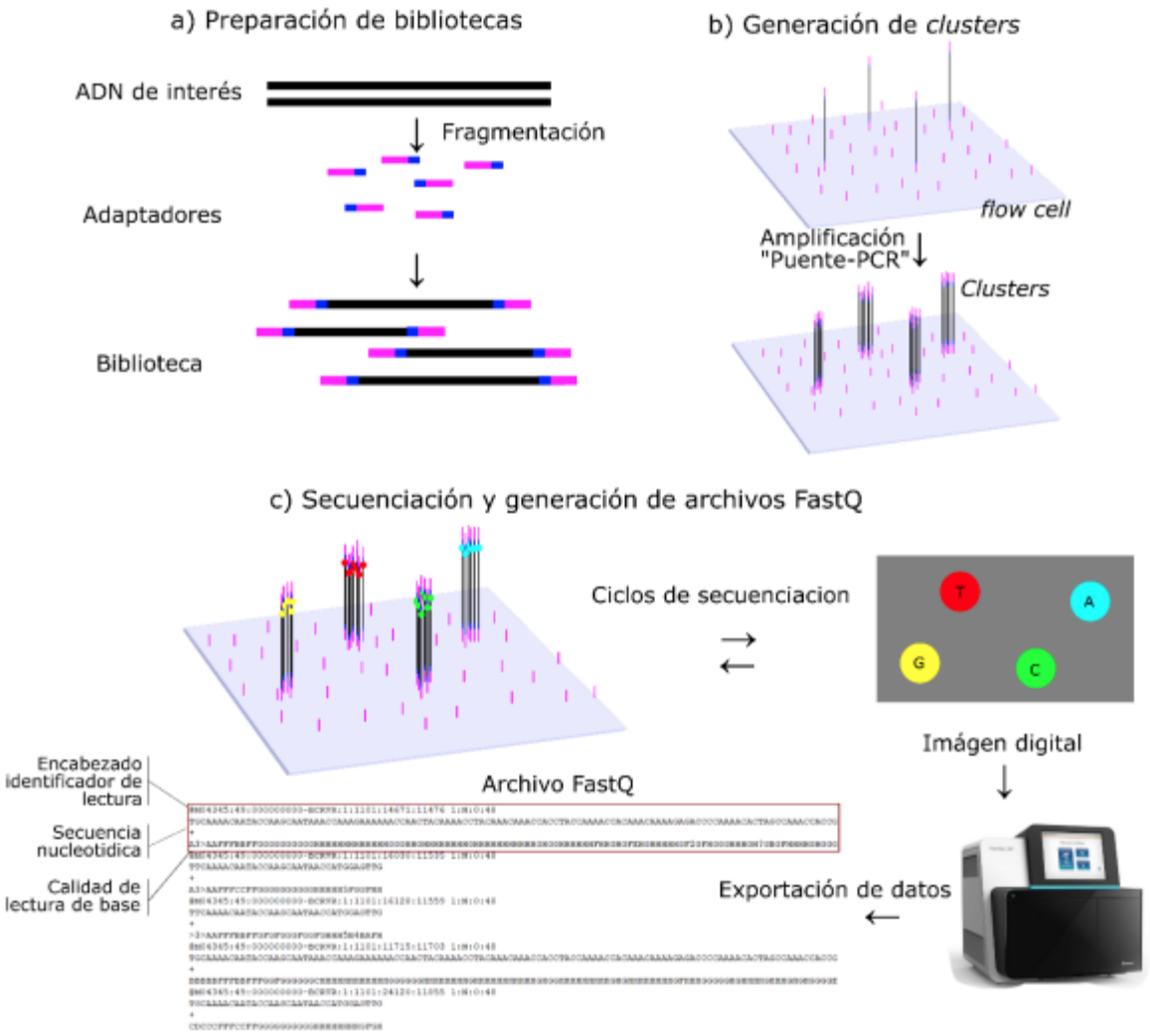



Figura 8: Metodología de trabajo illumina. (a) el primer paso es la generación de bibliotecas, donde el ADN de interés se fragmenta y se le adicionan los adaptadores para su unión a la base del equipo. (b) allí se amplifica cada fragmento a partir de la amplificación en puente, generando *clusters*. Posteriormente, comienza la secuenciación, agregando un nucleótido marcado con fluorescencia por ciclo. (c) Se toman imágenes de la celda y cada cluster realiza emisiones fluorescentes correspondientes al nucleótido sintetizado. Luego la información de cada *cluster* se traduce a un nucleótido. Este ciclo se repite N cantidad de veces, para crear una lectura de N nucleótidos de longitud, que serán recopilados junto a la información de calidad de dicho reconocimiento. El resultado final se vuelca en un archivo de texto en formato FastQ.

Por otro lado, actualmente tenemos técnicas de secuenciación de tercera generación. Estas estrategias se basan en la idea de eliminar la etapa de síntesis y obtener lecturas más largas que faciliten el posterior ensamblado. Algunas de estas tecnologías son Single Molecule Real-time (SMRT), desarrollada por Pacific Biosciencies (Flusberg et al. 2010) y la secuenciación por nanoporos [38], desarrollada por Oxford Nanopore. Los pasos para procesar las secuencias obtenidas con estas técnicas son muchas veces análogas respecto a las utilizadas en NGS

(segunda generación), sin embargo, su mayor tasa de error, lecturas más largas y naturaleza de cómo se procesa la molécula, hacen que las inferencias, modelos, validaciones y algoritmos utilizados (y por ende los programas que los implementan) sean distintos.

## Alineamiento

Una de las herramientas más utilizadas en la bioinformática y que ha sido usada intensivamente en este trabajo de tesis es el alineamiento de secuencias [39]. Un alineamiento de secuencias (ya sea de proteínas, ADN o ARN) consiste en el apareamiento de dos (o más) secuencias, de modo que cada una de las letras de una secuencia tenga un correspondiente en las otras, o un espacio llamado "*gap*" cuando no se tiene correspondencia. La inserción de *gaps* se hace de manera que se maximice la cantidad de caracteres de una secuencia cuyos homólogos a priori en las otras tengan el mismo carácter. Dado que es posible alinear secuencias insertando *gaps* de muchas maneras distintas, se establece un sistema de puntuación que beneficia las coincidencias entre letras de las secuencias ("match") y penaliza la inserción de gaps y las letras no coincidentes ("mismatch"). Los algoritmos de alineamiento de secuencias buscan aquel alineamiento que maximice este puntaje. Estos algoritmos tienen distintos parámetros para determinar el sistema de puntuación que se usará para calificar el alineamiento, entre ellos la matriz de sustitución que determina la probabilidad de cambio de una letra por otra (donde la letra puede ser tanto un nucleótido o un aminoácido, dependiendo de la secuencia), y los parámetros de apertura y extensión de gaps ("GOP" y "GEP" respectivamente). Cuando se alinean sólo dos secuencias se dice que el alineamiento es pareado o "*pairwise*" y cuando se alinean tres o más, se dice que es un alineamiento múltiples secuencias (AMS). Los algoritmos de generación de alineamientos son heurísticos, es decir, que no se obtiene el mejor alineamiento posible de manera absoluta, sino que se obtiene el óptimo entre los que pudo evaluar el algoritmo con el sistema de puntuación usado. Entre los programas más reconocidos y utilizados, para alineamientos pareados se encuentran Blast [40] y Diamond [41], y para MSAs Clustal [42] , Muscle [43] y MAFFT [44] .

## Bases de datos

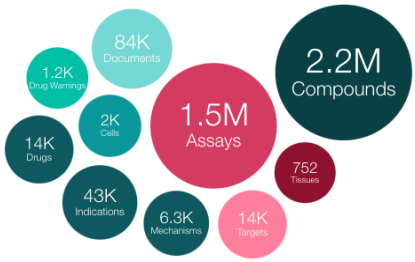
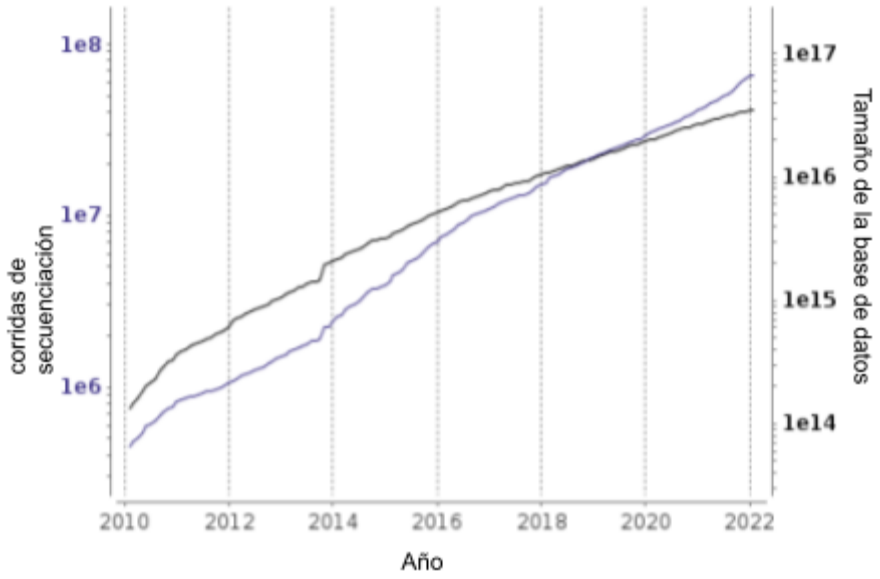
Una base de datos (DB), como concepto general, es un banco de información que contiene registros desglosados en propiedades y vinculados de una manera que permite relacionarlos como conjunto, ejercer sobre dicho conjunto búsquedas en función del valor de las propiedades, clasificarlos en función de filtros a las mismas, etc. En la actualidad, la mayoría de las bases de

datos (DBs) se encuentran almacenadas de una manera digital porque esto permite su rápido procesamiento debido a la capacidad de cómputo de los dispositivos en la actualidad. Podemos definir como base de datos biológica a cualquier colección de información cuyos registros consisten en datos acerca de entidades biológicas. La misma puede provenir de experimentos científicos, literatura publicada, tecnología de experimentación de alto rendimiento o análisis computacional, entre otras fuentes. Una base de datos biológica puede contener información de muy variadas áreas de investigación incluyendo genómica, proteómica, metabolómica, expresión génica mediante (proveniente tanto de microarreglos como de RNA-seq) y filogenética. La información contenida en bases de datos biológicas incluye funciones, estructura y localización (tanto celular como cromosómica), efectos clínicos de mutaciones, así como similitudes de secuencias y estructuras tridimensionales de moléculas [45].

Las publicaciones científicas tradicionales en su conjunto, durante mucho tiempo, funcionaron como un gran repositorio de información donde los investigadores podían consultar información y conocimiento derivado de datos experimentales. Pero con la influencia del campo computacional y las tecnologías de la información a principios de la década del 70, el desarrollo de repositorios basados en sistemas computacionales comenzó a crecer, alcanzando su auge a principios de los 80s con la creación de los grandes consorcios que ocuparon el rol de mantener y distribuir los datos. Para la década del 90 las bases de datos continúan creciendo y empiezan a llamar la atención de la comunidad científica en general, con las primeras secuenciaciones de genomas completos, tanto de procariotes como de eucariotes. Para la década del 2000, las bases de datos se vuelven bastante difíciles de manejar, no sólo por su tamaño, sino además, porque los distintos experimentos biológicos fueron dando lugar a distintos tipos de datos, al punto que en los últimos años fueron indispensables las iniciativas de integración de datos que permitan cruzar información entre varias bases de datos de forma unificada [46]. En la Figura 9 se presentan distintos ejemplos de bases de datos. En (A) se presenta el ejemplo de ChEMBL [47], donde en su versión 31, se procesaron más de 85 mil publicaciones científicas, de manera manual por varios curadores y se sistematizó dicha información a una estructura que permite el estudio de miles de datos bibliográficos en conjunto, a través de un lenguaje de consulta estándar (SQL). En (B) vemos el ejemplo de EBI/ENA (*Nucleotide Archive del European Bioinformatics Institute*) actualizado a principios del 2022[1]. Dicha base de datos se puede categorizar como primaria, ya que almacena datos secuencias sin ningún tipo de procesamientos o curación manual. En dicho gráfico se observan los avances en la secuenciación de cientos de miles de secuencias de todo

[1] https://www.ebi.ac.uk/ena/browser/about/statistics

tipo en los últimos 10 años. El valor de esto no sólo es histórico, sino que estas bases permiten replicar experimentos y sobre todo reutilizar los datos, para distintos análisis más completos, enfocados en distintos grupos (como ser poblaciones locales) o comparar entre ellos. Finalmente en (C) vemos el ejemplo de UNIPROT, la base de referencia de proteínas, donde están cargadas todas los productos génicos proteicos procesados hasta el momento, con cruces a numerosas bases de datos (como PDB, Reactome, GO[48], EC[49], ProSite [50] o Interproscan[51]).

| | |
|---|---|
| A) Captura de pantalla de las estadísticas de datos de ChEMBL versión 30 |  |
| B) ENA Crecimiento de las lecturas y tamaño de la base de datos al 17-feb-2022: (66 millones de corridas y 35 petabytes de información) |  |

| | |
|---|---|
| C) Uniprot (https://www.uniprot.org/statistics/) |  <br>n |

Figura 9: Ejemplos de crecimiento y volúmenes de datos en DBs Bioinformáticas

Hoy día, las bases de datos constituyen uno de los elementos básicos de cualquier investigación biológica dado que resultaron ser la forma más eficiente de almacenar datos biológicos. Una evidencia de esto, es el hecho de la aparición de importantes revistas científicas que publican artículos referidos exclusivamente a bases de datos [52]. Se han llevado adelante numerosos esfuerzos por manejar toda esta información para que esté accesible y, dependiendo de los datos disponibles, los distintos tipos de repositorios ofrecen distintas funcionalidades para poder acceder a los datos. Sin embargo, la integración de todos estos datos siempre se ha mostrado problemática y constituye uno de los principales elementos de análisis a la hora de poner a punto los datos en un proyecto de manera de poder trabajar con ellos [46].

Las DBs utilizadas en cada capítulo de este trabajo se encuentran descritas en la introducción de cada uno.

## Definiciones computacionales importantes

Para entender muchas de las herramientas y resultados de esta tesis, que en sí también son herramientas informáticas, es importante tener claros los siguientes conceptos.

- Algoritmo: especificación de un conjunto de instrucciones, que dado un conjunto de datos de entrada, obtiene un resultado para resolver un problema dado. Casi todos los algoritmos, al igual que una función matemática, tienen parámetros o información auxiliar.
- Script: en particular para esta tesis, nos referiremos a un tipo de programa, que en general implementa uno o más algoritmos. Está orientado a ejecutarse sin una interfaz gráfica y frecuentemente se aplica al procesamiento masivo de datos.
- *Pipeline* (o flujo de trabajo): En general compuesto de uno o más scripts o programas y relaciona/procesa las salidas de un script para ser la entrada de otro/s. Se puede considerar un tipo de script, solo que en general, no implementa un algoritmo, sino que llama a programas que los implementan, sigue prácticas estandarizadas y aplica algún conocimiento puesto a punto para conjuntos de datos con determinadas características. Por su complejidad, en general requieren algún seguimiento en su ejecución y distintas optimizaciones, por ejemplo que ante un error en la corrida de un *pipeline*, éste vuelva a ejecutarse desde donde falló. Gran parte de los resultados de esta tesis fueron *pipelines* y los datos resultantes de su aplicación. Éstos fueron cargados en una base de datos.
- Aplicación Web: es un tipo de programa, el cual se ejecuta en una computadora llamada Servidor, que es accesible por una red de comunicación, en general internet y ofrece funcionalidades a un usuario final (pantallas gráficas, fáciles de entender y utilizar) a través de un navegador (como ser Firefox, Chrome, Edge/Explorer). Uno de los principales resultados de esta tesis, el desarrollo de Target-Pathogen, es una aplicación Web.

# Hipótesis y Objetivos de esta tesis

La hipótesis de trabajo de la presente tesis sostiene que combinando de manera adecuada información proveniente de las distintas ciencias ómicas es posible avanzar en la comprensión de las bases moleculares inherentes a la adquisición de resistencias a antimicrobianos de las cepas XDR en Argentina y en el desarrollo de nuevos fármacos para combatirlas.

El objetivo general de esta tesis es desarrollar herramientas que permitan la búsqueda *in silico* de blancos moleculares para el desarrollo de nuevos fármacos antituberculínicos y permitan la caracterización molecular de resistencia para los aislamientos de *Mtb* XDR a nivel local.

Para cumplir con el objetivo general se plantearon los siguientes objetivos específicos:

1) Desarrollar una herramienta para caracterizar el mecanismo molecular de resistencia de aislamientos clínicos de tuberculosis.
    a) Desarrollar e implementar una herramienta que permita mapear las lecturas de un experimento de secuenciación de *Mtb*, llamar las variantes y cruzar esa información con bases de datos de resistencia.
    b) Utilizar la herramienta desarrollada en el punto anterior para estudiar los mecanismos moleculares de resistencia en aislamientos locales de *Mtb* XDR.
    c) Inferir la filogenia de los aislamientos *Mtb* XDR circulantes en nuestro país en el contexto de genomas representantes de los distintos linajes del mundo.
2) Desarrollar una herramienta para la priorización de blancos proteicos en patógenos
    a) Desarrollar e Implementar una herramienta web que permita la integración datos provenientes de distintas fuentes en escala genómica para el ranqueo de proteínas según su potencial como blanco molecular.
    b) Armar una base de datos de patógenos de importancia local y mundial con información genómica, estructural y metabólica utilizando la herramienta desarrollada en el punto anterior.
    c) Utilizar la herramienta y la base de datos anterior para inferir blancos moleculares de relevancia para el desarrollo de fármacos antituberculínicos.
    d) Utilizar la herramienta y la base de datos anterior para inferir blancos moleculares en otros patógenos de relevancia clínica: *Klebsiella pneumoniae, Bartonella bacilliformis* y *Listeria Monocytogenes.*

# Capítulo 1: Bases moleculares de la resistencia a antibióticos de los aislamientos XDR circulantes en Argentina

## Introducción

En los años 90, nuestro país fue identificado por la OMS como un *hotspot* de TB MDR. A mediados de esa década se documentaron brotes hospitalarios de TB MDR asociados al síndrome de inmunodeficiencia adquirida (SIDA) en la Ciudad Autónoma de Buenos Aires, el conurbano bonaerense, La Plata y Rosario. El brote se inició en el Hospital F. J. Muñiz y se extendió a localidades cercanas fue el de mayor magnitud. Con más de 800 casos diagnosticados entre 1992 y 2004, se puede considerar que adquirió proporciones epidémicas. En los primeros años del nuevo milenio, en ese hospital se documentó el ascenso de la TB MDR en pacientes sin infección por VIH y sin antecedentes de tratamiento para TB [53]. En el Hospital F. J. Muñiz, la TB MDR/SIDA se mantiene en niveles bajos pero estables a través del tiempo. A partir del brote original, la cepa responsable, denominada cepa M, se introdujo en hospitales de los distritos aledaños, provocando transmisión secundaria. En 2002, la cepa M fue aislada de los dos primeros pacientes con TB XDR identificados en el país.

La evolución de *M. tuberculosis (Mtb)* es modulada principalmente por la aplicación de estrategias de control de la TB, y también por factores socio-económicos, ambientales y biológicos. En particular, la biología del patógeno ha demostrado ejercer una notable influencia sobre la diseminación global de la enfermedad. Estudios moleculares revelaron que existe una insospechada diversidad genética intra-especie en *M. tuberculosis* que permitió diferenciar a *M. tuberculosis* en 6 linajes principales, los cuales poseen afinidad por determinadas regiones geográficas y etnias humanas. Estos linajes se definieron como: Indo-Oceánico (linaje 1), Este Asiatico (linaje 2, incluye el sublinaje Beijing), Este Africano-Indio (linaje 3), Euro-Americano (linaje 4), Oeste-Africano (linaje 5, M. africanum I) y Oeste Africano (linaje 6; M. africanum II). Los linajes 1, 5 y 6 son considerados “antiguos” y los linajes 2, 3 y 4 son considerados “modernos” en relación a la presencia o ausencia de la región del genoma TbD1, ausente en los linajes modernos [54] [55] [56] (Figura 10).

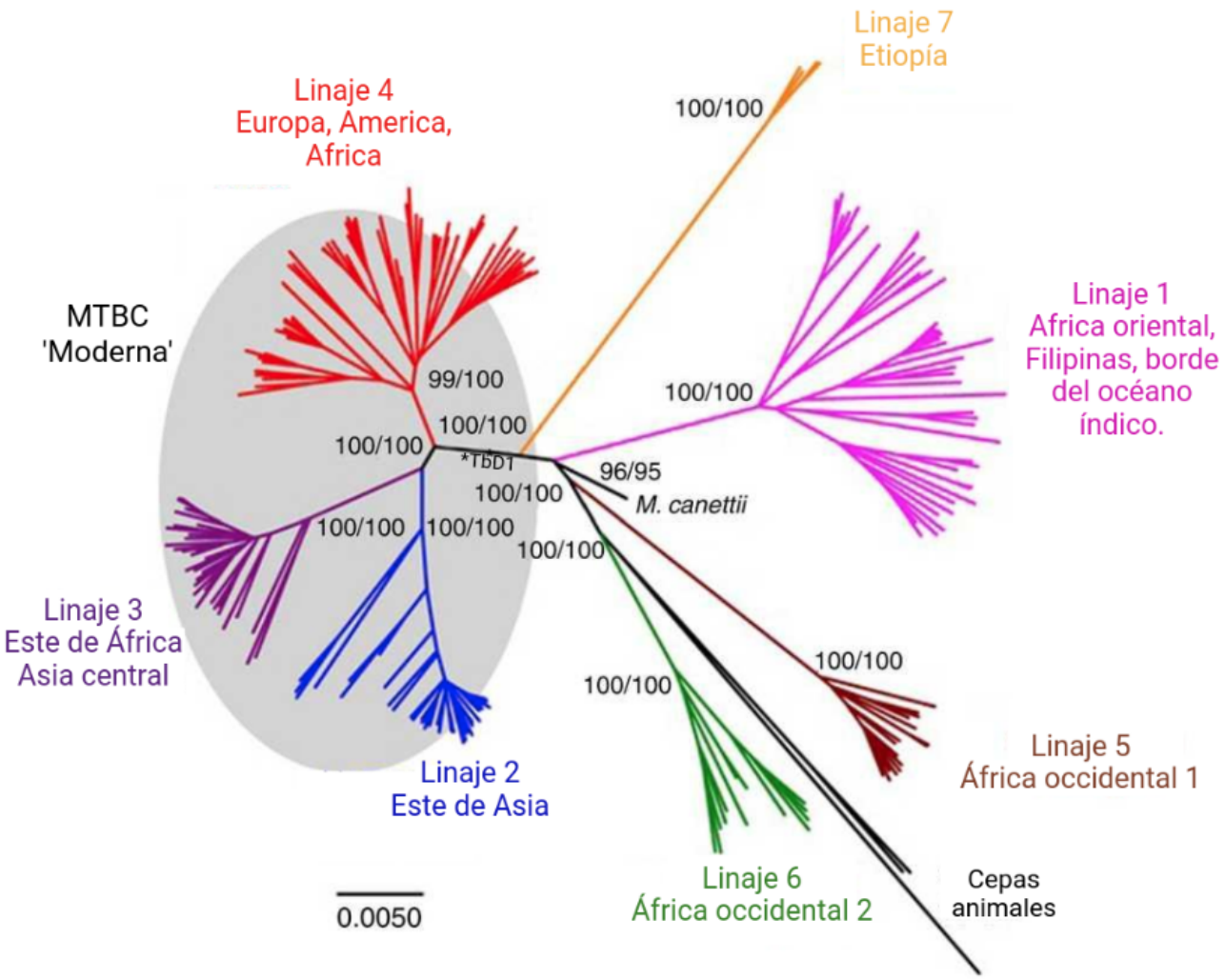



Figura 10: Filogenia global de *M. tuberculosis* en base a LSPs y su relación con la estructura de población global y distribución geográfica. Tomado de Niemann and Supply 2014 [55]

Las cepas de *M. tuberculosis* circulando actualmente en América fueron traídas por los europeos en la colonización, y el linaje Euro-Americano (IV) es el predominante.

Como diferentes linajes de *Mtb* prevalecen en distintas áreas del mundo, la adquisición de resistencia a drogas puede estar guiada por la carga genética pre-existente de las cepas y la heterogeneidad en el mundo puede explicarse por la variabilidad de estas mutaciones [57,58] . En otras palabras, los perfiles genéticos pre-existentes de ciertas cepas de *M. tuberculosis* podrían

asociarse preferencialmente a determinadas mutaciones responsables de resistencia, y el efecto de estas asociaciones podría modular la aptitud biológica de las cepas [59].

## Mapeo y llamado de variantes

A la hora de analizar variantes que confieren resistencia a antibióticos, el objetivo de un experimento de secuenciación no es obtener el genoma de un organismo si no encontrar variantes (diferencias) con respecto a un genoma de referencia. El genoma de *M. tuberculosis* que se utiliza como referencia para la detección de variantes de resistencia, y el utilizado a lo largo de esta tesis, es el de la cepa H37Rv. En la etapa de llamado de variantes, las lecturas se utilizan para identificar aquellos sitios en los cuales el genoma en estudio difiere del genoma de referencia y, en consecuencia, determinar el genotipo de la muestra en cada sitio. En este sentido, habiendo alineado las lecturas de uno o más organismos a un genoma de referencia (mapeo), el llamado de variantes identifica los sitios variables (Figura 11).

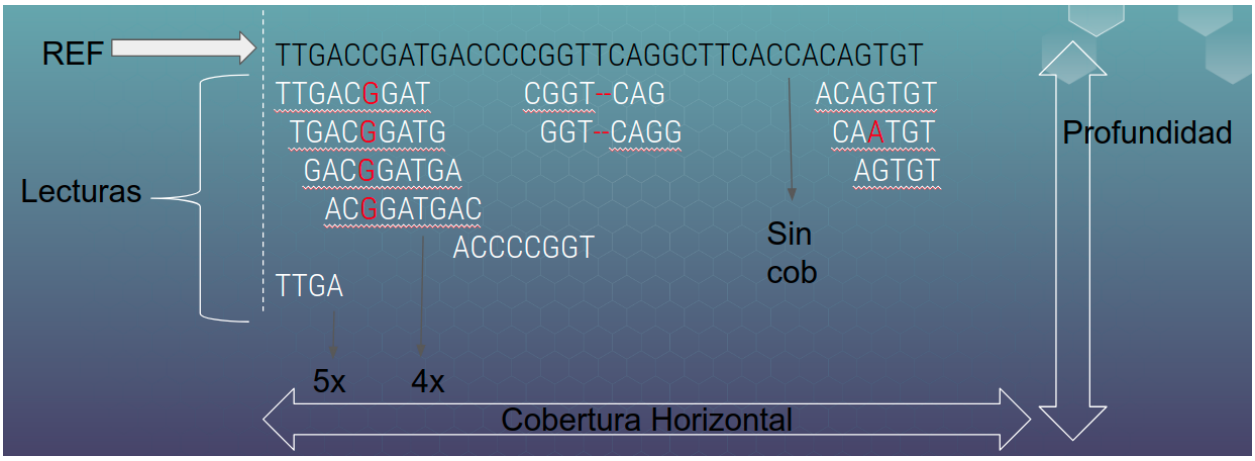



Figura 11: Mapeo contra genoma de referencia y llamado de variantes. En el llamado de variantes, se identifican aquellas bases que difieren del genoma de referencia, marcadas en rojo, y se decide dependiendo de la evidencia si se trata de una variación real o corresponde a un error de secuenciación.

La tasa de error en el reconocimiento de los nucleótidos en una lectura se debe a múltiples factores, incluyendo errores en la identificación de las bases y en el alineamiento. Bajo tales circunstancias, un llamado de variantes preciso no es trivial y frecuentemente hay una incerteza considerable asociada a los resultados, pero se puede reducir y cuantificar la misma [60]. La mayoría de los algoritmos contemporáneos utilizan un marco probabilístico. Las probabilidades del alelo alternativo, incorporan parámetros tales como la calidad de la base, métricas del alineamiento de

la lectura a la referencia, sesgos de lectura de cada hebra, etc. La precisión del alineamiento tiene un rol crucial en la detección de variantes. Lecturas incorrectamente alineadas pueden llevar a errores en el llamado de variantes, por lo que es importante que los algoritmos de alineamiento sean capaces de lidiar con errores de secuenciación, así como con las potenciales diferencias reales (tanto SNPs como indels) entre el genoma de referencia y el genoma secuenciado que se deben a polimorfismos. Además, es importante que los algoritmos de alineamiento produzcan valores de calidad bien calibrados, ya que el llamado de variantes y sus posteriores probabilidades dependen de esos puntajes.

La cantidad de identidad de secuencia requerida entre cada lectura y la secuencia de referencia es determinada por un balance entre precisión y profundidad de lecturas. La elección óptima del número tolerable de nucleótidos no apareados (*mismatches*) puede diferir entre distintos organismos.

El llamado de variantes tiene como objetivo determinar en qué posiciones hay polimorfismos o en qué posiciones al menos una de las bases difiere de una secuencia de referencia.

## Bases de datos de resistencia

Entre las bases que relacionan fenotipos y genotipos, las de resistencia son de particular importancia para esta tesis. En principio, son recopilaciones de trabajos bibliográficos, donde se describen mutantes con polimorfismos y/o genes, donde su presencia/ausencia provoca que dicho organismo sea resistente a una determinada droga. Más allá del mecanismo molecular involucrado, podemos hablar de 2 tipos de entradas en las bases: los genes de resistencia, que se movilizan en plásmidos, fagos, transposones, cuya presencia/ausencia inducen el fenotipo resistente, y por otro lado las variantes (SNPs + inserciones y deleciones cortas), las cuales afectan a genes (o su expresión o traducción) más estables dentro del genoma.

El genoma de *Mtb* no se ve afectado por procesos de transferencia horizontal de ADN, por lo que suele ser bastante estable. Esto hace que la mayoría de las muestras de un mismo linaje tengan prácticamente los mismos genes. En la figura 12 y la tabla 1, podemos ver las drogas más comunes, sus blancos y mecanismos de resistencia.

Existen bases de datos que recopilan variantes y resistencias de todas las bacterias, como CARD [61] y otras enfocadas en *Mtb*, como TBProfiler [62] y TBDR [63]. En la Tabla 2 se observan registros de ejemplo de ese tipo de bases de datos.

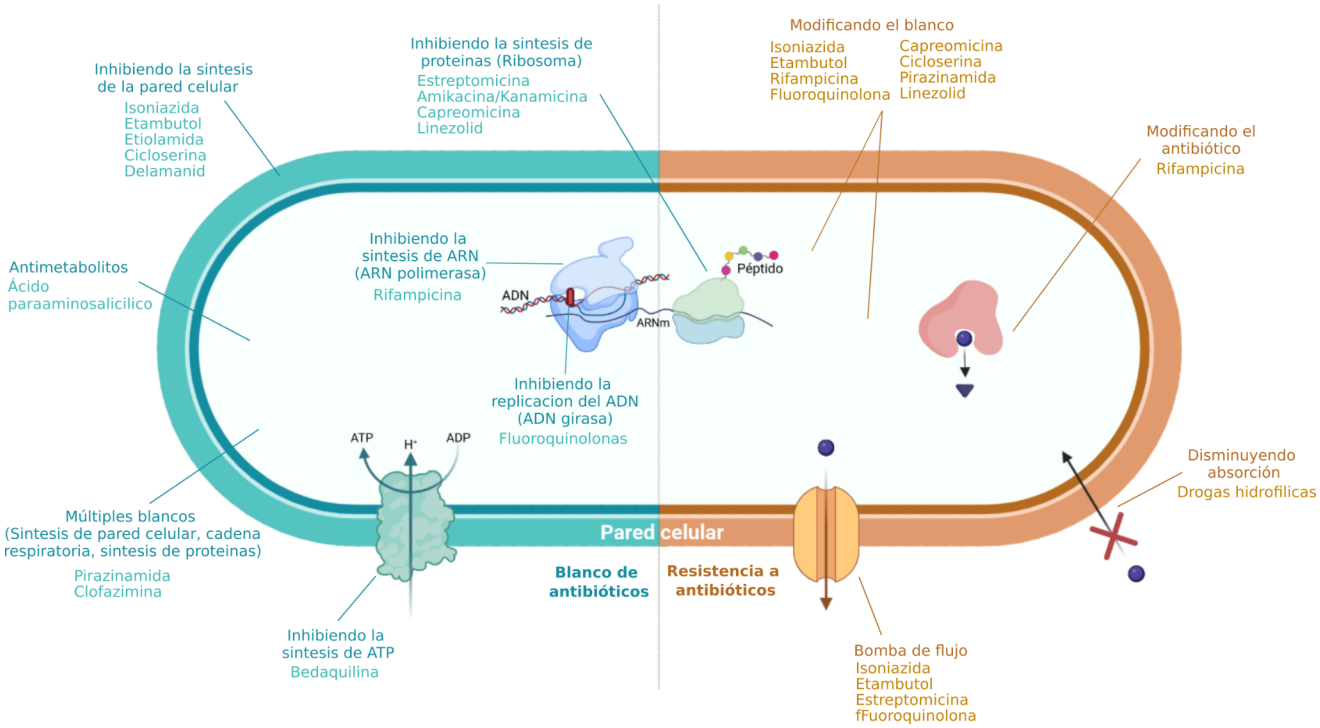



Figura 12: Mecanismos de acción de los fármacos antimicrobianos y mecanismos moleculares de resistencia de *Mtb*. Figura adaptada con Biorender [64].

Tabla 1: Modo de acción y blancos de drogas utilizadas en el tratamiento de TB. En verde se marcan las drogas de primera línea, en azul las de 2da y en naranja las de 3ra. Obtenido de Gagneux, S. et al. 2006 [65]

| Droga | Modo de Acción | Blanco* | Genes relacionados |
|---|---|---|---|
| Isoniazida | Inhibición de la síntesis del ácido micólico | Enoil-ACP reductasa (inhA) | *katG, inhA, Ndh kasA, ahpC, fabG1 (promotor)* |
| Rifampicina | Inhibición de la RNA polimerasa | Subunidad β RNA polimerasa (rpoB) | *rpoB* |
| Pirazinamida | Inhibición de la producción de energía y trans-traducción. | sintasa-I de acido graso, proteína ribosómica S1 | *pncA, rpsA, panD* |
| Etambutol | Inhibición de la síntesis de arabinogalactan | Arabinosil transferasa (Operón embCAB) | *embB* |
| | Inhibición de la síntesis proteica | Subunidad ribosomal 30S | *rpsL rrs* |
| | Inhibición de la síntesis proteica | Subunidad ribosomal 30S | *rrs Eis (promotor)* |
| Amikacina / Kanamicina / Capreomicina | Inhibición de la síntesis proteica | Subunidad ribosomal 30S | *rrs Eis (promotor)* |
| Fluoroquinolonas | Inhibición de la ADN girasa | ADN Girasa | *gyrA gyrB* |
| Etionamida | Inhibición de la síntesis del ácido micólico | Enoil-ACP reductasa (inhA | *ethA, inhA, Ndh, mshA* |
| Ácido paraamino-salicílico | Inhibición de la síntesis de folato | Timidilato sintasa, dihidrofolato sintasa, dihidrofolato reductasas | *thyA, folC, ribD* |
| Cicloserina | Inhibición de la síntesis de peptidoglicano | Alanina racemasa, D-alanina-D-alanina ligasa, L-alanina deshidrogenasa | *alr, ddl, cycA, Ald* |
| Clofazimina | Interferencia con el ciclo redox, provocando la desestabilización de la membrana y la producción de ERON. | NADH dehidrogenasa | *rv0678 rv2535c rv1979c* |
| Linezolid | Inhibición de la síntesis proteica | Dominio V de la subunidad ribosomal 50S | *rrl rplC* |
| Bedaquilina | Inhibición de la síntesis de ATP | ATP sintasa | *atpE* |
| Delamanid | Inhibición de la síntesis de lípido y proteína de la pared celular | Deshidrogenasa, nitroreductasa | *Rv0407 Rv3547* |

*Los blancos mencionados son afectados por la presencia del agente de forma documentada. A pesar de esto, en algunos casos puntuales, se desconoce el mecanismo de acción.

Tabla 2: Ejemplo de entradas en bases de datos de resistencia. En este caso, se ve primero un polimorfismo en el gen *inhA* (blanco de la INH) que causa resistencia a INH y luego una inserción, la cual, causa un cambio del marco de lectura en el gen (*FRAMESHIFT*). En este caso el cambio de marco de lectura en *katG* produce una Catalasa-Peroxidasa truncada. Esta proteína es necesaria para la activación de la INH, por lo que su ausencia genera resistencia a este antibiótico.

| **Locus Tag** | **Gen** | **Droga** | **Posición** | **REF** | **ALT** | **AA POS** | **AA ref** | **AA alt** | **Efecto** |
|---|---|---|---|---|---|---|---|---|---|
| Rv1484 | inhA | ISONIAZID | 118 | G | T | 40 | G | W | G40W |
| Rv1908c | katG | ISONIAZID | 1849 | A | AGT | 617 | L | - | *FRAMESHIFT* |

## Filogenia

La filogenia es la relación de parentesco entre especies o taxones en general [66] y actualmente se utilizan estructuras llamadas árboles para expresarlas. En este marco, se sugiere que todos los organismos guardan relaciones genealógicas, tal que dos organismos cualesquiera tienen un cierto grado de parentesco porque descienden de un antecesor común. Por ello, los organismos comparten características en común, con lo que podemos establecer que aquellos que tienen un antecesor común más reciente presentarán, mayor número de características compartidas que otros organismos que tienen un ancestro común más antiguo. En el contexto de esta tesis, las características utilizadas pueden ser morfológicas o de secuencias, donde estas últimas son las utilizadas en esta tesis.

Llamaremos homologías a las semejanzas causadas por descendencia de un ancestro común, homoplasias a respuestas adaptativas independientes a causas comunes, generalmente ambientales y consideraremos que la mejor filogenia inferida es la que mejor distingue estos 2 tipos.

Para poder realizar una reconstrucción filogenética, se tienen varias metodologías [67]. El método de las distancias, donde para generar un árbol primero se construye una matriz de distancias entre organismos (distancia entre pares, de todos contra todos). La construcción de esta matriz es fundamental (que características utilizar) y frecuentemente puede perderse cierto detalle al construirla. Luego mediante algún método de agrupamiento, por ejemplo UPGMA (*unweighted pair group method with arithmetic mean*) o unión de vecinos (*Neighbor joining*) se construye el árbol a partir de la matriz. El método cladístico, está basado en el principio de parsimonia, considera que la mejor aproximación es aquella reconstrucción que requiera el menor número de cambios evolutivos, minimizando las homoplasias. El método de máxima verosimilitud [67] (maximum likelihood), basado en el concepto de que la verosimilitud (L) de una hipótesis (H),

para un grupo de datos (D), es proporcional a la probabilidad condicional de observar esos datos (D) dado que la hipótesis (H) es correcta: L ∝ P(D/H). En nuestro caso los datos son alineamientos múltiples y las hipótesis son los árboles y un modelo de evolución molecular [68], y se seleccionan los árboles que maximicen el valor de **L**, dados las secuencias del MSA y el modelo de evolución elegido. Finalmente, los métodos bayesianos, que al igual que los métodos de máxima verosimilitud son probabilísticos, pero en lugar de buscar el árbol más probable de acuerdo a los datos observados, produce el mejor conjunto de árboles, dados los datos y el modelo evolutivo especificado. En general, los árboles son calculados por algoritmos MCMC (Markov Chain Monte Carlo), lo cual permite muestrear un gran espacio por su eficiencia computacional y estimar valores de distintos parámetros, como ser reloj molecular, tamaño de la población, etc.

En este contexto, en la presente tesis utilizaremos los MSAs de los genomas *Mtb* XDR, mediante la metodología de máxima verosimilitud para reconstruir su filogenia molecular y caracterizar sus variantes de resistencia y de linaje.

## Tipificación con oligos espaciadores

Las repeticiones palindrómicas cortas agrupadas y regularmente interespaciadas (CRISPR) comprenden una familia de elementos de ADN repetitivos ampliamente encontrados. Estos elementos se identificaron en aproximadamente el 40 % de las bacterias y el 90 % de las arqueas [69]. Los loci CRISPR generalmente consisten en una secuencia líder rica en A/T no codificante y un número variable de repeticiones directas idénticas (DR) intercaladas con un espaciador único. Se han identificado loci CRISPR en varias especies de micobacterias [70], en particular se han encontrado CRISPR largos en *M.tuberculosis, M. bovis* y *M. avium*.

La tipificación de oligonucleótidos espaciadores (*spoligotyping o espoligotipo*) es una técnica basada en PCR para la diferenciación de cepas de MTBC que aprovecha la estructura y el polimorfismo del locus DR. Las cepas se diferencian por la presencia o ausencia de espaciadores individuales en el conjunto completo de 43 espaciadores [71]. Dado que los resultados del spoligotyping pueden presentarse como un sistema binario (presente/ausente), pueden interpretarse, digitalizarse y compararse fácilmente entre diferentes laboratorios [72].

Debido a su simplicidad, formato de resultados binarios y alta reproducibilidad, el *spoligotyping* se usa ampliamente para investigaciones de epidemiología molecular de MTBC [73].

Dentro del linaje Euro-Americano se distinguen varios sub linajes o familias: Latinoamericana & Mediterránea (LAM), Haarlem (H), familia T, familia X y familia S, definidas por e*spoligotipos* en Jagielski, T. et al 2016 [73]. Este tipo de tipificación, no tiene tan buena resolución como el uso de SNPs específicos, y en la tabla 3 puede verse la relación entre muestras de distintos linajes y su espoligotipo.

Tabla 3: Relación entre espoligotipo y Linaje de *Mtb*, obtenido de datos suplementarios de Coll, F. et al 2014 [74]

| **Sub linaje** | **Nombre** | **Espoligotipo** |
|---|---|---|
| 2.2 | East-Asian (Beijing) | Beijing |
| 4 | Euro-American | S;T;X;LAM;H |
| 4.1 | Euro-American | T;H;X families |
| 4.1.1 | Euro-American (X-type) | X family |
| 4.1.1.1 | Euro-American (X-type) | X2 |
| 4.1.1.2 | Euro-American (X-type) | X1 |
| 4.1.1.3 | Euro-American (X-type) | X3;X1 |
| 4.1.2 | Euro-American | T1;H1 |
| 4.1.2.1 | Euro-American (Haarlem) | T1;H1 |
| 4,2 | Euro-American | LAM7-TUR;H3;H4;T1 |
| 4.2.1 | Euro-American (Ural) | H3;H4 |
| 4.2.2 | Euro-American | LAM7-TUR;T1 |
| 4.2.2.1 | Euro-American (TUR) | LAM7-TUR |
| 4,3 | Euro-American (LAM) | LAM |
| 4.3.1 | Euro-American (LAM) | LAM9 |
| 4.3.2 | Euro-American (LAM) | LAM3 |
| 4.3.2.1 | Euro-American (LAM) | LAM3 |
| 4.3.3 | Euro-American (LAM) | LAM9;T5 |
| 4.3.4 | Euro-American (LAM) | LAM11-ZWE;LAM9;LAM1;LAM4 |
| 4.3.4.1 | Euro-American (LAM) | LAM1 |
| 4.3.4.2 | Euro-American (LAM) | LAM11-ZWE;LAM9;LAM1;LAM4 |
| 4.3.4.2.1 | Euro-American (LAM) | LAM11-ZWE |
| 4,4 | Euro-American | S;T1;T2 |
| 4.4.1 | Euro-American | S;T1 |
| 4.4.1.1 | Euro-American (S-type) | S |
| 4.4.1.2 | Euro-American | T1 |
| 4.4.2 | Euro-American | T1;T2 |
| 4,5 | Euro-American | H3;H4;T1 |
| 4,6 | Euro-American | LAM10-CAM;T2 |
| 4.6.1 | Euro-American (Uganda) | T2-Uganda;T2 |
| 4.6.1.1 | Euro-American | T2-Uganda |
| 4.6.1.2 | Euro-American | T2 |
| 4.6.2 | Euro-American | LAM10-CAM; T3 |
| 4,7 | Euro-American (mainly T) | T1;T5 |
| 4,8 | Euro-American (mainly T) | T1;T2;T3;T4;T5 |

| | | |
|---|---|---|
| 4,9 | Euro-American (H37Rv-like) | T1 |

# Materiales y Métodos

## Selección de cepas

La selección de casos se realizó en conjunto con los profesionales de salud de las instituciones que colaboraron en la presente tesis: el Hospital de Infecciosas “Francisco Javier Muñiz” y de la Administración Nacional de Laboratorios e Institutos de Salud - “Dr. Carlos Malbrán”. Se iniciaron cultivos de 120 aislamientos que corresponden a la totalidad de casos reportados de pacientes con *Mtb* XDR entre los años 2006 y 2011 de los cuales en 49 se logró extraer ADN suficiente para secuenciar (6 fueron descartadas como se explicará más adelante). El listado puede verse en el Anexo 1. En todos los casos la interacción con el paciente fue exclusiva del médico y la evaluación para su inclusión en el proyecto se realizó en función de la recomendación del mismo y el análisis de la historia clínica. En todos los casos las muestras fueron anonimizadas y tomadas del cepario correspondiente a cada hospital. Las mismas fueron secuenciadas en el equipo Illumina Miseq del Instituto Malbrán. Finalmente, la caracterización como *Mtb* XDR fue realizada por los hospitales e institutos que proveyeron las muestras.

## *Pipeline* del procesamiento para el llamado de variantes de resistencia

Para procesar cada una de las muestras, se siguió el protocolo que se detalla en la Figura 13. Primero se verifica la calidad de las lecturas de cada experimento de secuenciación utilizando FastQC versión v0.11.5 [75] y se recortan las bases afectadas por sesgos y baja calidad con el programa Trimmomatic [76] . Seguido de eso, se alinean las lecturas de cada cepa contra el genoma de referencia H37Rv (NCBI NC_000962.3) utilizando BWA [77] y procesándolo en formato BAM con Samtools [78]. Luego se realiza el llamado de variantes utilizando GATK [77,79] y finalmente se anotan las variantes con SnpEff [80].

Para determinar las resistencias genotípicas de cada cepa, se cruzaron las variantes obtenidas en el punto anterior con una base de datos propia que combina las variantes presentes en TBProfiler [62], TBDream [81] y CARD [61] con datos bibliográficos.

Esto se realizó con scripts desarrollados en el laboratorio durante esta tesis.

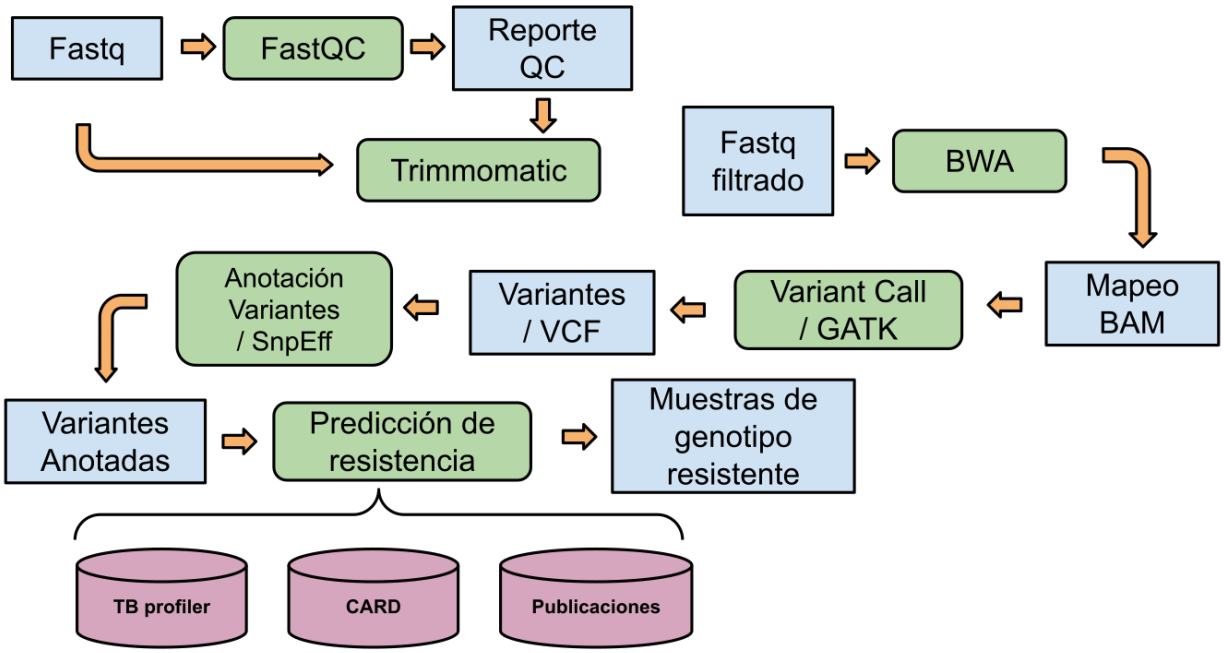



Figura 13: *Pipeline* para el procesamiento de cada muestra (FastQ) hasta generar el listado de variantes preliminar. En verde se ven los procesos y en azul los resultados del procesamiento.

## Análisis filogenético de los aislamientos XDR circulantes en Argentina

Para analizar el conjunto de secuencias genómicas, se optó por realizar un análisis filogenético siguiendo los pasos que se indican en la Figura 14. Primero fue necesario ensamblar las secuencias de cada muestra. Eso fue realizado aplicando las variantes obtenidas en cada una de ellas al genoma de referencia, utilizando Samtools. Luego se enmascararon las zonas de baja cobertura, y las regiones repetitivas IS6110, PGRSs, CRISPRs y VNTR, dado que en dichos lugares suelen ocurrir muchos errores de secuenciación/mapeo. Después se realizó un alineamiento múltiple con todas las secuencias utilizando el programa MAFFT, con parámetros por defecto. Antes de realizar la filogenia, se calcularon las opciones óptimas (modelo evolutivo, sustitución, frecuencias y otros pesos) para generar la misma, utilizando ModelTest [82]. Finalmente utilizando el MSA se realizó un árbol de ML utilizando RaXML (Stamatakis 2014), con los parámetros determinados por jModelTest.

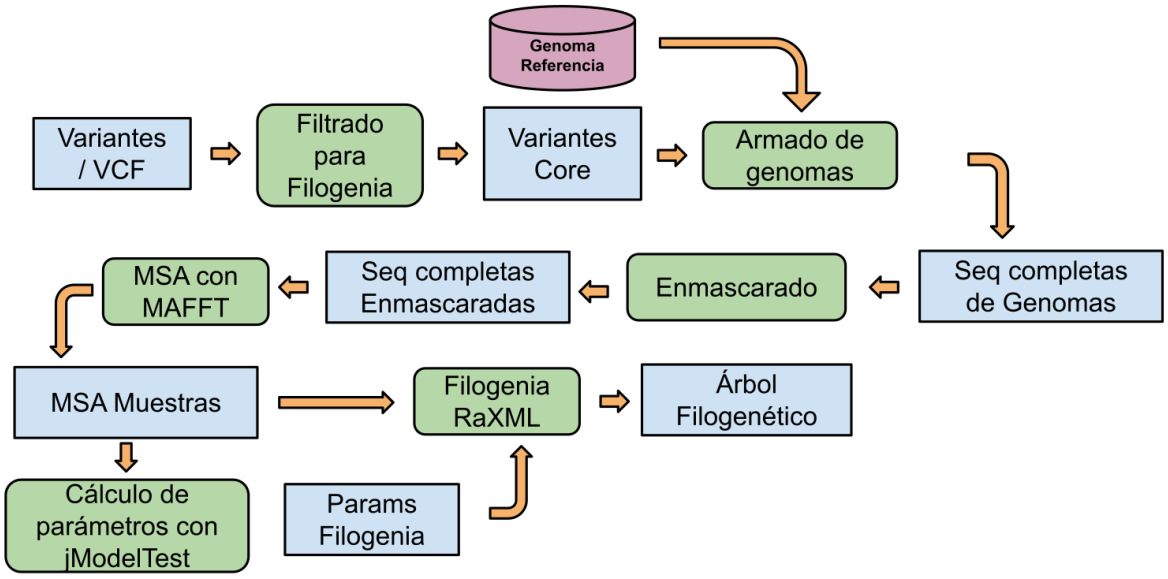



Figura 14: *Pipeline* de análisis comparativo. Va desde las variantes de cada muestra, hasta la obtención de un árbol filogenético que las compara. En verde se ven los procesos y en azul los resultados del procesamiento.

Finalmente, para agregar como columnas complementarias a la filogenia, se predijo el linaje de *Mtb* y espoligotipo de cada muestra in silico, utilizando la herramienta TBProfiler [62]

# Resultados y Discusión

## *Pipelines* de procesamiento

Utilizando el lenguaje python, se implementaron los *pipelines* de la Figura 13. El procesamiento de aislamientos se dividió en 3 scripts:

- process_sample.py: procesa las lecturas illumina hasta obtener variantes crudas (sin anotar)
- process_group.py: toma todas las variantes crudas, arma un único archivo multimuestra de las mismas y las anota (cruza la información con la anotación del genoma)
- resistance_analysis.py: toma el fasta multimuestra y lo cruza contra los datos de resistencia. Genera como resultado una lista de registros, que poseen la información de la variante, la droga asociada y a que variante corresponde.

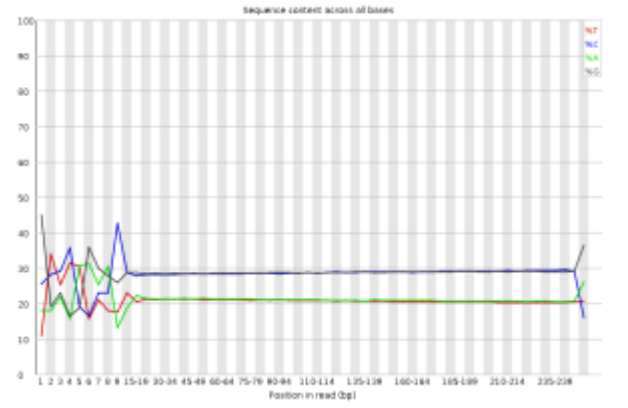
Todos los programas externos son llamados a través de *Docker*, una plataforma para distribuir y utilizar programas de manera transparente, con lo cual los requerimientos de instalación son mínimos: tener Python y Docker instalado. El código y las instrucciones de uso se dejó disponible online en un repositorio[2] de GitHub.

El procesamiento del grupo de muestras, que toma los vcfs crudos y genera una filogenia, no se desarrolló de manera automatizada, ya que el mismo es un proceso exploratorio. Esto quiere decir que se realizan varias selecciones de variantes (con y sin filtro de zonas repetidas, usar solo SNPs o también inserciones y deleciones cortas, usar un programa para predecir los parámetros de filogenia o probar los utilizados en bibliografía para Mtb en trabajos similares). En este trabajo finalmente se siguieron los pasos de la Figura 14, ya que fue la que agrupa correctamente las muestras según su linaje y espoligotipo.

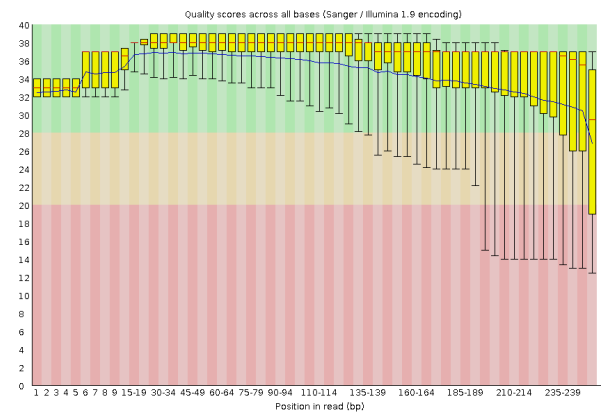
[2] https://github.com/florenciacastello/tb_resistance

## Mapeo, llamado de variantes y anotación

La calidad de las secuencias obtenidas por FastQC fue más o menos uniforme en todas las muestras, se muestra un ejemplo del reporte en la Figura 15.




A) Proporción de nucleótidos leídos en cada posición. Dado que la fragmentación es al azar, teóricamente, debería ser constante. En general existe un sesgo en las primeras bases de cada lectura debido a un sesgo técnico en los equipos de secuenciación Illumina y a los kits de armado de librerías.




B) Puntuación del promedio de la calidad de las bases leídas por posición. El mismo generalmente disminuye hacia el final de las lecturas debido a una pérdida de la actividad de la polimerasa durante la secuenciación.

Figura 15: Reporte parcial del programa FastQC. El mismo genera varios gráficos / distribuciones que sirven para evaluar la calidad de las lecturas, base por base, cuya información es valiosa para aplicar recortes, filtrados u otros procesamientos, de manera tal de no hacer análisis sobre información errónea.

Luego de evaluar la calidad, se realizó un recorte de las lecturas con Trimmomatic, usando 2 filtros importantes. El primero se utiliza para corregir los artefactos introducidos por la técnica de secuenciación debido al sesgo de lectura que poseen algunos kits de armado de bibliotecas de ADN (Figura 15 A), que se fragmentan preferencialmente en algunas regiones. Para corregir este error, se recortan los primeros 15 nucleótidos de cada una de las lecturas. Por el otro, al final de las secuencias (Figura 15 B), las polimerasas comienzan a desincronizarse y la calidad de lectura baja. En este caso se aplicó un filtro para recortar las bases de baja calidad, desde el “final” de la lectura hasta que llega a un conjunto de bases de buena calidad (puntuación de calidad >20).

Seguidamente se procedió al mapeo de lecturas sobre H37Rv utilizando el programa BWA. Luego de esto es importante chequear qué porcentaje de las lecturas pudieron ser alineadas contra el genoma y qué cobertura y profundidad se obtuvo, ya que estas son medidas importantes

para la calidad del llamado de variantes. En la Tabla 4 se ejemplifica con el resultado de una de las muestras.

Tabla 4: Estadísticas del alineamiento de una de las muestras. El programa Samtools, en particular el subcomando “flagstat” arroja esta información sobre el mapeo. En este caso, estamos viendo estadísticas sobre un archivo BAM, el cual solo posee lecturas mapeadas.

| **Estadística** | **Descripción** | **Cantidad de lecturas** |
|---|---|---|
| Total | Cantidad total de lecturas alineadas | 1896464 |
| Mapean en más de un lugar | Lecturas que tienen un buen alineamiento en más de un lugar del genoma | 10552 |
| Mapeados | Lecturas mapeadas, en este caso, coincide con el total, ya que el resto es eliminado. En otros análisis no realizados aquí, lo que no se mapeó puede analizarse para determinar si existe contaminación o presencia de otro/s organismo/s en las muestras | 1896464 |
| Pareados correctamente | Cuando ambas lecturas de un mismo fragmento pudieron ser alineadas | 1866572 |
| Singletons | Cuando sólo una lectura de un fragmento pudo ser alineada | 10968 |

A partir de esos resultados se calculó la profundidad promedio del mapeo de las distintas muestras (bases alineadas / longitud del genoma). Debido a su baja profundidad promedio (< 15X) y/o baja cobertura horizontal (muchas partes del genoma sin leer) se descartaron los datos de 6 de las 49 aislamientos. Con lo cual todo el resto de los análisis se realizará con **las 43 restantes**. En general la profundidad mínima recomendada depende del análisis a llevar adelante. Si observamos las recomendaciones del proveedor de la tecnología utilizada, ilumina, para la secuenciación del genoma humano recomienda 30X [83] (estimado a generar por el secuenciador), nosotros tomamos la mitad del valor, ya que estamos midiendo las lecturas recortadas y alineadas. También es importante analizar la profundidad a lo largo del genoma, ya que la misma no es uniforme, por lo que en general, zonas de baja cobertura no solo se traducen en incertidumbre debido a los errores, también implica que muchas regiones del genoma directamente no fueron secuenciadas. Tomamos como límite inferior 10 por base, es decir, si una posición no fue leída al menos 10 veces, se marca como una zona de baja cobertura y la información se enmascara (es decir no se toma en cuenta para el análisis de llamado de variantes). En el histograma Figura 16 se representa la profundidad promedio en todos los aislamientos analizadas.

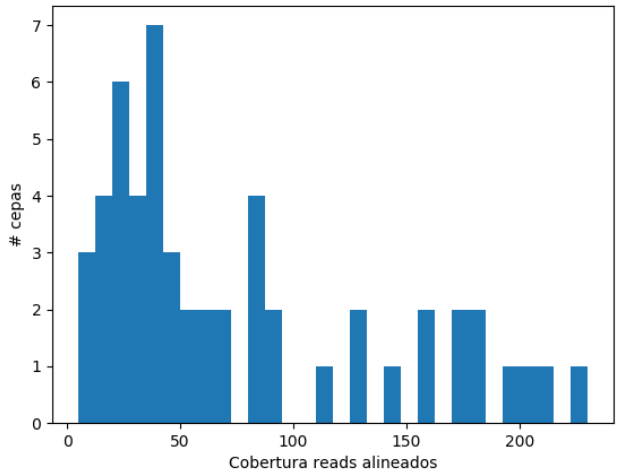



Figura 16: Histograma de la profundidad promedio de las muestras analizadas.

A continuación, se ejecutó el llamado de variantes, utilizando una ploidía de 2. Si bien este concepto no aplica para *Mtb* (por ser haploide), esto le permite a GATK interpretar un “alelo” minoritario (en lugar de tomarlo como un error), lo cual se puede interpretar como más de una población dentro de la misma muestra [3,84]. La presencia de más de una población en la muestra obtenida de un paciente puede deberse a principalmente a evolución intra paciente o co-infección. Determinar el motivo queda fuera del alcance de este trabajo. En el caso de que se detecte más de un “alelo” en una posición, se utilizó el que aparecía en mayor proporción (> cantidad de lecturas que lo respaldan y al menos un 75% de las veces), dado que es la más probable que pertenezca a la población mayoritaria. Este tipo de convenciones también fueron tomadas en el armado de genomas consenso para SARS-CoV-2 [85] y detectar SNPs de baja frecuencia Mtb [3]. Finalmente se anotaron las variantes utilizando la anotación de NCBI de H37Rv, los resultados obtenidos pueden verse en la Tabla 5. Dado que aproximadamente el 90 por ciento de la secuencia genómica de las bacterias son zonas codificantes, se encuentran muchas más variantes génicas (sinónimas y no sinónimas) que intergénicas. Por otro lado, la mayor parte de las variantes del alto impacto (se asume se pierde la función de la proteína), como ser pérdida de codones start/stop o frameshifts (corrimiento del marco de lectura), son muchos menos frecuentes.

Tabla 5: Distribución de variantes por región (obtenida de la anotación de SnpEff)

| Tipo de Variante | Número de variantes respecto a H37Rv para los 43 aislamientos analizadas |
|---|---|
| Corrimiento del marco de lectura | 106 |
| Pérdida de codón de inicio | 27 |
| Codón de terminación prematuro | 263 |
| Mutación Sinónima | 81069 |
| Intergénica | 5991 |
| Cambio de Aminoácido (mutación no sinónima) | 33138 |

## Mutaciones de resistencia

La isoniazida es una prodroga que se activa luego de que la enzima *katG* actúe sobre la misma. De esta manera, uno de los mecanismos de resistencia que actúan sobre la misma son mutaciones que bajen la expresión o la actividad de dicha enzima. Por otro lado, el blanco de acción de la INH es *inhA* cuyas variantes pueden también desencadenar resistencia a esta droga. Particularmente se ha reportado que mutaciones en la región promotora del operón *inhA* ( *fabG1*, *inhA* y *hemZ*) pueden ser causantes de una sobreexpresión de esta enzima, desencadenando la resistencia a INH. De las variantes causantes de resistencia a INH en los aislamientos analizados, 31/43 corresponden a *katG* Ser315Thr (Serina por Treonina en la posición 315 de *katG*), 13/43 promotor *fabG1-inhA* −C15T (mutación en el promotor del operón *inhA* C por T, que se encuentra 15 nucleótidos antes del gen *fabG1*) y 3/43 tenían ambas. Este resultado se encuentra en concordancia con lo observado previamente a nivel local en Monteserin, J. et al. 2017 [86]. En dicho trabajo se muestra que estas 2 mutaciones explican la resistencia a INH del 80% de los 362 aislamientos resistentes en Argentina analizados. En dicho trabajo también se analizaron aislamientos XDR de las cuales el 70 % presenta la variante *katG* Ser315Thr sin *fabG1-inhA* −C15T coincidentemente con nuestros resultados También, se detectó las variante de resistencia *inhA* ile194Thr y *fabG1-inhA* C-17T y C-8T, pero cada una en un solo aislamiento.

En el caso de la RIF, que actúa inhibiendo la síntesis de ARN mensajero, a cargo de la ARN polimerasa (gen *rpoB*) [87], su mecanismo de resistencia implica mutaciones que modifican el

blanco de acción que conllevan la pérdida de afinidad con la droga. Las mutaciones encontradas para este gen fueron Ser450Trp (28/43), Asp435Val (10/43), Gln432Pro (3/43), Asp435His (2/43) y Asp435His junto con His445Asp en el aislamiento 24830. También se encontraron mutaciones reportadas en *rpoC* (Val483Gly, Trp484Gly, Phe452Ser) y *rpoA* (Val183Ala), pero siempre acompañadas de *rpoB* Ser450Trp. Por un lado vemos que los resultados coinciden con la vigilancia local realizada en Imperiale et al. 2019 [88], donde se estudiaron diversas muestras de *Mtb* MDR circulantes en Argentina entre Junio de 2010 y Agosto de 2018 mediante técnicas moleculares (amplificación y Sanger). En la misma, se detectó que la mayoría de las muestras MDR poseía la variante *rpoB* Ser450Trp (45/77). Previamente se ha observado que aislamientos MDR que poseen esta mutación presentan altos valores de CIM a RIF [89]. La siguiente variante más frecuente, *rpoB* Asp435Val, también fue reportada [90], pero detectada en baja proporción en cepas Beijing. Hay que tener en cuenta que la proporción de la variante Beijing en nuestro país es muy baja [91]. Sin embargo, un estudio más reciente [92], que recopiló aislamientos resistentes a RIF de amplia distribución geográfica (Brasil, India, Vietnam, Sudáfrica, China y Pakistán), reportó que la posición *rpoB* Asp435 es una de las posiciones más mutadas, en particular en China e India. Por último cabe remarcar que en dicho trabajo se han reportado mutaciones en ese posición de distintos aminoácidos, mientras que en los aislamientos analizados en el presente trabajo solo se vio Asp435Val.

En el caso particular de la resistencia a rifampicina, muchas veces se reportan las variantes en el gen de *rpoB* utilizando las posiciones aminoacídicas del genoma de *E. coli* [88] por una cuestión histórica, dado que las primeras asociaciones variante/resistencia se reportaron en dicha bacteria. En este sentido es importante aclarar que para compatibilizar las posiciones con las de Mtb H37Rv se siguió la Tabla 6.

Tabla 6: Mutaciones en rpoB, tanto en *E. coli* como en H37Rv [89] y su fitness. La columna “RIF CIM (mg/litro) > 400” muestra los resultados de Jagielski, T. et al. [89] indicando si los aislamientos que encontró con esa variante poseían una CIM mayor a 400 mg/l

| **Variante** | **aa (*H37Rv*)** | **aa (*E. coli*)** | RIF CIM (mg/litro) > 400 [89] | Fitness relativo a cepa Haarlem [93] |
|---|---|---|---|---|
| C1349G | Ser450Trp | Ser531Trp | SI | 0.67 |
| C1349T | Ser450Leu | Ser531Leu | SI | 0.84 |
| A1304T | Asp435Val | Asp516Val | SI | no reportada |

PZA es una prodroga activada por la pirazinamidasa (*pncA*). Su mecanismo de resistencia está asociado con la falta de actividad de la misma. En este estudio las 2 mutaciones más frecuentes resultaron Gln10Arg (reportada como la gran causante de resistencia en el brote de cepa M [94] ) y Arg154Gly, que solo representan un tercio del total de las XDR analizadas (18/43). El resto de las variantes se encuentra solo una o dos veces por muestra. Esto coincide con el hecho de que la región de *pncA* es la más polimórfica del genoma [95].

El etambutol actúa inhibiendo la transferencia de los ácidos micólicos a la pared celular y el mecanismo de resistencia está asociado con variantes en *embB*. Las tres variantes más frecuentes en el estudio fueron Met306Ile (19/43), Gly406Asp(10/43) y Gln497Arg (2/43). Se encontraron otras mutaciones dentro del gen, pero solo aparecían en una única muestra. Además en la secuencia promotora, se encontraron dos mutaciones asociadas a resistencia en 3 aislamientos distintos pero siempre acompañadas de una mutación dentro del gen. Respecto a antecedentes en Argentina, la posición Gly406 pero con la variante Gly406Ala en lugar de la reportada en este trabajo (Gly406Asp), fue la mayor causa de resistencia en el brote de la cepa M [94]. En dicho brote, Met306Ile fue encontrada en solo un aislamiento (resistente) y asociada en otros trabajos con un CIM medio de 8 $\mu$g/ml [96]. En otro estudio [97] también se midieron valores de CIM de 20$\mu$g/ml para Met306Ile y Met306Val de 40 $\mu$g/ml, sin embargo esta última fue encontrada en un solo aislamiento de XDR en Argentina.

La estreptomicina, al ser un aminoglucósido, impide el inicio de la síntesis proteica, uniéndose a las subunidades ribosomales 30S, específicamente en las proteínas ribosomales S12 y 16S ARNr codificadas por los genes *rpsL* y *rrs*, respectivamente. La unión al ribosoma interfiere con la elongación de la cadena peptídica [98]. La resistencia fenotípica se asocia principalmente a mutaciones en *rrs* [99]. La variante encontrada con mayor frecuencia en las muestras analizadas fue *rrs* A1401G (32/43). La misma fue la principal causa de resistencia en el brote de la cepa M [94]. En el gen *rpsL* se detectaron las variantes Lys43Arg (típica en aislamientos Beijing mono resistentes a estreptomicina [100] ) y Lys88Arg (también característica de la cepa Beijing, tanto en mono resistentes como MDR [100] ), en 4 y 3 muestras respectivamente.

Las fluoroquinolonas (levofloxacina, ofloxacina, moxifloxacina), actúan por la inhibición de la ADN girasa o la topoisomerasa IV. Su mecanismo de resistencia está relacionado con mutaciones en las proteínas blanco (genes *gyrA* y *gyrB* que codifican para las subunidades de la girasa y *parC* y *parE*, que codifican para las subunidades de la topoisomerasa), lo cual reduce su afinidad con el fármaco [101]. En los aislamientos estudiados en este trabajo sólo se detectaron

variantes reportadas en *gyrA* (38/43) y en *gyrB* (6/43). La mayor parte de las mutaciones se encontraba en la posición *gyrA* Asp94, cambiando el aminoácido por alanina o asparagina. A nivel global, estas mutaciones son las más comunes en la resistencia a Levofloxacina [102]. Mientras que a nivel local se detectaron en muy baja proporción en la cepa M y ya fueron reportadas en aislamientos clínicos resistentes a LVR en Buenos Aires durante el 2006 - 2010 en el Hospital Dr. Antonio Cetrangolo[103]. La segunda mutación más frecuente en *gyrA* fue Ala90Val, la cual también fue reportada en dicho estudio [103]. Es importante notar que se han reportado valores más altos de CIM para la fluoroquinolona levofloxacina para las variantes *gyrA* 94 que para las variantes *gyrA* 90 [103]. Un estudio realizado recientemente por nuestro laboratorio [84], que consistió en el seguimiento de un paciente para estudiar el fenómeno microevolutivo, mostró la coexistencia de ambas mutaciones en poblaciones diferenciadas de *Mtb* a lo largo del tratamiento. Mientras que la población que presentaba la mutación en *gyrA en la posición* 90 fue disminuyendo y finalmente desapareciendo, los clones que portaban la mutación D94H se seleccionaron dentro de la población bacilar. Las variantes *gyrA* Ala74Ser y Gly88Cys fueron encontradas en 4/51 aislamientos y una de ellas coexistiendo con Asp94Ala. En cuanto a *gyrB*, las mutaciones encontradas fueron Arg446Cys (4/43 total, 1 aislada y, 2 en compañía de *gyrA*94 o *gyrA*90), Ala504Val 2/43 y Asp461Asn (1/43). Todas estas variantes fueron reportadas previamente en la cepa M, pero nuevamente, en muy baja proporción.

Los aminoglucósidos inyectables de segunda línea (kanamicina, amikacina, capreomicina) al igual que la estreptomicina (STR) inhiben la síntesis proteica a través del bloqueo del ribosoma. La mutación *rrs* A1401G es la alteración más reportada en muestras resistentes a KAN y se ha asociado con resistencia cruzada a AMK, KAN y CAP [104]. Al ser aminoglucósidos como la estreptomicina, comparten su blanco con *rrs*, con lo cual las observaciones realizadas para STR se mantienen. Para KAN en particular, se detectaron mutaciones en el promotor del gen *eis* G-12A y C-10T, relacionadas con menores niveles de resistencia a la de las variante *rrs* A1401G [105].

El antibiótico oral etionamida (ETH) es una prodroga activada por la monooxigenasa *ethA*. Este fármaco inhibe la síntesis de ácidos micólicos [106]. Está reportado que alteraciones en el promotor del gen *fabG1*, involucrado en el primer paso de reducción del ácido micólico, otorga resistencia a esta droga [107]. La variante reportada más encontrada en nuestras muestras fue la del promotor *fabG1* C-15T en 13/43 muestras. La misma está frecuentemente asociada con el fenotipo resistente. Esta variante es de penetrancia completa, es decir que su presencia produce un fenotipo resistente. Sin embargo existen aislamientos que no portan esta mutación y tienen mecanismos de resistencia alternativos [107]. Por último, se encontraron las variantes *fabG1* C-17T y

C-8T e *inhA* Ile194Thr, en una sola de las muestras estudiadas. Una última mutación que vale la pena destacar es la mutación *rplC* Cys154Arg, reportada como mutación de resistencia para linezolida, un antibiótico de tercera línea, en dos aislamientos correspondientes al grupo T2.

## Análisis Filogenético

Luego se generó un árbol local, (figura 18), que es un subárbol del global, tomando solo los aislamientos XDR circulantes en nuestro país para estudiar si las mutaciones asociadas a resistencia se distribuyen homogénea o heterogéneamente en los distintos grupos.

Tabla 7: Frecuencia de Espoligotipos de Mtb en BsAs. N=816, uno por paciente. Tomado de Monteserin J. et al. 2018 [91]

| **Subfamilia** | **%** |
|---|---|
| T | 35.9 |
| LAM | 33.2 |
| Haarlem (H) | 19.5 |
| S | 3.2 |
| X | 1.5 |
| U | 0.7 |
| Beijin | 0.2 |
| Bov | 0.2 |
| Otros | 0.2 |
| Desconocido | 5.3 |

El Grupo H es corresponde a todos los aislamientos caracterizadas molecularmente como H2 o H3. Solo la muestra 10900 está caracterizada como H3 pero no está incluida en este grupo. Dentro de él, se observan las mismas mutaciones de resistencia, excepto para las fluoroquinolonas (Tabla 8). A nivel global se agrupan con las muestras de la cepa M, y están agrupadas en la misma rama que las muestras representativas de los linajes 4.1.2.1, 4.1.2 y 4.1 (la cepa M se reportó como del linaje 4.1.2.1 [94]). Los aislamientos 1074, 10900 y 17270, no se asignaron al grupo H, pero también se asocian con el sublinaje 4.1.2, sin embargo, no comparten ningún SNP de resistencia entre los 3. En la misma rama, se encuentran las muestras 20394 y 24830, con espoligotipo ambiguo entre T2-T3 la primera y espoligotipo “X” la segunda (a su vez agrupada con un representativo del lineage correspondiente 4.1.1,Tabla 3). Las mismas comparten

la mutación para EMB *embB* Met306Ile y 24830 es la única que posee las variantes en el promotor de *fabG1* C-8T y *rpoB* His445Asp.

Tabla 8: Comparativa de mutaciones de resistencia del Grupo H vs Cepa M [94]

| **Droga** | **Mutación** | **Frecuencia (M 225 muestras, H 8 muestras)** | **¿Existe la mutación en un grupo distinto al H?** |
|---|---|---|---|
| INH | *katG* S315T | M: 90% ><br>H: 100% | no |
| RIF | *rpoB* S450L<br><br>*rpoB* Q432K<br>*rpoB* D435V<br>*rpoB* H445Y/R | M: 90% ><br>H: 100%<br>M: <1% H: 0%<br>M: <1% H: 0%<br>M: <1% H: 0% | no<br><br>no<br>Si, en Grupo LAM5<br>no |
| EMB | *embB* G406A<br><br>*embB* M306I<br>*embB* M306V | M: 90% ><br>H: 100%<br>M: <1% H: 0%<br>M: <1% H: 0% | no<br><br>Si, Grupo LAM5 y otros |
| PZA | *pncA* AQ10P | M: ~85%<br>H: 100% | no |
| KAN | *rrs* 1401A>G (93%) | M: ~99%<br>H: 100% | no |
| FLQ | *gyrB* A504V<br><br>*gyrA* A90V<br>*gyrB* R446C<br><br>*gyrA* D94A<br><br>*gyrA* D94G/N<br>*gyrB* A504V<br><br>*gyrB* R446S<br>*gyrA* L105R<br>*gyrB* D461V<br>*gyrA* R292G | M: <1%<br>H: 25%<br>M:<1%<br>M: <1%<br>H: 12.5%<br>M:0%<br>H: 25%<br>M: <1% H: 0%<br>M: <1%<br>H: 12.5%<br>M: <1% H: 0%<br>M: <1% H: 0%<br>M: <1% H: 0%<br>M: <1% H: 0% | no<br><br>no<br>no<br><br>no<br><br>no<br>no<br><br>Si, Muestra 18712 -> R446C<br>no<br>no<br>Si, Muestra 1074 |

El grupo LAM5 agrupa aislamientos caracterizados como tales. El aislamiento 11880 es el único tipificado como LAM5 que queda fuera de este grupo. El mismo presenta variantes de resistencia diferentes a aquellas que portan las caracterizadas en el grupo LAM5. Todos los aislamientos que forman parte del grupo LAM5 presentan los mismos genotipos de resistencia para INH, RIF, PZA y EMB. En cuanto a los aminoglucósidos de segunda línea, el único aislamiento que no posee la variante *rrs* C1401T es 13429. Tanto 13429 como 13431 y 18712 son los únicos aislamientos que no presentan la variante *gyrA* Asp94Ala para fluoroquinolonas dentro del grupo. El árbol filogenético es consistente con la hipótesis de que las variantes *gyrA* Asp94Ala

y *rrs* C1401T se hayan incorporado en un ancestro común a todas los aislamientos LAM 5 que las portan, *rrs* C1401T podría haberse perdido en un evento evolutivo posterior en el aislamiento 13429.

El grupo T2, se compone de 3 aislamientos caracterizados experimentalmente como T: 22372, 16306 y 20246. Todas ellas conforman un mismo grupo monofilético con el genoma representativo de 4.3.3, que se corresponde con los espoligotipos LAM9 y T5 (tabla 3), con lo cual aunque el grupo de espoligotipo concuerda con el correspondiente para ese linaje (T), no lo hace con el subgrupo (T1). En cuanto a las variantes de resistencia, los aislamientos de este grupos son de los pocos que no poseen las variantes de resistencia a INH en *katG*, sino que el mecanismo de resistencia se da a través de de variantes en el promotor de *fabG1***.** A parte de dicha mutación, también comparte *rrs* A1401G para aminoglucósidos, *embB* Met306Ille para EMB y *gyrA* Asp94Ala, compartido también con el grupo LAM5, pero a diferencia de ese grupo, no se detectaron mutaciones de resistencia a PZA.

Respecto al grupo LAM3, vemos en el árbol global que se agrupa con las muestra del sublinaje de 4.3.2 (la misma con la que es caracterizado *in silico*) y en particular, el aislamiento 25203, está cercana al brote de la cepa Ra (caracterizada como sublinaje de 4.3 - LAM 3 [108]) de perfil de resistencia MDR. En cuanto a sus variantes de resistencia, todas comparten las mismas variantes para INH y RIF. Los aislamientos 11401 y 11880, también comparten los mecanismos para STR (*rpsL* Lys88Arg) y variantes en el promotor de *fabG1-inhA* C-15T, causante de la resistencia a INH y ETH. Estos dos aislamientos comparten una variante de resistencia para KAN en el promotor eis G-12A, que difiere de la variante asociada a esta droga en la mayoría de los aislamientos estudiados (*rrs* A1401G). Cabe destacar que si bien no se encontraron variantes de resistencia a FLQ, se sabe que son fenotípicamente resistentes. Por último es interesante observar que si bien 11880 fue clasificada experimentalmente como LAM5, su espoligotipo *in silico* y contexto filogenético sugieren que es LAM3.

Finalmente, 8 aislamientos fueron asociados al grupo T1. Todas fueron caracterizadas con distintos valores de espoligotipo de T, tanto *in silico* como experimental, sin embargo el aislamiento representativo del linaje 4.8 es la más cercana y coincide con su linaje *in silico*. Las variantes de resistencia son bastante heterogéneas en este grupo para todos los antibióticos. La única compartida por todos los aislamientos es *rpoB* Ser450Trp, asociada a resistencia a RIF. Este grupo puede ser subdividido en dos subgrupos. El primero, T1, agrupa a los aislamientos, 22468 y 10010. Estos aislamientos portan la mutación *katG* Ser315Thr como mecanismo de resistencia a ISO y presentan variantes genotípicas relacionadas la resistencia a EMB. Por otro lado, el

subgrupo T1, conformado por los aislamientos 17817, 17591, 23100 y 20483, presentan resistencia a INH mediada por *fabG1* C-15T, no poseen mutaciones de resistencia a EMB y tienen la variante más representativa de resistencia a aminoglucósidos. También se puede observar que en 10010 no se encontró la mutación que le confiere resistencia a KAN (nuevamente sabemos que es resistente por su perfil experimental) y no pudimos establecer su mecanismo. Su vecino 22468 porta la variante *rrs* G1484T, pero es el único aislamiento que lo posee, por lo que esta subrama parece haber tomado un camino distinto a *rrs* A1401G

En las ramas vecinas a este grupo T1, encontramos varios aislamientos. 16561 está clasificado como 4.7 in silico, agrupada con un representante de 4.7 y clasificado experimentalmente como T5 (que coincide con su linaje). Es una de las 2 muestras de las cuales no se encontró su mutación de resistencia a INH y la variante *rpoB* Gln432Pro. El resto de las muestras no son tan eficientemente clasificadas por los métodos tradicionales, probablemente debido a que sean autóctonas:

- 20811: El linaje *in silico* es 4.8 o 4.3.3, se clasificó como LAM5 experimentalmente pero el espoligotipo *in silico* es T1. Por el otro lado, no se agrupó cercanamente con otros representantes de linajes de TGS. La más cercana es la 4.8 dentro del grupo T1. Es una de los dos aislamientos que tiene 2 mutaciones distintas por droga para 4 de ellas (INH, RIF, PZA y FQL). Se corroboró que la misma no sea una mezcla de 2 aislamientos producto de contaminación), verificando las proporciones de alelos en las variantes. Como se mencionó antes, ocurre frecuentemente que una variante tenga 2 alelos posibles y que esto no sea debido a un error. Se contó la cantidad de variantes con más de un alelo de todas las muestras y esta no tuvo significativamente más variantes respecto al resto (el mismo criterio se aplicó para detectar potenciales coinfecciones en SarS-Cov2 en un trabajo en prensa Goya-Sosa et. al. 2022, aunque con un N mucho más grande ).
- 12821: Al igual que la anterior, presenta 2 mutaciones asociadas a diferentes drogas (INH, RIF y ETH): Se realizó la misma validación respecto al caso anterior y tampoco se encontraron evidencias de contaminación. Tampoco pudo obtenerse el espoligotipo experimental y su clasificación a nivel linaje in silico es ambigua, aunque incluye a 4.8, en su vecindad.
- 14698: si bien coincide su patrón de espoligotipo in silico, experimental y linaje, se encuentra en una rama donde todos menos el 16561 presentan un linaje 4.8. Por

otro lado, es uno de los 2 aislamientos a los cuales no se le pudo detectar la mutación que explica su resistencia a INH.

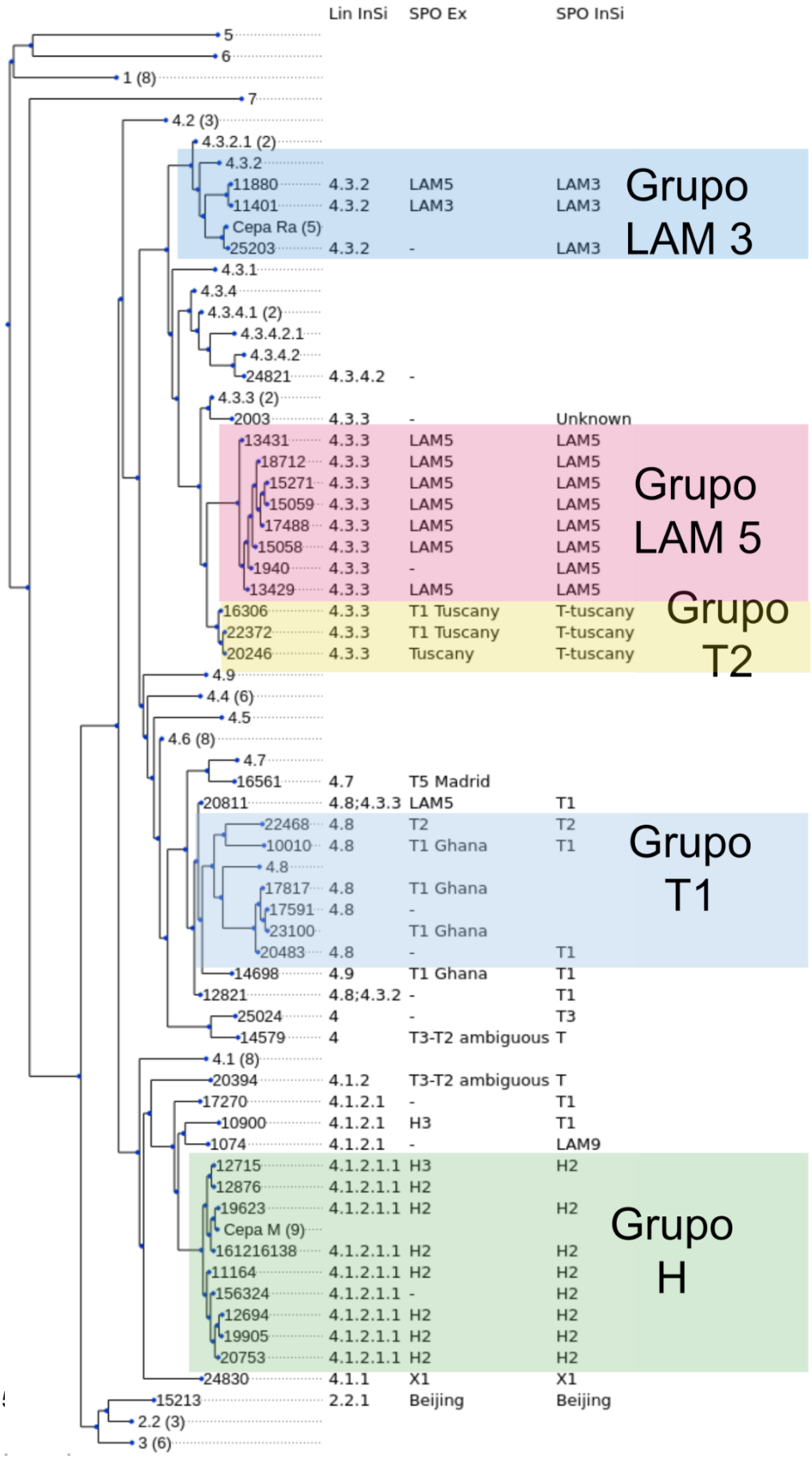
- Las muestras 14579 y 25024 no parecen agruparse claramente con ningún representante de un linaje y de hecho, la primera tiene un patrón de espoligotipo experimental ambiguo.

Teniendo en cuenta la filogenia obtenida con nuestro conjunto de datos, podemos realizar distintas observaciones, contrastando el espoligotipo y los perfiles de resistencia, siempre teniendo en cuenta el tamaño reducido de muestras XDR con las que contamos. Primero, vemos que el patrón de espoligotipo experimental y el linaje, no siempre se agrupan bien. Sin embargo, esto ya se había observado en trabajos previos, por ejemplo la tabla 3 muestra que un mismo patrón de espoligotipo puede corresponder a distintos linajes. Por el otro, vemos una gran prevalencia de las mutaciones en *rpoB* 450 y 432 para resistencia a RIF y de las mutaciones en (*katG* 315, promotor *fabG1*) para INH, y como se marcó, con estudios que respaldan sus elevados valores de CIM, reportado tanto a nivel local como global. Si bien no todos los aislamientos poseen una de estas alternativas, podemos establecer que son comunes en los linajes de Argentina por los altos niveles de resistencia que confieren y por eso probablemente son más frecuentemente observados que otros. Consideramos un caso similar a la mutación *rrs* 1401, la cual aparece independientemente del grupo al que pertenece, pero en menor medida que el resto. Si bien la resistencia a EMB es variable en todos los aislamientos analizados la mutación. La variante *embB* Gly406Asp está presente en todos los aislamientos del grupo H, mientras que todos los aislamientos del grupo LAM5 portan la variante Met306Ile. En estos casos podemos inferir que el mecanismo de resistencia tendría un origen común en estos grupos. En referencia a PZA, como se mencionó anteriormente, el gen *pncA* es uno de los más variables a nivel mundial. Sin embargo se ven fijadas mutaciones en el grupo H (Gln10Arg) y el grupo LAM5 (Arg154Gly), las cuales no se ven en otras ramas. Esto se debe a que varias mutaciones en *pncA* están relacionadas a resistencia, sin embargo la presencia de una de estas mutaciones no garantiza un fenotipo resistente. De manera similar a lo que ocurre con las variantes que confieren resistencia a etambutol, podemos inferir que las mutaciones que confieren resistencia a PZA en los grupos H y LAM 5 provienen de una mutación en un ancestro común a los aislamientos que forman parte de cada uno de los grupos mencionados. En el resto de los aislamientos observamos mutaciones aisladas ( se observan solo en aislamiento) que probablemente se trate de eventos de adquisición de resistencia independientes. En el caso de las FQL, los mecanismos observados parecen ser más recientes, ya que no están relacionados con la filogenia. En particular el caso de *gyrA* 94Val o *gyrA* 94Ala,

parecen ser homoplasias, ya que aparecen en varias ramas, los estudios antes citados hablan de una prevalencia a nivel local y valores elevados de CIM. Finalmente para ETH, también parece estar emergiendo la variante *fabG1* C-15T, pero no parece estar asociado de manera exclusiva con ninguna rama particular de la filogenia.

Está fuera del alcance del trabajo hacer una reconstrucción ancestral (que requeriría mayor diversidad de aislamientos, perfiles de resistencia y tiempos), en la cual pueda observarse de manera más precisa, cuando la variante de resistencia surge como una homoplasia / convergencia o cuando la misma estaba en un ancestro común y se perdió en una de los aislamientos observados (afirmación también admisible, particularmente en este caso, ya que se sabe que las mutaciones de resistencia afectan el fitness de *Mtb* [109,110] en un ambiente sin presión de selección).

Figura 17: Árbol global. En la misma figuran: los aislamientos del proyecto, un subconjunto de muestras de la cepa M, otro de la cepa Ra y las del proyecto TGS. Todas los aislamientos externos fueron descargados de NCBI. Las del proyecto TGS poseen como nombre el linaje que representan y entre paréntesis, si hay más de un representante, la cantidad, lo mismo para los aislamientos M y Ra. Los aislamientos del proyecto tienen 3 columnas Lin InSi, SPO ex y SPO InSi que se corresponden al Linaje obtenido *in silico*, el espoligotipo experimental e *in silico* respectivamente. Cuando el espoligotipo experimental está en blanco, quiere decir que no se realizó. Cuando el cálculo *in silico* está en blanco, quiere decir que no fue concluyente.


Lin InSi
SPO Ex
SPO InSi
5
6
1 (8)
7
4.2 (3)
4.3.2.1 (2)
4.3.2
11880 4.3.2 LAM5 LAM3
11401 4.3.2 LAM3 LAM3
Cepa Ra (5)
25203 4.3.2 - LAM3
Grupo LAM 3
4.3.1
4.3.4
4.3.4.1 (2)
4.3.4.2.1
4.3.4.2
24821 4.3.4.2 -
4.3.3 (2)
2003 4.3.3 - Unknown
13431 4.3.3 LAM5 LAM5
18712 4.3.3 LAM5 LAM5
15271 4.3.3 LAM5 LAM5
15059 4.3.3 LAM5 LAM5
17488 4.3.3 LAM5 LAM5
15058 4.3.3 LAM5 LAM5
1940 4.3.3 - LAM5
13429 4.3.3 LAM5 LAM5
Grupo LAM 5
16306 4.3.3 T1 Tuscany T-tuscany
22372 4.3.3 T1 Tuscany T-tuscany
20246 4.3.3 Tuscany T-tuscany
Grupo T2
4.9
4.4 (6)
4.5
4.6 (8)
4.7
16561 4.7 T5 Madrid
20811 4.8;4.3.3 LAM5 T1
22468 4.8 T2 T2
10010 4.8 T1 Ghana T1
4.8
17817 4.8 T1 Ghana
17591 4.8 -
23100 T1 Ghana
20483 4.8 - T1
14698 4.9 T1 Ghana T1
12821 4.8;4.3.2 - T1
25024 4 - T3
14579 4 T3-T2 ambiguous T
Grupo T1
4.1 (8)
20394 4.1.2 T3-T2 ambiguous T
17270 4.1.2.1 - T1
10900 4.1.2.1 H3 T1
1074 4.1.2.1 - LAM9
12715 4.1.2.1.1 H3 H2
12876 4.1.2.1.1 H2
19623 4.1.2.1.1 H2 H2
Cepa M (9)
161216138 4.1.2.1.1 H2 H2
11164 4.1.2.1.1 H2 H2
156324 4.1.2.1.1 - H2
12694 4.1.2.1.1 H2 H2
19905 4.1.2.1.1 H2 H2
20753 4.1.2.1.1 H2 H2
Grupo H
24830 4.1.1 X1 X1
15213 2.2.1 Beijing Beijing
2.2 (3)
3 (6)

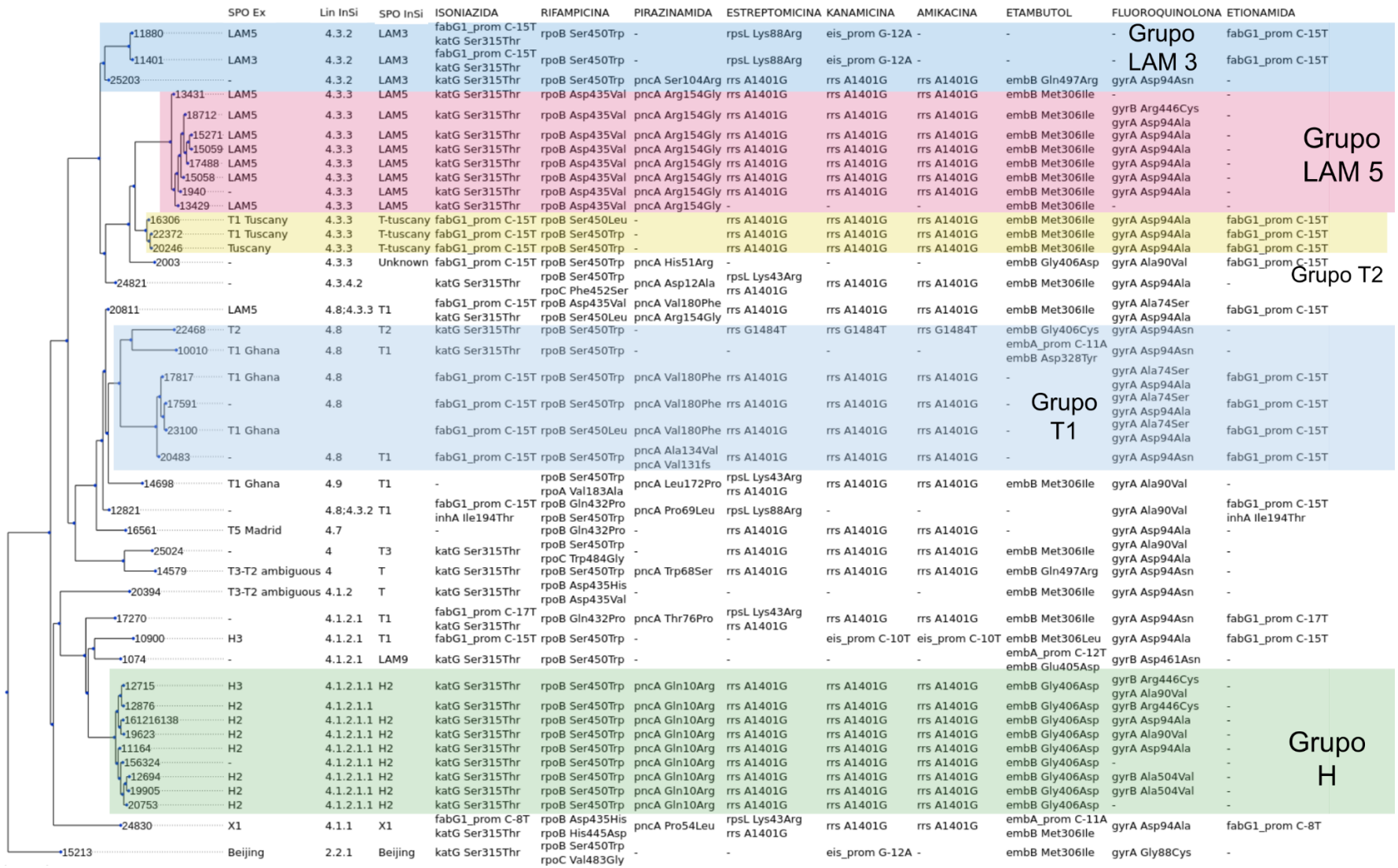



Figura 18: Árbol de máxima verosimilitud obtenido con RaXML de los 43 aislamientos XDR. En las columnas se marca el nombre de cada droga y en cada celda se marcan las variantes de resistencia. Se detectaron entre 1 y 2 por muestra/droga. En colores se marcan los mismos grupos que en el árbol global. Los aislamientos del proyecto tienen 3 columnas Lin InSi, SPO ex y SPO InSi que se corresponden al Linaje obtenido in silico, el espoligotipo experimental e in silico respectivamente. Cuando el espoligotipo experimental está en blanco, quiere decir que no se realizó. Cuando el cálculo in silico está en blanco, quiere decir que no fue concluyente.

# Conclusiones

1) Se caracterizaron las bases genéticas de los mecanismos de resistencia a antibióticos (Estreptomicina, Isoniazida, Rifampicina, Etambutol, Kanamicina, Amikacina, Capreomicina, Pirazinamida, Etionamida y Fluoroquinolona en general) de los aislamientos clínicos de tuberculosis XDR obtenidas en Argentina entre 2003 y 2015.
2) Las variantes asociadas a resistencia obtenidas no cambian respecto a las conocidas en reportes previos para aislamientos MDR o con algún tipo de resistencia en nuestro país.
   a) Isoniacida: las mutaciones *katG* Ser315Thr y *fabG1* C-15T, fueron las más frecuentes, como ya fue reportado en varios trabajos
   b) Rifampicina: las 2 más frecuentes fueron *rpoB* Ser450Trp, como ya fue reportado en otros trabajos a nivel local, sin embargo aparece frecuentemente Asp435Val, que otros estudios identificaron dicha posición como la 3ra más frecuente.
   c) Pirazinamida: las mutaciones de *pncA* son variadas como es de esperarse, pero parece mantenerse dentro de una rama monofilética establecida, por ejemplo los grupos H y LAM5.
   d) Aminoglucósidos de segunda línea: Se mantiene mayoritariamente la mutación *rrs* A1401G, aunque no se conserva a lo largo de toda la filogenia.
   e) Etambutol: las variantes *embB* Gly406Asp y Met306Ile, ya reportadas a nivel local, siguen siendo las predominantes en este conjunto de XDR, donde la primera es exclusiva del grupo H
   f) Fluoroquinolonas: las mutaciones de resistencia de *gyrA* se distribuyeron de manera más variada en el árbol que el resto de los antibióticos, sin embargo, en la lista de las más frecuentes se encontraron las variantes Ala90 y Asp94, ya reportadas como causantes de una alta CIM y encontradas a nivel local en aislamientos resistentes a las FQL
   g) Etionamida: las variantes encontradas fueron sobre el promotor de *fabG1* ( también asociado a la resistencia a INH) y solo una cepa con su blanco, *inhA*. La mutación en *fabG1* C-15T parece quedar fija en varios grupos, pero sin un origen común claro.
   h) La aparición de más de una mutación de resistencia para una dada droga no fue tan poco común como lo esperado, ya que aproximadamente un 50% tenían al menos una. Sin embargo, fueron muy comunes las mutaciones que aparecen en una sola cepa.
3) Se confirmó que los aislamientos XDR no forman un único grupo monofilético en un árbol global pero sí se observaron grupos de aislamientos XDR, que parecen ser del mismo linaje que las cepas M y Ra, los 2 brotes más importantes de Argentina.

# Capítulo 2: Selección y priorización de blancos moleculares proteicos para el desarrollo de drogas capaces de combatir *M. tuberculosis* latente

## Introducción

Si bien los antibióticos se encuentran entre las intervenciones humanas que más han salvado vidas desde el siglo XX, los proyectos de desarrollo de nuevos fármacos antiinfecciosos han presentado dificultades en las últimas décadas por razones que van desde la mala selección de los blancos terapéuticos hasta la reducción de los esfuerzos de descubrimiento de antimicrobianos por parte de las compañías farmacéuticas [111,112].

Actualmente, se acepta que la identificación y validación de blancos apropiados son pasos críticos para diseñar nuevos medicamentos. En este sentido, los métodos de secuenciación masiva han permitido tener una gran colección de genomas de microorganismos disponibles creando nuevas oportunidades para el diseño y desarrollo de drogas para combatir a los agentes infecciosos. La integración de datos provenientes de distintas fuentes ómicas resulta clave para la evaluación de la función, esencialidad, contexto metabólico, drogabilidad y otras características de interés para analizar el potencial de las proteínas para ser utilizadas como blanco en el desarrollo de fármacos. Sin embargo, en la actualidad hay escasez de recursos en línea (WEB) que permitan la integración y consolidación de estos datos para definir una lista de potenciales blancos. Aquí presentamos la base de datos Target-Pathogen (target.sbg.qb.fcen.uba.ar/patho), una aplicación en línea que permite clasificar e identificar blancos moleculares a partir de la integración y ponderación los datos, centrándose en la esencialidad, el contexto metabólico y la predicción estructural de las proteínas, facilitando la identificación y priorización de blancos candidatos adecuados para nuevos proyectos de desarrollo de fármacos.

### Anotación

El primer paso para obtener información de las proteínas que componen el genoma de un patógeno es la anotación. La anotación de un genoma hace referencia al proceso de identificación de las posiciones, funciones e información relacionada de todos los genes y regiones de interés

del genoma, por ejemplo las codificantes, promotores, sitios de unión, etc. Este proceso puede dividirse fundamentalmente en 4 etapas:

- Anotación estructural: Intenta localizar la posición física de los genes y otros sitios biológicamente relevantes (promotor, sitios de fosforilación, señales de localización, etc…).
- Anotación funcional: Se ocupa de dilucidar las funciones de los genes y otros sitios predichos en la etapa anterior.
- Control de calidad de la anotación: Verifica que se hayan encontrado un conjunto mínimo de genes. Dicho conjunto está dado por el conjunto de genes bien conocidos encontrados a un determinado nivel taxonómico. Por ejemplo, se sabe que todas las bacterias tienen una RNA polimerasa, o que todas las micobacterias tienen *katG*. COG/KOG [113] o BUSCO[114] son bases de datos que compilan esa información y tienen conjuntos de genes para diversos niveles taxonómicos (bacteria, eucariota, arqueas).
- Refinación y curación manual: una vez que se tiene una anotación inicial de un genoma, esta puede ser mejorada, ya sea por nueva información existente (RNA-Seq, datos de proteómica, resultados de PCR, Sanger, etc) o porque un curador se enfoca en curar una familia de genes, proceso biológico o vía metabólica.

En la práctica, las 2 primeras etapas suelen encontrarse superpuestas parcialmente. Y el proceso no siempre es secuencial, ya que se pueden obtener varias anotaciones, combinarse, o mejorarse con otras realizadas con bases de datos actualizadas.

La predicción de genes ha sido estudiada desde diversas perspectivas y se han desarrollado múltiples herramientas bioinformáticas tanto para la predicción de regiones codificantes como ARN no codificantes. También es importante remarcar que cada región de interés tiene su complejidad, por lo que hay herramientas específicas que atacan cada problema: dominios transmembrana [115] , selenoproteínas [116], miRNAS (Rfam [117] + Infernal [118] )

En el caso de genomas procariotas, la predicción de genes codificantes de proteínas resulta más sencilla pues los genomas procariotas cuentan con gran densidad génica y ausencia de ciertos elementos como los intrones. Sin embargo, la predicción no resulta ser 100% eficaz debido a la presencia de numerosas regiones solapantes y en general se genera un gran número de falsos positivos.

Los métodos usados para identificar genes pueden dividirse en dos categorías: métodos *ab initio* y métodos basados en homología. En general los anotadores combinan ambas estrategias y ponderan la evidencia obtenida por cada uno.

Los métodos *ab initio* buscan predecir las estructuras genéticas a partir de características del propio genoma, como ser señales conservadas (codones de inicio y de stop, sitios de splicing, sitios de unión a ribosomas, etc.) o a partir de las características de las secuencias codificantes de cada microorganismo (contenido de G+C, el uso diferencial de codones, proporción de aminoácidos, etc.). Estos métodos no utilizan ningún tipo de información de secuencias externas conocidas. La mayor parte de los falsos positivos se debe a este procedimiento, sin embargo es una herramienta valiosa para captar potencialmente nuevos genes en especies / linajes no tan estudiados.

Los métodos basados en homología utilizan algún método de comparación de secuencias (alineamiento directo o perfiles de secuencias conocidas), pero sea cual sea el método empleado tienen en cuenta la estructura básica del gen para delimitar el alineamiento, es decir un marco de lectura suficientemente largo entre un codón de inicio y uno de terminación. Los análisis de homología pueden llevarse a cabo con genes o proteínas de otras especies o alternativamente con evidencia de transcripción del propio organismo.

En el caso de predicción de genes que codifican para ARNr y ARNt, los resultados alcanzados son mejores ya que en la actualidad existe una amplia gama de programas que funcionan de manera casi perfecta con niveles de exactitud cercanos al 100%. Para los ARNt el programa más utilizado es el tRNAscan-SE [119], este programa consiste en la detección de genes ARN de transferencia presentes en la secuencia génica a partir de modelos estadísticos de secuencias de ARNs de transferencia conocidos. Para ARNr, en cambio, el más utilizado es RNAmmer [120], basado en modelos ocultos de Markov. Estos programas prácticamente no han recibido modificaciones y sus predicciones son de altísima confianza. En los últimos años los modelos utilizados para su funcionamiento fueron incorporados a otras bases de datos, como ser Rfam [117], que es específica de RNAs, pero general en el sentido que predice RNAs de distintos tipos.

Para el caso de secuencias que codifican proteínas, existen diversos programas para localizar genes codificantes dentro del genoma bacteriano, entre ellos, Genemark [121,122] - con él y sus variantes están predichas las proteínas del 99% de los genomas anotados en NCBI - GLIMMER 3 [123] y Prodigal [124]. El proceso de predicción consiste en detectar los marcos abiertos de lectura (ORFs, del inglés *open reading frames*) mediante el reconocimiento de patrones

estadísticamente relevantes producto de entrenar modelos predictivos usando regiones codificantes conocidas. Cuando esto no es posible, puede intentarse con genes de organismos relacionados anotados previamente. Sin embargo, en los casos en que los organismos anotados no son lo suficientemente cercanos, esta última opción no es posible. Como la frecuencia de codones de terminación al azar esperada es de 1 cada 21 codones (3 de 64 posibles), una secuencia de gran extensión (entre un codón de inicio y uno de terminación), con ausencia de codones de terminación internos, es una excelente candidata para contener regiones codificantes de proteínas.

Seguidamente a las regiones del genoma predichas como genes (anotación estructural) se las anota funcionalmente. Para tal fin, dichas regiones son traducidas al lenguaje de proteínas y mediante la utilización de distintas bases de datos se buscan similitudes de secuencias con proteínas anotadas previamente, es decir, con función conocida. Hay muchas bases de datos y modelos para la anotación de proteínas, con lo cual los anotadores actuales ensamblan el resultado de uno o más programas y bases de datos. Los 2 más utilizados son InterproScan [51] y Eggnog [113].

Es importante destacar que aunque un genoma puede estar anotado y disponible en una base de datos, no resulta extraño que cada proyecto particular tenga interés en analizar distintos factores para los cuales la anotación previa no sea suficiente o no resulte útil. Es por esto que cada proyecto genómico suele seleccionar distintas estrategias de búsqueda así como también distintas bases de datos, dependiendo de las incertidumbres que posean acerca del genoma secuenciado, para de esta manera, responder sus preguntas biológicas.

## Vías Metabolicas

El metabolismo es el conjunto de reacciones bioquímicas y procesos fisicoquímicos que ocurren en una célula. Estos procesos complejos interrelacionados son la base de la vida, a escala molecular, y permiten las diversas actividades de las células, entre ellas, crecer, reproducirse, mantener sus estructuras y responder a diversos estímulos. Entre las principales moléculas del metabolismo se encuentran los aminoácidos, proteínas, lípidos, carbohidratos y nucleótidos. Dentro del metabolismo tienen particular importancia las enzimas, moléculas de naturaleza proteica (en general), catalizadoras de reacciones químicas y las proteínas que presentan funciones reguladoras del metabolismo, es decir, aquellas proteínas con capacidad de activación y/o inhibición de diferentes componentes metabólicos. Es por eso que el genoma cumple un rol

preponderante sobre el metabolismo, codificando la información necesaria para la síntesis de las enzimas y otros elementos regulatorios.

Toda base de datos de metabolismo, tiene al menos los siguientes elementos: metabolito o compuesto (elemento químico, sustrato o producto de una reacción de una reacción), la reacción en sí, catalizada por una enzima o complejo enzimático, que acelera la transformación de un sustrato en el producto correspondiente. Finalmente está el concepto de vía metabólica, que nos sirve para agrupar conjuntos de reacciones, facilitando su estudio, comprensión y sobre todo, dando un objetivo o función biológica específico.

Dado que las enzimas están codificadas en el genoma, es posible inferir automáticamente una parte del potencial metabolismo de un organismo patogénico a partir de su secuencia. Esto es de particular importancia cuando se trata de cepas biotecnológicas o clínicamente relevantes, ya que el metabolismo representa un factor clave para comprender su fisiología. En las últimas dos décadas han sido desarrolladas diferentes bases de datos que brindan la posibilidad de llevar a cabo análisis metabólicos y alimentan los algoritmos desarrollados para la reconstrucción automática del complemento metabólico de los organismos. KEGG [125], MetaCyc/BioCyc [126,127] y Reactome [128] representan algunas de las bases de datos con curación manual más utilizadas en la actualidad.

Tanto KEGG como Pathway Tools [129] representan los principales métodos automatizados de reconstrucción metabólica. A lo largo de esta tesis, nos concentramos en el uso de Pathway tools. El algoritmo Pathologic [130], implementado como subherramienta de Pathway Tools se utiliza para crear una base de datos (PGDB, Pathway/Genome Database) que contiene las vías metabólicas, y las reacciones, genes y proteínas asociadas, predichas para un organismo dado a partir de su genoma anotado en formato Genbank. Pathologic, en una primera etapa, busca dilucidar las reacciones enzimáticas presentes en la red a través de la comparación entre la anotación del genoma en estudio con las reacciones y vías metabólicas presentes en la base de datos Metacyc. La misma, es una base de datos altamente curada que contiene información de vías metabólicas, reacciones, enzimas y metabolitos de todos los dominios de la vida.

La etapa automática de reconstrucción metabólica incluye la determinación de asociaciones entre genes (proteínas) y reacciones, basadas principalmente en el número EC (*Enzyme Commission*) [49] - que suele formar parte de las anotaciones de un gen dentro de un genoma anotado - o de manera alternativa, en otro tipo de anotaciones como el nombre proteico o los términos GO (vale aclarar que los genomas a procesar deben estar anotados con dichas ontologías). La numeración EC es un esquema de clasificación numérica para las enzimas,

basado en las reacciones químicas que catalizan. Como sistema de nomenclatura de enzimas, cada número EC está asociado a las reacciones específicas catalizadas por cada proteína. Enzimas diferentes (por ejemplo que procedan de organismos diferentes) que catalizan la misma reacción recibirán el mismo número EC. Cada código de enzimas consiste en las dos letras EC, seguidas por 4 números separados por puntos. Estos números representan una clasificación progresivamente más específica, como puede verse en la Figura 19

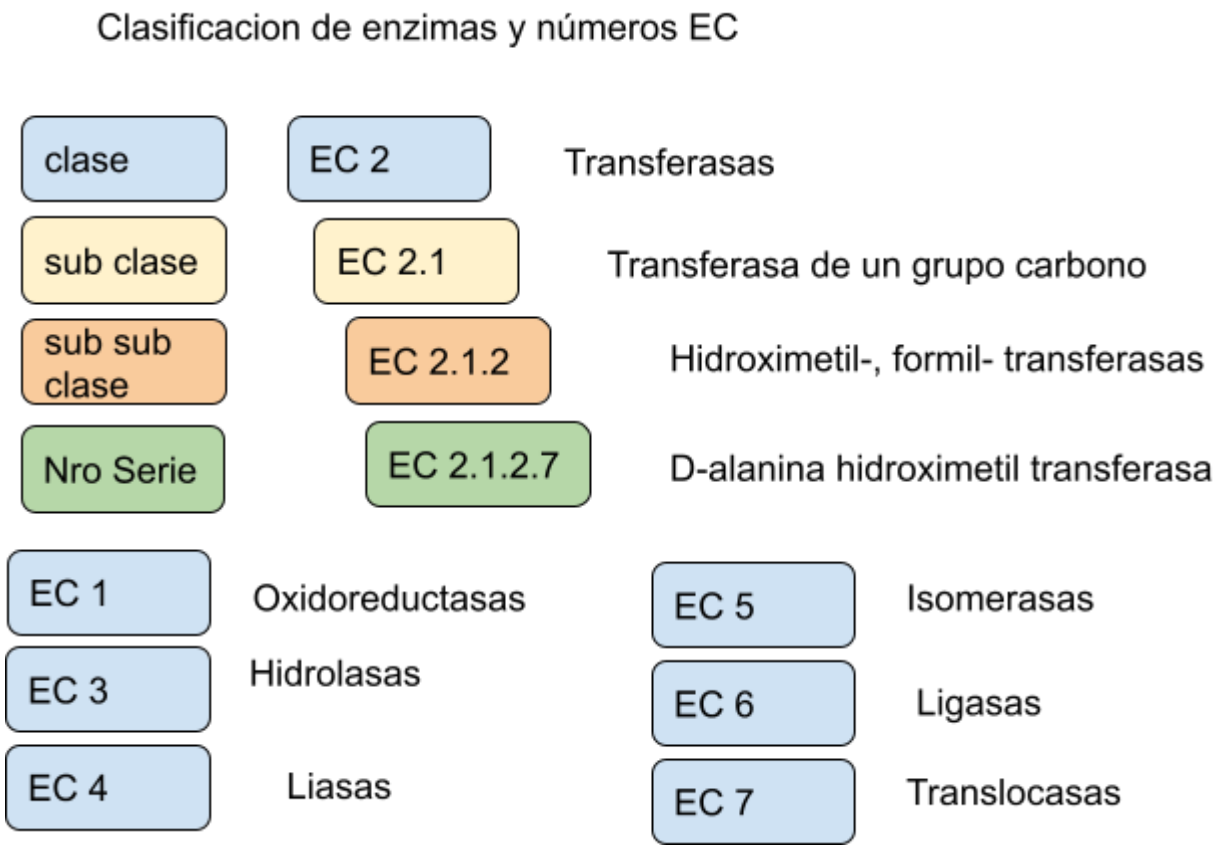



Figura 19: Ejemplo del código numérico de clasificación de enzimas (EC).

Una vez establecidas las asociaciones entre los genes y sus correspondientes reacciones, Pathway Tools mapea las reacciones con sus correspondientes vías metabólicas.

El resultado de un metabolismo, puede representarse de varias maneras [131], en este trabajo escogeremos hacerlo con un grafo dirigido, donde los nodos son las reacciones, y las aristas interconectan 2 nodos cuando el producto de una reacción es utilizado como sustrato de la reacción subsiguiente.

La representación en formato de grafos permite evaluar ciertas medidas topológicas de relevancia para analizar el contexto metabólico de las distintas enzimas. La centralidad es una

medida de la contribución de cada nodo según su ubicación en la red. Existen diversos tipos de medidas de centralidad entre ellas, la centralidad de intermediación (*betweenness centrality*) [132] y la centralidad de proximidad (*closeness centrality*). El análisis de estas medidas topológicas permiten detectar nodos que resultan biológicamente relevantes en el contexto de la red metabólica. De particular importancia para esta tesis es la relación centralidad-esencialidad [133], y si bien hay otras aproximaciones al problema, como la de interacciones [134] o complejos [135] esenciales, la elegida por nosotros nos permite identificar reacciones potencialmente esenciales con la información disponible, específica del genoma que se esté estudiando.

Por esto decimos que nodos (reacciones) con alta centralidad de intermediación son relevantes (en particular usaremos *betweenness*) y se consideran blancos atractivos, debido a que la inhibición de una enzima que participa en una reacción de gran centralidad podría afectar, de manera simultánea, diferentes vías metabólicas del patógeno ejerciendo una acción de amplio espectro en el contexto metabólico **(**Figura 20**)**.

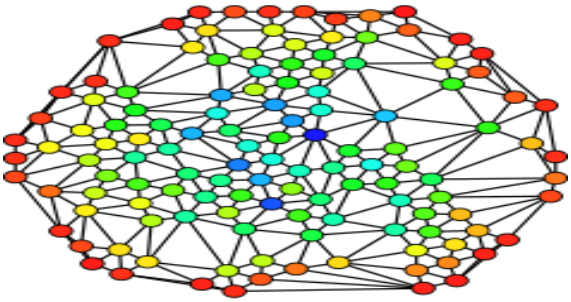


Figura 20: Red en forma de grafo donde se representa la de centralidad de intermediación. Nodos azules presentan centralidad máxima, mientras que en los rojos es mínima.

El otro criterio que también es tenido en cuenta a la hora de identificar enzimas o reacciones cruciales para la supervivencia de los organismos, es el de *chokepoint* (reacciones que consumen o producen de forma única un sustrato o producto dado, respectivamente). Dicho análisis brinda la posibilidad de identificar posibles proteínas desde la perspectiva metabólica, pues un bloqueo de este tipo de enzimas podría conducir a la acumulación de metabolitos potencialmente dañinos o a la incapacidad de producir por parte del organismo compuestos esenciales para la vida celular **(**Figura 21**)**.


Figura 21: Representación de tipos de choke-points. a) Producción: Reacciones tales que sólo una reacción R1 produce el metabolito M, y al menos una reacción consume M. (b) Consumo: Reacciones tales que sólo una reacción R2 consume el metabolito M, al menos una reacción produce M. Tomado de https://bioinformatics.ai.sri.com/.

## Estructuras proteicas

### La relación estructura-función en las proteínas

La mayor parte de los roles estructurales y funcionales esenciales para la célula son mediados por proteínas. En esta tesis tomaremos el principio fundamental en la ciencia de las proteínas que establece que “la estructura proteica determina la función”, sin embargo para los tiempos que corren se sabe que esto es una simplificación de la realidad [136], pero dentro de nuestro marco de trabajo, es un modelo adecuado para tomar decisiones.

Las diferentes estructuras permiten a las proteínas actuar como catalizadores (en el caso de las enzimas) en una variedad sorprendente de reacciones químicas, desempeñar importantes funciones estructurales, de transporte y de regulación en los organismos. Dado que la estructura proteica conduce a la función, y las funciones proteicas son diversas, no es sorprendente que las estructuras proteicas sean semejantemente diversas. Además, la diversidad funcional de las proteínas es expandida a través de su interacción con moléculas pequeñas, denominadas ligandos, donde en nuestro caso nos enfocaremos en los compuestos tipo droga.

### Cómo obtener estructura de una proteína

La estructura de una proteína, se puede determinar con distintas técnicas experimentales o inferir con metodologías in silico. Experimentalmente se puede conocer la estructura a través de diversas técnicas, como cristalografía de rayos X, resonancia magnética (RMN) o Microscopia

electrónica (y más recientemente *CryoEM*). Cada una de ellas presenta distintas ventajas y desventajas a la hora de resolver un modelo 3D. A través de métodos computacionales la inferencia se realiza a través de la aplicación de aprendizaje automático (en el caso de AlphaFold) o modelado por homología.

Vale la pena hacer dos aclaraciones importantes, primero que en general, los métodos computacionales se basan en las estructuras obtenidas experimentalmente, es decir, estas son el método más confiable y preferido. Sin embargo, son difíciles de lograr a gran escala y la dificultad de obtenerla varía muchísimo de proteína en proteína. Segundo, las proteínas tienen una dinámica y la estructura obtenida es una "foto" o conformación de la proteína (en el caso de RMN varias), que se toma como útil para estudiar y analizar la función de la proteína, como ser interacciones con potenciales ligandos.

### Modelado por homología

Debido a que la cantidad de secuencias disponibles crece a un ritmo mucho mayor que el número de estructuras (esto se puede observar comparando los gráficos de crecimiento de la Figura 9 A y D), se hace necesaria la aplicación de métodos in silico para la obtención de estructuras proteicas para compensar esta diferencia. El modelado por homología es el proceso por el cual una o más proteínas de estructura conocida, cuya secuencia es similar a la proteína de interés que carece de estructura conocida, se utilizan como moldes para modelar la estructura desconocida [137,138]. En la actualidad, dada la aparición de AlphaFold [35], el modelado por homología deja de ser el adecuado cuando no se conoce una secuencia con alta identidad (con estructura conocida) de la proteína que se quiere modelar. Sin embargo se explica a continuación, debido a que fue intensamente utilizado para esta tesis.

El modelado por homología se basa en la observación de que las proteínas que tienen estructura primaria estrechamente relacionada (es decir, que han divergido de una proteína ancestral común durante la evolución) comparten segmentos de conformación similar. Se asume que si las secuencias aminoacídicas están cercanamente relacionadas, luego la estructura 3D de la proteína puede ser predicha partiendo de las estructuras 3D de otras proteínas que pertenezcan a la misma familia. Es importante remarcar, que lo contrario no es cierto, por ejemplo, proteínas/dominios con muy baja identidad de secuencia (<10%), pueden tener una estructura 3D similares [139].

La metodología de modelado por homología puede dividirse en cuatro pasos principales, que se muestran en la Figura 22 [140]. A continuación se presenta una breve descripción de cada uno de ellos.

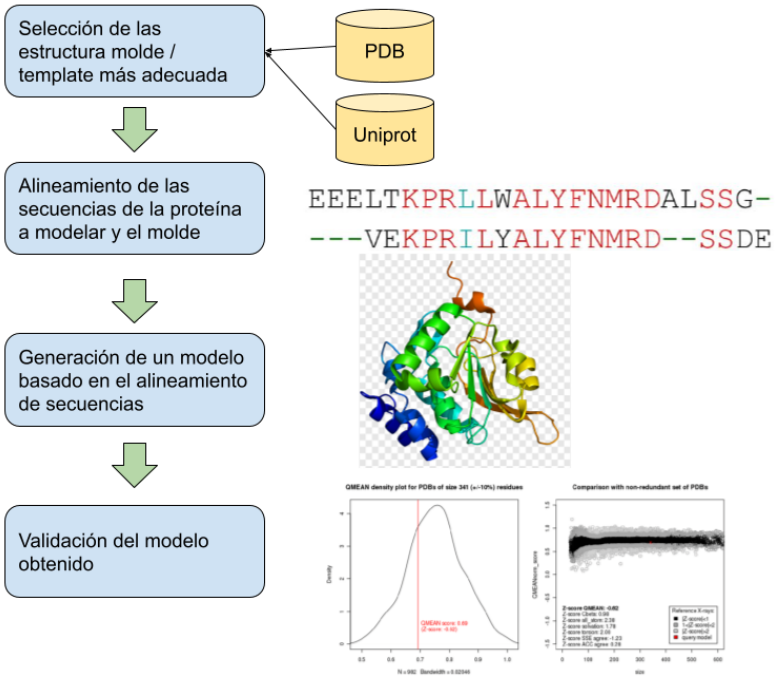



Figura 22: Pipeline de modelado por homología con Modeller

## Selección de la estructura molde más adecuada

La calidad del modelado por homología es altamente dependiente de la elección de la estructura que se utilizará como molde o *template*. Como regla general, una estructura proteica puede proveer un modelo más o menos acertado para otra proteína si entre ellas comparten una identidad de secuencia superior al 50%. Aunque aplicando conocimiento de aminoácidos en posiciones conservadas, como se explicará más adelante, dicha identidad puede ser muy inferior, ya que se hace hincapié en la conservación de dichos aminoácidos.

En el caso de tener más de un molde (que cumpla los criterios de identidad y cobertura), se pueden evaluar otras características del experimento, como ser la estructura cuaternaria, si el mismo se cristalizó con o sin ligando y la resolución del cristal.

### Alineamiento de las secuencias de la proteína a modelar y el molde

Es la etapa más importante y la más delicada, ya que la construcción del modelo se realizará conforme a este alineamiento. Existe una gran variedad de métodos de alineamiento de secuencias proteicas, y los AMS suelen ser más precisos que los de a pares. Por otro lado, cuando se construye un alineamiento para modelado por homología, se tiende a ubicar los "gaps" fuera de las regiones de estructura secundaria de la proteína molde. En general, es necesario mejorar el alineamiento obtenido in-silico por un curador que conozca bien la proteína / familia en cuestión, pero en el caso de esta tesis, al realizar un procesamiento masivo, realizamos todos los análisis sin esas correcciones.

### Generación de un modelo estructural basado en el alineamiento de secuencias

En este trabajo de tesis se utilizó el programa MODELLER [137], que hasta la aparición de AlphaFold, era uno de los programas más ampliamente utilizados, y pertenece al conjunto de métodos que implementa el modelado por homología por satisfacción de restricciones espaciales. Los métodos de esta clase generan un conjunto de restricciones en la estructura de la secuencia a modelar, basándose en su alineamiento contra las estructuras molde. Estas restricciones en general se obtienen asumiendo que las correspondientes distancias y ángulos entre residuos alineados en las estructura modelada y molde son similares. Estas restricciones derivadas de la homología típicamente se complementan con restricciones estereoquímicas sobre las longitudes de los enlaces, ángulos de enlace, ángulos diedros, y contactos de no-unión átomo-átomo obtenidas de un campo de fuerzas de mecánica molecular. El modelo se deriva luego minimizando las violaciones de todas las restricciones.

### Validación del modelo obtenido

Para evaluar los modelos obtenidos por homología, se necesitan métodos que permitan testear la calidad de los mismos. Estos métodos deberían ser capaces de verificar la confiabilidad del modelo, es decir, distinguir un modelo apropiadamente plegado de uno inapropiadamente plegado y evaluar propiedades geométricas y estéricas de los modelos. Existen numerosos programas de testeo de calidad de los modelos con diferentes criterios. La mayoría de los métodos han sido desarrollados utilizando datos empíricos de proteínas globulares de estructura conocida. Las estructuras modeladas en esta tesis fueron testeadas utilizando el método QMEAN [141]. Brevemente, el mismo tiene precalculada una serie de propiedades importantes, sobre PDBs

de buena calidad. Esas propiedades se hacen converger a un score, del cual se tienen varias distribuciones, agrupadas por el tamaño de la proteína / cadena. Lo que hace el programa es volver a calcular esas propiedades y en función de ellas el score, sobre el cual, junto con la longitud de la proteína moldeada, determinan qué tan “lejos” (malo) o “cerca” (bueno) están las propiedades de mi estructura comparadas con las experimentales de buena calidad.

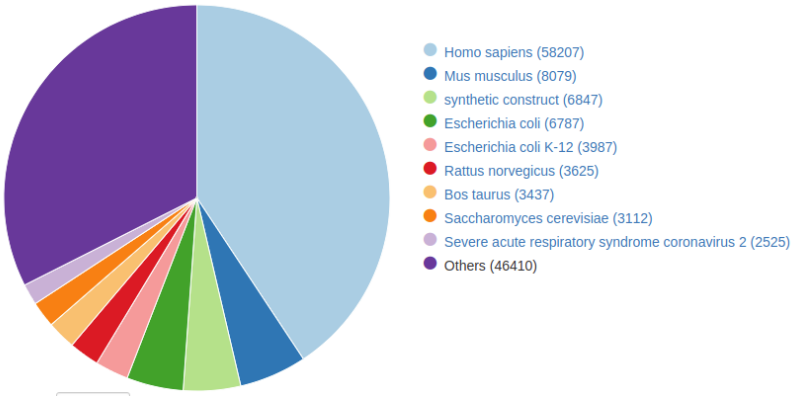
## Bolsillos y Drogabilidad

Para cualquier diseño racional de una droga, es prerrequisito conocer la estructura del blanco y su capacidad para unirse a un compuesto. Al ser un problema complejo, existen distintas aproximaciones para estudiarlo.

Partiendo de la estructura proteica, ya sea experimental o modelada es posible predecir sus cavidades (*pockets*) drogables. Estos *pockets*, son superficies dentro de la estructura proteica con alta probabilidad de unir un compuesto tipo droga. Asociadas a los mismos, pueden calcularse una serie de propiedades, a partir de las características fisicoquímicas del conjunto de los residuos que lo componen. Para nosotros será de particular interés la drogabilidad, que establece que tan posible es que una un compuesto tipo droga (independientemente de las propiedades o estructura del compuesto). Para calcular las cavidades y caracterizarlas, utilizamos el programa Fpocket [142] y su score de drogabilidad propuesto. Brevemente, el mismo se basa en el algoritmo de teselación de Voronoi (criterio geométrico) para identificar cavidad y luego computa descriptores fisicoquímicos (densidad hidrofóbica, superficies polares y apolares, cantidad de carga, tamaño) que son utilizados para generar el score de drogabilidad, basado en el análisis de una distribución para todos los pockets que unen a una droga en el PDB [142,143].

## PDB

Para la obtención y cálculo de estructuras en esta etapa, se utilizó la base de datos PDB [144], la cual provee a los investigadores y organizaciones, la posibilidad de almacenar y recuperar las estructuras obtenidas experimentalmente (coordenadas en formato PDB o CIF), junto con todos los datos asociados (proteína, experimento, resolución, sistema de expresión, etc). La misma, al momento de la escritura de esta tesis contiene 193455 estructuras depositadas y el desglose por organismo puede verse en Figura 23, en particular hay 1481 de Mtb (se remarca que muchas proteínas tienen más de una estructura, por ejemplo una por ligando cristalizado).




Show 10 entries

| Scientific Name of Source Organism | Number of Entries |
|---|---|
| Influenza A virus | 289 |
| Klebsiella pneumoniae | 563 |
| Lama glama | 586 |
| Legionella pneumophila subsp. pneumophila str. Philadelphia 1 | 208 |
| Leishmania major | 216 |
| Methanocaldococcus jannaschii | 291 |
| Methanocaldococcus jannaschii DSM 2661 | 200 |
| Mus musculus | 8079 |
| Mycobacterium tuberculosis | 1357 |

Figura 23: estructuras experimentalmente determinadas en la base de datos PDB agrupada por organismos.

## Selección y priorización de blancos para el desarrollo de drogas.

Tradicionalmente el descubrimiento de moléculas con actividades antimicrobianas se ha basado en estudios de laboratorio experimentales, que involucran diseños complejos y reactivos costosos y equipos, o por pura casualidad, como fue el descubrimiento de la penicilina, la primera molécula con actividad antibiótica, identificada por Sir Alexander Fleming en la década de 1920 [145]. Si bien estos últimos pueden considerarse excepciones, el camino habitual hacia el descubrimiento de fármacos está lleno de advertencias experimentales e implica una extensa prueba y error.

Metodologías experimentales de alto rendimiento, donde bibliotecas químicas del orden de cien mil un millón de moléculas se examinan para identificar posibles moléculas similares a fármacos, han permitido la identificación de "nuevas" moléculas antibióticas candidatas, pero la utilidad de esta “fuerza bruta” se cuestiona actualmente debido a que la tasa de descubrimiento de nuevos compuestos potenciales se desplomó [146]. Las limitaciones en la representación de las bibliotecas químicas en sí mismas dificultan la utilidad del enfoque HTS, especialmente considerando que las estimaciones del número total de moléculas interesantes para el

descubrimiento de fármacos son del orden de $10^{60}$ moléculas [147]. En este contexto, surgen metodologías alternativas, que se basan en el uso de la información disponible en escala genómica gracias a la irrupción de metodologías de secuenciación masiva.

Estas tecnologías ómicas cambiaron el enfoque del descubrimiento de nuevos fármacos hacia un enfoque centrado en el blanco. Esto permitirá tomar decisiones basadas en la enorme cantidad de información biológica disponible para seleccionar blancos con mayor probabilidad de éxito en las subsiguientes etapas de desarrollo experimental, a las que se pueden dirigir más recursos experimentales y económicos. Por ejemplo, empresas biofarmacéuticas llevaron a cabo varias campañas de detección de alto rendimiento, cuyos blancos candidatos que fueron priorizados a través de criterios derivados del genoma; luego se expresaron, purificaron y se buscaron inhibidores mediante la utilización de bibliotecas químicas [148]. Sin embargo, después de extensos esfuerzos de minería del genoma, centrado en el blanco, los enfoques fueron fuertemente criticados debido a su bajo rendimiento en términos de generar nuevos potenciales blancos o inhibidores [149]. Parte de estas críticas se relacionaban con la expectativa de que la 'era' de la genómica abriera las puertas a una gran cantidad de nuevos antimicrobianos, una creencia claramente ingenua. El incremento de las estrictas políticas de las agencias reguladoras de medicamento en los últimos años, la relación riesgo-beneficio desfavorable para productos farmacéuticos y la falta de atractivo general del mercado antibacteriano (en comparación con otros enfermedades) se encontraban entre los factores clave que contribuyeron al bajo rendimiento de la investigación en nuevos antimicrobianos durante este período, particularmente por parte de las compañías farmacéuticas [150]. A pesar de las advertencias mencionadas, existen ejemplos exitosos de este enfoque en la literatura, uno de los primeros es el desarrollo de inhibidores de la péptido deformilasa (basados en ácidos N-alquil urea hidroxámicos) [151]. Otro ejemplos fructíferos incluyen el desarrollo de inhibidores del VIH [152], la identificación de inhibidores dirigidos a proteínas involucradas en la biogénesis del ácido teicoico de la pared en *Staphylococcus aureus*, que potencian fuertemente los antibióticos β-lactámicos contra *S. aureus* resistente a la meticilina [153]; el descubrimiento de inhibidores del regulón MvfR (implicado en la autoinducción) en Pseudomonas aeruginosa [154]; y el desarrollo reciente de mAbs dirigidos al ensamblado del β-barril montaje en *E. coli*, con notable actividad bactericida [155]. El rápido avance de los antibióticos cepas de resistencia y la falta de nuevos compuestos contra patógenos bacterianos ha re-instalado la idea de que los enfoques basados en blancos, particularmente cuando se integran múltiples capas de información (por ejemplo, estructura 3D, esencialidad, localización subcelular, criterios *off-target*, contexto metabólico, conservación de secuencias y propiedades funcionales), todavía ofrecen una vía de

investigación prometedora en hacer frente a las enfermedades infecciosas. Sin embargo, en la actualidad existe una escasez de herramientas bioinformáticas para la integración y consolidación de datos provenientes de las diferentes ómicas que permita la selección de blancos adecuados en el proceso de diseño racional de drogas. En este capítulo presentaremos Target-Patogen (http://target.sbg.qb.fcen.uba.ar/patho), un plataforma web diseñada y desarrollada específicamente como un recurso online que permite la integración y valoración de estas capas de datos para la identificación y ranqueo de proteínas blanco [156].

# Materiales y Métodos

## *Pipeline* de procesamiento

Todos los datos presentes en Target-Pathogen se basan en el cálculo de propiedades *in silico* seleccionadas para cada proteína o en la integración y análisis de datos disponibles públicamente. Todos las proteínas de cada genoma se descargaron de la base de datos UniProt [157]. Los mismos se analizaron con el programa HMMER [158] para asignar las correspondientes familias o dominios PFAM [159]. El esquema general de Target-Pathogen se muestra en la Figura 24.

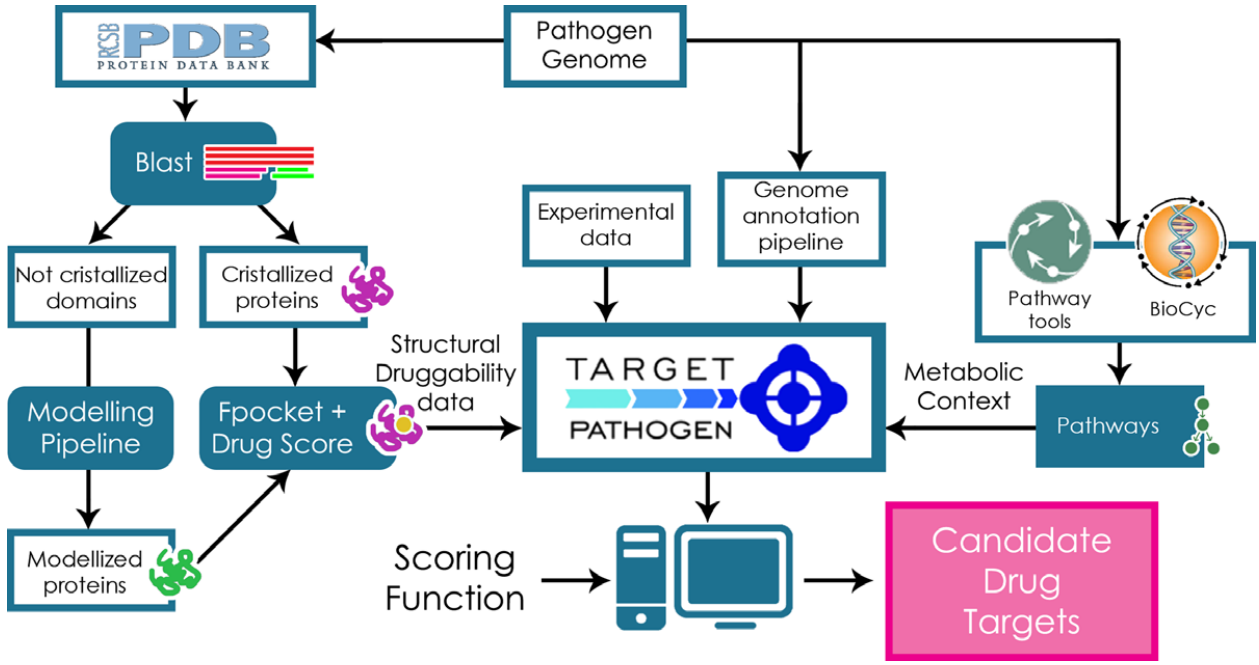



Figura 24. Esquema general de Target-Pathogen. Los análisis metabólicos y de drogabilidad estructural se integran con los datos experimentales disponibles y de análisis *in silico*. Una vez que los mismos se integran en Target-Pathogen, se utiliza una función de puntuación diseñada por el usuario para ponderar diferentes características a fin de obtener una lista clasificada de posibles targets farmacológicos.

## Generación de modelos basados en homología estructural

Se buscaron estructuras experimentales en el PDB. Se asignaron a la consulta estructuras de proteínas completas o dominios PFAM presentes en el PDB con una identidad de secuencia del 95 %. Para todas las proteínas restantes, se intentó construir modelos basados en homología usando MODELLER. Para todas las estructuras (cristales y modelos), luego se calcularon propiedades estructurales como: (i) *Score* de drogabilidad (DS), (ii) Residuos del sitio activo [160] y

(iii) residuos relevantes de la familia PFAM (posiciones con un alto aporte de información, en general conservados en el dominio).

## Evaluación de la drogabilidad estructural

Como se estableció en la introducción, la drogabilidad es un concepto que describe la capacidad de una proteína determinada para unirse a un compuesto, que a su vez modula su función [161,162]. Las proteínas blanco de un fármaco deben tener un *pocket* bien definido con atributos físico químicos adecuados para permitir la predicción de los sitios de unión.[162,163].

Se evaluó la drogabilidad estructural de cada potencial target mediante el uso del programa FPocket para detectar bolsillos y luego se calculó el correspondiente DS. Las proteínas se clasificaron según el score de drogabilidad de su *pocket* más drogable de la siguiente manera: entre 0.2 y 0.5 pobremente drogable, entre 0.5 y 0.7 drogable y mayor a 0.7 como muy drogables. Las que poseían DS menor a 0.2 se descartaron . En la Figura 25 puede verse la distribución de las proteinas de *Mtb* H37Rv según su *pocket* más drogable, donde las zonas en verde, amarillo y rojo muestran la clasificación mencionada.

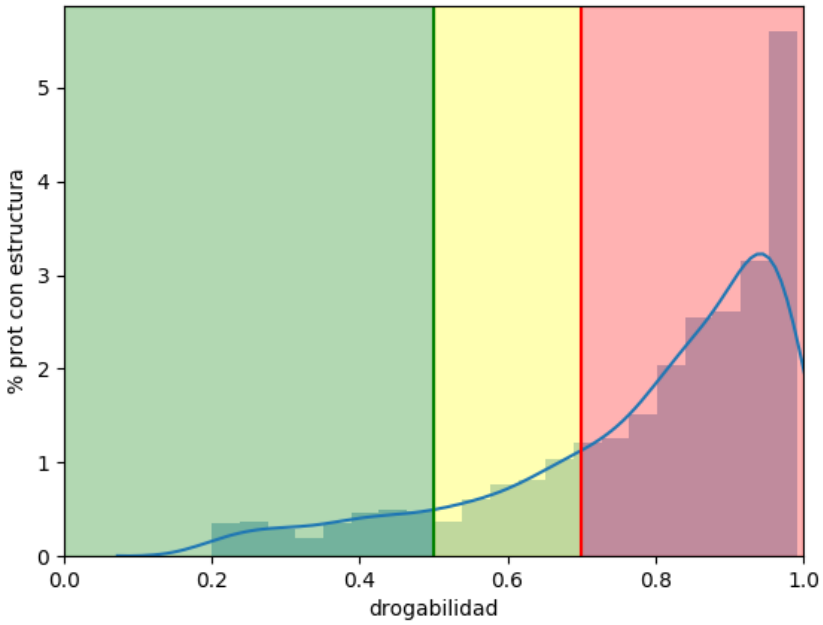



Figura 25: Distribución de la drogabilidad para las estructuras de H37Rv, tomando el *pocket* más drogable de cada una. Se observan 295 proteínas pobremente drogables, 289 drogables, 1462 altamente drogables. En 335 proteínas no se detectó ningún pocket relevante.

## Criterios *off-target*

Gran parte del diseño racional de una droga apunta a mejorar la afinidad "*on-target*", es decir, contra el blanco deseado, sin embargo, durante un tratamiento (condición in vivo), el compuesto interactúa contra múltiples "*off-targets*" [164]. Con el fin de evitar efectos adversos producidos por un inhibidor de un blanco cuyo homólogo se encuentre en humano, se realizaron las siguientes búsquedas: utilizando el algoritmo blastp se compararon el proteoma de *Mtb* y de humano y luego el de *Mtb* y los proteomas de la microbiota intestinal. La selección de los 226 proteomas de la microbiota fueron tomados del Human Microbiome Project [165].

Todas las proteínas en la base de datos se sometieron a NCBI-BLAST (mi-valor<10 −7 ) contra proteoma humano (acceso de ensamblado ncbi GCF 000001405.36) para descartar aquellas proteínas que tuvieran blancos cercanos en el hospedador humanos . La puntuación de los metadatos (*off-target* humano) refleja este valor con la escala (1 - identidad de alineación máxima). Como criterios *off-target* una proteína bacteriana se consideró un homólogo humano cuando la similitud de secuencia fue>50% con cobertura de más del 50% de la proteína de consulta del patógeno y con un valor inferior a $10^{-4}$ utilizando la matriz BLOSUM62.

## Esencialidad

Todos los proteomas se procesaron contra la base de datos de genes esenciales (DEG) [166,167] para análisis de homología [166,167,168]. Los valores de corte de BLASTp utilizados fueron: evalue E=$1e^{-05}$, identidad ≥ 80% [169]. Si se encuentran genes homólogos esenciales a un gen de interés, es posible que el gen en cuestión también sea esencial, ya que las funciones codificadas por genes esenciales se conservan ampliamente en los microorganismos [167–169]. Los genes esenciales también se pueden identificar mediante experimentos específicos en escala genómica y se pueden agregar al patógeno target como metadatos. En este sentido, hemos incluido en Target-Pathogen datos de genes esenciales de *M. tuberculosis*, *Klebsiella pneumoniae* y *Staphylococcus aureus*, identificados por la hibridación del sitio de transposones [170,171].

## Análisis de redes metabólicas

Como se habló en la introducción, la reconstrucción de redes metabólicas permite identificar las reacciones de los puntos de acumulación o chokepoints [172,173] y establecimos que altos valores de centralidad de intermediación de nodos desde la perspectiva metabólica, reflejan

la participación de una reacción como intermediaria en muchas otras transformaciones, y su bloqueo generaría desequilibrio en muchas vías diferentes

Las redes metabólicas se crearon utilizando Pathway Tools v. 19.0 [174], utilizando como entrada los archivos Genbank de patógenos descargados de NCBI. Cada red reconstruida se exportó en formato de lenguaje de marcado de biología de sistemas (SBML). Con dicho archivo, se generó un grafo de reacción, donde los nodos representan reacciones y hay una arista entre dos nodos si el producto de una reacción se usa como sustrato en la reacción siguiente (como se describió en la introducción). Dicho grafo se analizó utilizando la librería de python NetworkX y con la misma se calcularon los valores de centralidad. Por otro lado se determinó las reacciones chokepoint utilizando scripts desarrollados durante este doctorado.

# Resultados y Discusión

El gran desarrollo de esta tesis fue la aplicación web Target-Pathogen (TP). Mediante su interfaz web http://target.sbg.qb.fcen.uba.ar/patho. TP permite seleccionar distintos genomas presentes actualmente en su base de datos para comenzar a explorarlos (listado preliminar en tabla 9). La interfaz ofrece un menú de búsqueda principal con varias opciones para recuperar los registros de genes (es decir, palabra clave, genes o vías). Finalmente, los genomas también se pueden explorar fácilmente por el número EC [49,175] o las diferentes categorías de Gene Ontology (GO) [176] usando Krona [177].

El sistema permite elegir el gen/proteína deseada y mostrar información específica de la misma. Dentro de sus correspondientes registros, se puede encontrar las distintas estructuras proteicas disponibles y al seleccionar un modelo o cristal en en dicha sección, el usuario será dirigido al módulo de visualización de estructuras que permite (i) seleccionar un bolsillo para visualización gráfica, (ii) mostrar heteroátomos, residuos relevantes PFAM y *Catalytic Site Atlas* (CSA)[160] asignados y (iii) mostrar residuos de unión a fármacos; permitiendo al usuario analizar la relevancia del bolsillo. La proteína mostrada está disponible para descargar como un archivo VMD [178]. Supongamos que encontramos una estructura interesante en el gen *pcaA*, así que simplemente escribimos *'pcaA'* en el campo 'Gene', para recuperar este registro. Dicha proteína de interés se ha cristalizado y su estructura fue depositada en PDB con el código 1l1e. Una vista de esta estructura se muestra en la Figura 26, donde no solo se puede observar la estructura, sino también sus pockets, la información de sus dominios y los residuos que se unen al ligando. Notar que el pocket número 2 es representado como esferas alfa (definen un volumen sin átomos de la proteína en su interior, correspondientes al pocket en cuestión) y las misma se superponen con el sitio de unión del fármaco cristalizado.

Tabla 9: Resultados de aplicar el pipeline de análisis en todos los genomas cargados en Target-Pathogen

| **Patógeno** | **Proteinas** | **Pathways** | **Proteínas con estructura** | **Cobertura Estructura** | **Proteínas con reacción** |
|---|---|---|---|---|---|
| *Mycobacterium tuberculosis H37Rv* | 4023 | 285 | 1999 | 49.69% | 902 |
| *Mycobacterium leprae Br4923* | 1956 | 221 | 756 | 38.65% | 361 |
| *Klebsiella pneumoniae Kp13* | 5736 | 396 | 2772 | 48.33% | 1201 |
| *Leishmania major Friedlin* | 8400 | 142 | 6535 | 77.80% | 1682 |
| *Wolbachia endosymbiont TRS of Brugia malayi* | 946 | 114 | 595 | 62.90% | 234 |
| *Trypanosoma brucei DAL972* | 9895 | 234 | 3286 | 33.21% | 386 |
| *Shigella dysenteriae Sd197* | 4294 | 386 | 1486 | 34.61% | 774 |
| *Schistosoma mansoni Puerto Rico* | 11802 | 132 | 7507 | 63.61% | 358 |
| *Toxoplasma gondii ME49* | 8322 | 241 | 1380 | 16.58% | 259 |
| *Plasmodium vivax Salvador I* | 5586 | 207 | 4907 | 87.84% | 219 |
| *Pseudomonas extremaustralis 14-3 substr. 14-3b* | 5919 | 393 | 1393 | 23.53% | 1622 |
| *Bartonella bacilliformis* | 1143 | 148 | 882 | 77.17% | 355 |
| *Trypanosoma_cruzi CL Brener* | 10338 | 98 | 4518 | 43.70% | 575 |
| *Leishmania infantum* | 8526 | 145 | 4682 | 54.91% | 791 |
| *Meloidogyne incognita* | 43718 | 249 | 6109 | 13.97% | 6057 |
| *Staphylococcus aureus subsp. aureus N315* | 2776 | 255 | 1672 | 60.23% | 1247 |
| *Achromobacter xylosoxidans* | 6087 | 361 | 4586 | 75.34% | 1529 |
| *Achromobacter insuavis AXX-A* | 5920 | 359 | 4498 | 75.98% | 1412 |

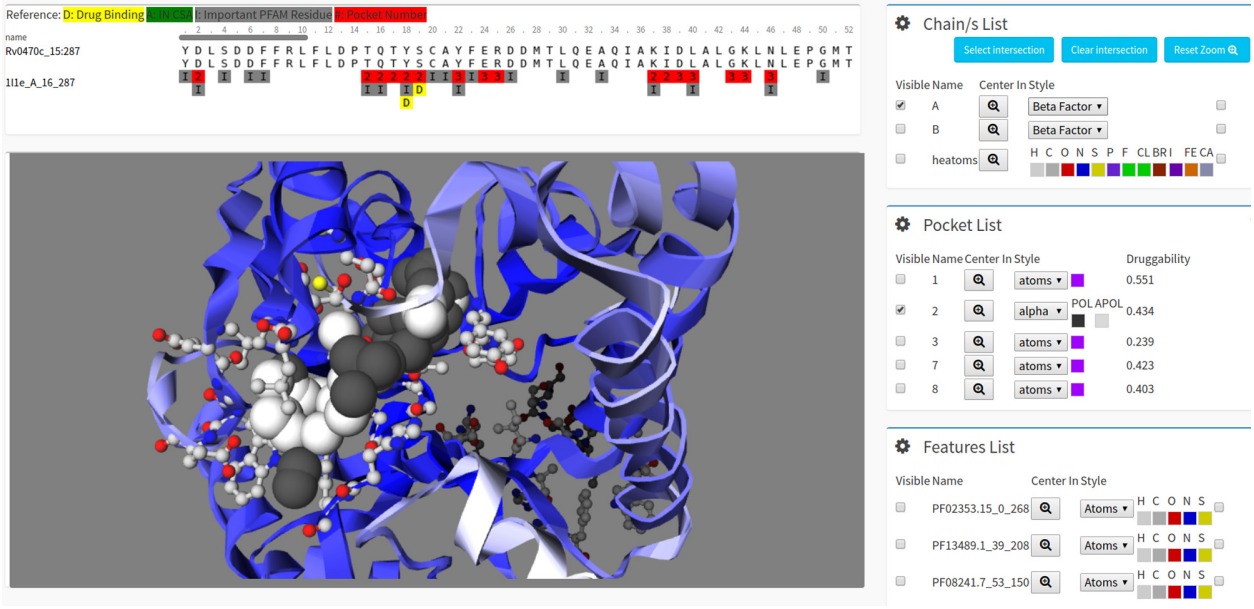



Figura 26.Visualización de estructuras proteicas en Target-Pathogen. Se muestran las diferentes pestañas de visualización que están disponibles al buscar proteínas. La tabla en el borde superior izquierdo muestra el alineamiento con la estructura cristalina correspondiente o el modelo de plantilla. Otras pestañas presentan datos relacionados con la estructura, incluido el módulo de visualización de bolsillo interactivo. El módulo de visualización permite (i) seleccionar qué bolsillo mostrar (marcando el campo Seleccionar bolsillo correspondiente), (ii) mostrar HETATMS presentes [179], CSA asignado o residuos relevantes de PFAM, (iii) mostrar la proteína en diferentes estilos. En el bolsillo apto para fármacos del ejemplo que se muestra a continuación, representamos las esferas alfa polares del bolsillo '1' en negro, mientras que sus esferas apolares están en blanco. Los HETATMS que se encuentran en la estructura cristalina se muestran con el estilo "*ball and sticks*" (bolas y palos) en diferentes colores.

## Selección de targets para patógenos

El enfoque de buscar en la literatura e intentar integrar mentalmente diversos criterios, puede volverse abrumador rápidamente. Alternativamente, Target-Pathogen puede ayudar al investigador a aplicar computacionalmente un conjunto de filtros para obtener una lista corta de proteínas que cumplan con los criterios predefinidos por el usuario, como la función de la proteína, el papel metabólico, la selección de targets, la drogabilidad estructural, la esencialidad y los experimentos ómicos disponibles. Para realizar esta tarea, se debe elegir un organismo específico y seleccionar una serie de filtros que permita seleccionar blancos en función de si cumplen o no un conjunto de criterios definidos. Como ejemplo, mostramos cómo obtener una lista de proteínas con características atractivas para el direccionamiento de fármacos en *Leishmania major* [180]. Simplemente, aplicando un conjunto de filtros obtuvimos 381 proteínas que son esenciales (tiene un hit/homólogo en DEG), altamente drogables (DS> 0.7), que tienen residuos catalíticos dentro del bolsillo informado en la base de datos CSA y no tienen homólogos cercanos en el genoma humano (*off-target*>0,6). Curiosamente, más de una cuarta parte (105) de estas proteínas fueron

anotadas con actividad quinasa (*GO Activity* GO:0016301), previamente reportadas como potenciales candidatas para el desarrollo de nuevos fármacos contra este protozoo [181]. Si filtramos las proteínas anotadas con el término GO Biological Process 'transducción de señales intracelulares', la base de datos arroja una breve lista de proteínas que cumplen todos estos criterios. La mayoría de estas proteínas son quinasas y entre ellos MKK se informó previamente como un target de drogas en Leishmania [182]. Mediante el uso de filtros simples, hemos podido seleccionar una breve lista de proteínas relevantes para desarrollar nuevos fármacos contra *L. major* con estructura disponible e información sobre los bolsillos drogables (Figura 27).

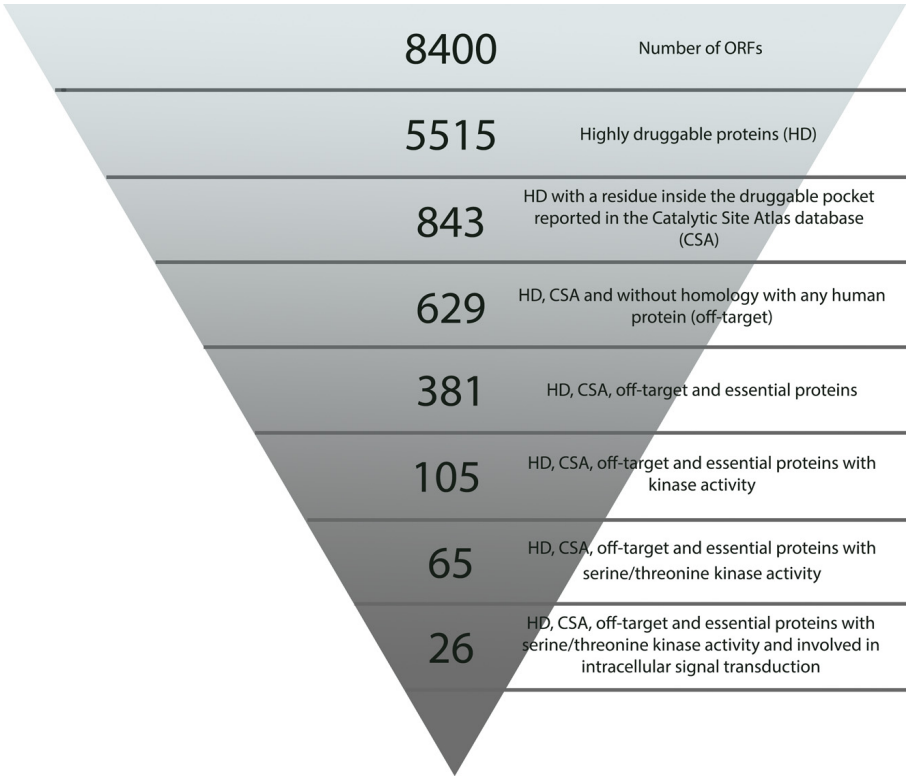



Figura 27.Número de proteínas en *Leishmania major* con propiedades deseables para blancos farmacológicos.

## Clasificación y priorización de blancos farmacológicos con Target-Pathogen

La base de datos Target-Pathogen no solo permite consultar y filtrar proteínas, también ofrece la funcionalidad de asignar un valor de peso numérico a diferentes propiedades de proteínas para crear una función de puntuación (SF o *scoring function*) definida por el usuario. Además, los usuarios pueden combinar diferentes filtros con el SF para obtener una lista particular

de genes clasificados según un criterio diseñado por el usuario. A continuación mostraremos un ejemplo aplicado a *Staphylococcus aureus*, sobre la cepa N315.

En relación a los filtros, primeramente se descartaron todas las proteínas para las cuales no se pudo obtener un modelo estructural. En segundo lugar, de las proteínas obtenidas a partir de este primer filtro sólo se consideraron aquellas que tuvieran por lo menos una cavidad drogable, es decir, con una puntuación de drogabilidad mayor a 0,5 (DS>0,5). Por último, las proteínas con ortólogos en el genoma humano se descartaron para minimizar las posibilidades de inhibición cruzada y toxicidad de un fármaco con proteínas del hospedador humano. Se puede ver el filtro final en la figura 28.

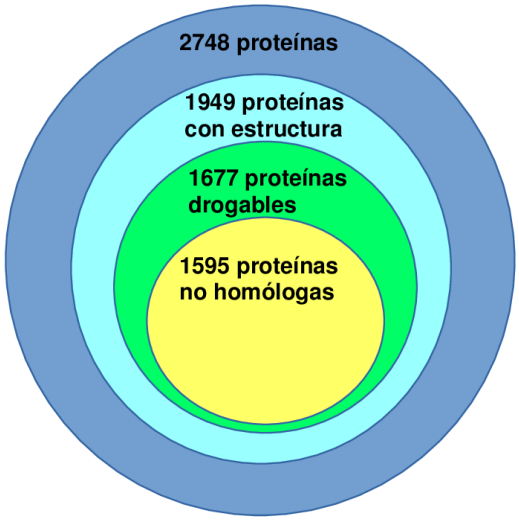



Figura 28: Resultado de los filtros de estructura, drogabilidad y *off-target* en *Staphylococcus aureus* N315

A las 1595 proteínas filtradas se les aplicó la función de scoring de la ecuación 1. Estas características incluyen alto grado de conservación en patógenos relacionados, su esencialidad y la contextualización de su función en una o varias vías metabólicas (centralidad de intermediación y reacciones *chokepoints*). Asimismo, también se tuvo en cuenta el grado de homología de dichas proteínas con las presentes en la flora intestinal humana. Es decir, si una proteína de *S. aureus N315* resulta homóloga de proteínas de un gran número de microorganismos de la microbiota, no será una buena candidata como blanco molecular debido a los posibles efectos adversos que puede ocasionar inhibir dicha proteínas ya que no sólo se afectaría a la proteína patógena sino que también a sus homólogas en la flora normal del huésped.

$$SF = \frac{Chk + C + E + Cons + (1 - \%Mb)}{5}$$

Ecuación 1: Función de scoring para la priorización de blancos en *Staphylococcus aureus* N315. Chk indica 1 si la proteína es *chokepoint* y cero si no lo es, C la centralidad de la reacción, E si es esencial o no (1 / 0 respectivamente), Cons es la conservación del gen en 32 cepas de *Staphylococcus Aureus* CC5 (con genoma cerrado a nivel cromosoma, descargado de NCBI) y %Mb en el cual dicho gen está conservado en la microbiota de intestino humana

A partir de estos datos incluidos en la función de puntaje fue que se construyó una lista de posibles proteínas de *S. aureus N315* candidatas a ser blancos moleculares para el desarrollo de agentes antimicrobianos que puede verse parcialmente en la tabla 10.

Tabla 10: Top 10 blancos de *Staphylococcus aureus* N315 aplicando SF de la ecuación 1

| Ranking | Gen | Producto | Vía metabólica | (Chk) | (C) | (E) | (Cons) | Presencia en microbiota (%) | Puntaje |
|---|---|---|---|---|---|---|---|---|---|
| 1 | *gatD* | glutamina amidotransferasa | Biosíntesis de peptidoglicano II | 1 | 0.71 | 1 | 1 | 24.00% | 0.9275 |
| 2 | *metG* | metionina - ARNt ligasa | Carga de aminoacil-ARNt | 1 | 0.25 | 1 | 1 | 16.00% | 0.8125 |
| 3 | *murG* | undecaprenil fosfato-muramil pentapéptido beta-N-acetilglucosaminiltransferasa | Biosíntesis de peptidoglicano II | 1 | 0.15 | 1 | 1 | 26.00% | 0.7875 |
| 4 | *glyQS* | glicina - ARNt ligasa | Carga de aminoacil-ARNt | 1 | 0.12 | 1 | 1 | 12.00% | 0.78 |
| 5 | *sdhA* | flavoproteína succinato deshidrogenasa | Ciclo ácido tricarboxílico I (TCA I) | 1 | 0.12 | 1 | 1 | 10.00% | 0.78 |
| 6 | *folK* | 2-amino-4-hidroxi-6-hidroximetildihidropteridina difosfoquinasa | Biosíntesis de tetrahidrofolato | 1 | 0.11 | 1 | 1 | 36.00% | 0.7775 |
| 7 | *tarL* | 2-C-metil-D-eritritol 4-fosfato citidililtransferasa | | 1 | 0.1 | 1 | 1 | 28.00% | 0.775 |
| 8 | *tar1* | 2-C-metil-D-eritritol 4-fosfato citidililtransferasa | | 1 | 0.1 | 1 | 1 | 26.00% | 0.775 |
| 9 | *ddlA/ddl* | D-alanina - D-alanina ligasa | Biosíntesis de peptidoglicano II | 1 | 0.1 | 1 | 1 | 18.00% | 0.775 |
| 10 | *prkC* | serina / treonina proteína quinasa | | 1 | 0.09 | 1 | 1 | 49.00% | 0.7725 |

En el listado obtenido, la proteína que aparece al tope del ranking glutamina amidotransferasa producto del gen *gatD* forma parte de una vía metabólica involucrada en la formación del peptidoglicano. Dicha enzima resultó drogable, esencial, conservada entre las distintas cepas analizadas, principalmente relacionada con reacciones importantes desde el punto de vista metabólico y además reacción *chokepoint*. También se encontró *murG*, implicada en la misma vía, y el cual ya ha sido blanco de desarrollo de algunos agentes antibacterianos [183]

También se listaron proteínas asociadas a vías de ARNt como es el caso de metionina-ARNt ligasa (enzima codificada por el gen *metG)* y la glicina-ARNt ligasa codificada por el gen *glyQS.* Las aminoacil-tRNA sintetasas son enzimas que catalizan la reacción de aminoacilación al unir covalentemente un aminoácido a su ARNt análogo en el primer paso de la

traducción de proteínas. Más allá de esta función clásica, también se sabe que estas enzimas tienen un papel en varias vías metabólicas y de señalización que son importantes para la viabilidad celular. El estudio de estas enzimas es de gran interés debido a su papel fundamental en el crecimiento y la supervivencia de un organismo [184].

## Priorización de vías para el desarrollo de fármacos

La reconstrucción bioinformática de la red metabólica patógeno proporciona datos útiles para explorar nuevos targets moleculares. Target-Pathogen permite a los usuarios seleccionar y estudiar proteínas en el contexto de sus vías metabólicas. Además, les permite clasificar y priorizar rutas completas como buenos candidatos para terapias novedosas con un criterio bien definido. Una ventaja fundamental de estudiar el contexto metabólico de blancos putativos es que se espera que los resultados permitan el diseño de posibles terapias combinadas (dirigidas a más de un target de la misma ruta metabólica). Para la priorización de la mismas podemos parametrizar una función de puntuación que tenga en cuenta las características metabólicas como la proporción de chokepoints, la centralidad y la integridad de la vía (es decir, el número total de reacciones de una vía asociada con un gen/número total de reacciones presentes en la ruta). Un ejemplo completo de esto se estudiará en la aplicación sobre Klebsiella pneumoniae

## Búsqueda de blancos en *Mtb*

La respuesta inmune del huésped a tuberculosis (TB) se basa en la fagocitosis de los bacilos por los macrófagos conducen a la formación de granulomas que detienen la replicación bacteriana. Dentro de los macrófagos, las bacterias se enfrentan a condiciones de hipoxia, óxido nítrico inducible derivado de la síntesis de NO y privación/falta de nutrientes. El bacilo, en respuesta, cambia a un estado no replicativo (latente), donde pueden permanecer ocultos y vivos durante décadas [185]. También es importante resaltar que los medicamentos antituberculosos existentes son ineficaces contra la *Mtb* latente, y que hay una falta de blancos bien definidos específicos para este estado [186,187].

En el presente trabajo [171], se realizó una priorización de blancos para tratar TB en estado latente haciendo uso de la actividad micobactericida dependiente de la concentración de especies reactivas de nitrógeno y oxígeno (ERON) [188]. Esto se apoya en el hecho que los compuestos similares al nitroimidazol bicíclico como la Pretomanida (aprobada el 2019) y Delamanida, matan las células de *Mtb* en estado no replicativo, por medio de la liberación intracelular de NO, subrayando la relevancia de ERON para combatir infecciones de TB [189–192] . Con base en estas observaciones, se procedió a intentar identificar los blancos enzimáticos de ERON, que posteriormente permitan diseñar drogas que actúen de manera conjunta con los macrófagos que atacan y matan la TB en la fase latente [193].

En primer lugar se reconstruyó la vía metabólica de *Mtb* para analizar el contexto metabólico de cada proteína del patógeno. La misma fue curada manualmente, (especialmente en las vías metabólicas donde se encontraron blancos de interés) mediante búsqueda en literatura y utilización de diferentes herramientas bioinformáticas y bases de datos (BLAST, UniProt, Metacyc, KEGG [125]) para evaluar homólogos funcionales de proteínas para su incorporación a las distintas vías de la red metabólica. Una vez exportada a formato SMBL por PathwayTools, antes de ser convertida en un grafo para el cálculo de centralidad y reacciones chokepoint, fue generada una lista de todos los compuestos presentes en la red, y recolectamos su frecuencia como participantes de reacción usando un script de Python. Los que más aparecieron con frecuencia como participantes de reacción que se consideran compuestos ubicuos (como ATP, cofactores, agua) y se descartaron de la red, ya que pueden crear enlaces artificiales en la representación basada en gráficos de la red, por estar involucrados en muchas reacciones que no están necesariamente relacionadas. Un total de 51 compuestos fueron filtrados antes de la transformación de la red metabólica en un grafo de reacción.

En la Tabla 11 se muestran algunas de las métricas que consideramos importantes en la red de *Mtb*, que nos hacen dar una idea de su tamaño y porcentaje del genoma (<25%) que la compone. También, en la Figura 29 se muestra su distribución y como solo unos pocos nodos poseen una alta centralidad.

Tabla 11: Números generales de la red metabólica construida

| **Proteínas** | **Reacciones** | **Chokepoints** | **Pathways** |
|---|---|---|---|
| 902 | 1229 | 452 | 285 |

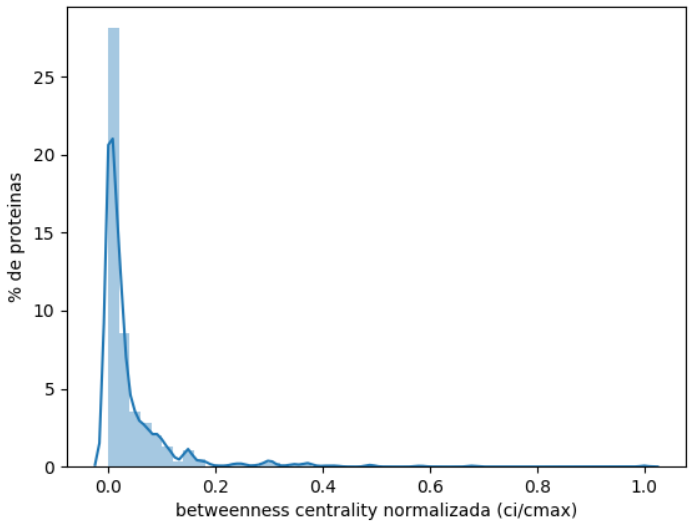



Figura 29: Distribución de centralidad para las proteínas en la red metabólica. Se puede ver que solo unas pocas tienen centralidad alta, y es en general una característica deseada para un buen blanco.

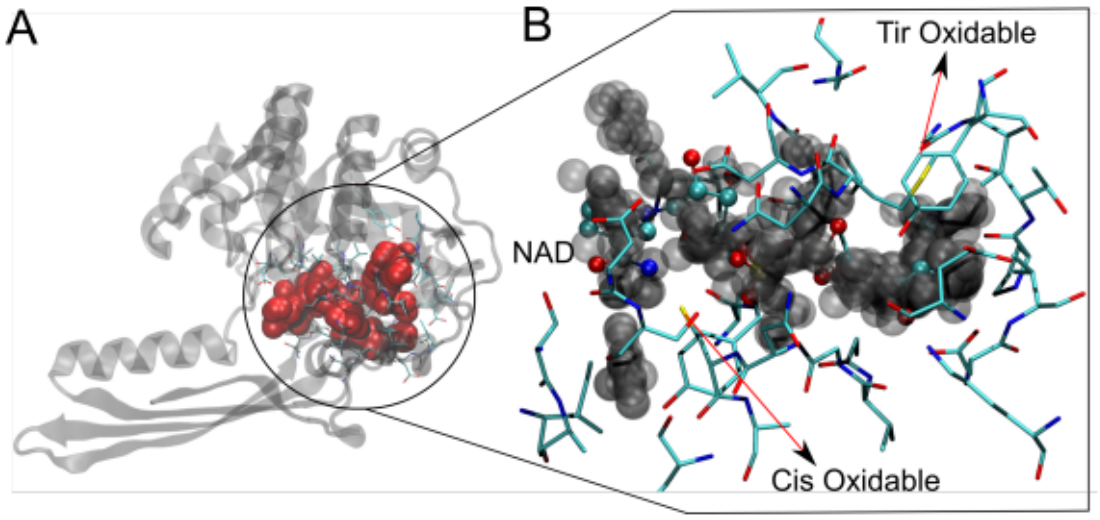
Para la selección de blancos, primero se realizó una priorización de vías metabólicas con una función de puntuación para determinar qué vías son relevantes en la fase latente de *Mtb*. Como filtro, se tomaron solo las pathways “drogables”, es decir, si al menos uno de las proteínas implicadas eran susceptibles de fármacos. Y como función de puntuación, se usó la ecuación 2.

$$SF = \frac{1}{2} * \left( \frac{NR + Cy + Chk + Es + C}{5} + \frac{S + H + St + I}{4} \right)$$

Ecuación 2: Función de puntuación para vías metabólicas de *Mtb*. NR es el número total de reacciones presentes en la vía correspondiente normalizada por la cantidad de reacciones que componen la vía con más reacciones dentro del metabolismo. Cy es la relación entre la centralidad del nodo y el nodo con la mayor centralidad en toda la red. Chk es la relación entre el número de choke-points y el número total de reacciones presentes en la vía. E es la relación entre el número de genes esenciales presentes en las vías y el número total de genes asociados a la vía. C = “completitud” es la relación entre el total número de reacciones de la vía asociada con un gen y el número total de reacciones enzimáticas

presentes en la ruta. S es la relación entre el número de genes asociados a condiciones de estrés [194,195,196] presentes en la vía y el número total de genes asociado a la vía. H es la relación entre el número de genes asociados a la hipoxia [197,198,199] y el número total de genes asociados a la vía. St es la razón entre el número de genes asociados a la inanición [200,201] y el número total de genes asociados a la vía. Finalmente, I es la relación entre el número de genes asociados a la infección [202,203,204,205] y el total de genes asociados a la vía.

Una vez priorizadas las vías, dentro de cada una se seleccionó el blanco con mayor potencial dentro de cada una, ponderando: sobreexpresión en hipoxia, hambruna, estrés ERON e infección y drogabilidad, dando lugar a nuestra lista final de blancos de la Tabla 12. De los mismos, muchos ya han sido destacados por Chandra's Lab y TDR Targets [206,207,208] y se marcaron en la tabla como revalidadas. Pero, hasta el momento de la publicación de esta tesis, ninguno de estos estudios apuntaba a proteínas en la vía de síntesis de micotiol (como *mshB* e *ino1*), relevantes para el balance redox en micobacterias, *ino1* es la proteína que convierte la glucosa-6-P en 1D-mio-inositol 3-fosfato, se ha reportado como esencial tanto en alto rendimiento [209,210] así como estudios mutantes individuales [211]. Es parte del regulón *DosR* y también es altamente sobreexpresado en condiciones de hambruna / falta de nutrientes. Como se muestra en Figura 30 el cristal de pdb 1GR0 (estructura de *ino1*) presenta un bolsillo drogable (puntuación DS de 0,719) que se superpone con el sitio de unión de NAD, un sitio que puede albergar compuestos similares a fármacos en otras proteínas como *InhA*. También muestra 2 residuos sensibles a ERON Tyr145 y Cys26, cuyo papel no ha sido completamente dilucidado. Claramente, *ino1* alberga todas las características de un blanco ideal (esencial, drogable, sobre expresada en hipoxia, ERON, hambruna y primer paso en la vía de biosíntesis de micotiol).



C

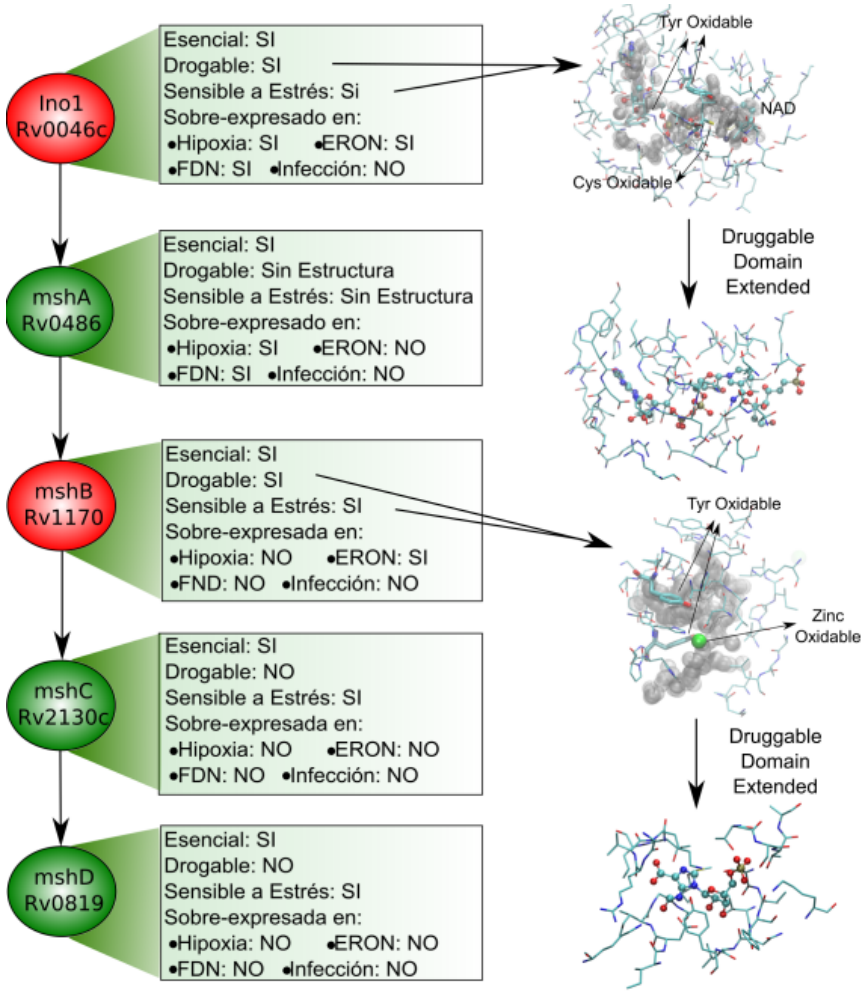



Figura 30: Estructura de Inositol-3-Phosphate Synthase. Vista del plegado de *ino1* con el bolsillo drogable destacado en esferas rojas (PDBID 1GR0). B) Acercamiento del bolsillo drogable superpuesto con la estructura de NAD. C) Vía de síntesis del micotiol. En rojo se muestran las proteínas más relevantes de la vía *ino1* y *mshB*. A un lado se muestra una representación del bolsillo drogable. FDN: falta de nutrientes. Tomado de De felipe et. al. 2016 [171]

Otro caso, no mencionado anteriormente, es ejemplificado por L-D transpeptidasa (LDTP1) que participa en la formación del peptidoglicano en la pared celular de las micobacterias (y por lo tanto relacionado con la biosíntesis del ácido micólico) fundamental para la resistencia *in vivo*. LDTP1 es una enzima involucrada en la formación de enlaces de reticulación de peptidoglicano y, por lo tanto, es esencial. Se sobreexpresa en todas las condiciones, particularmente en presencia de ERON. El bolsillo drogable (DS puntuación de 0,701), que también es el sitio activo que contiene un residuo Cys en la posición 226, que es el nucleófilo en la reacción enzimática, fuertemente argumentando a favor de su potencial inhibición por ERON. También se ha sugerido que la actividad catalítica de LDTP1 podría ser inhibida por compuestos betalactámicos [212].

Por otro lado, hay blancos predichos por TargetTB que son no predichos con la metodología aplicada en este organismo. En general, esto se debe al hecho de que nuestro análisis prioriza la drogabilidad estructural y la expresión de blancos para la fase latente de la infección por *Mtb*. Por ejemplo, ambos *DdlA* (*Rv2981c*) y *EmbA* (*Rv3794*), blancos conocidos de la cicloserina [213] y etambutol [214], respectivamente, son drogables y esenciales, pero carecen de regulación positiva durante la infección que imita las condiciones de fase latente. Por otro lado, dos proteínas que pertenecen a la vía de biosíntesis de micolil-arabinogalactano-peptidoglicano, *AftA* (*Rv3792*) y *AftB* (*Rv3805c*), no fueron predichos por nuestra metodología debido a la falta de estructuras o modelos disponibles.

Tabla 12: Blancos moleculares revalidados y novedosos, resultado de la utilización de nuestra metodología para la priorización de blancos en *Mtb*. DS: Score de drogabilidad..

| Rv | Proteína | Antecedentes | DS | Vía metabólica |
|---|---|---|---|---|
| *Rv0046c* | *Inositol-3-phosphate synthase* | nuevo | 0.719 | Síntesis de myo-inositol |
| *Rv3227* | *3-phosphoshikimate 1-carboxyvinyltransferase* | nuevo | 0.724 | Biosíntesis de corismato a partir de 3-deshidroquinato |
| *Rv3340* | *O-acetylhomoserine aminocarboxypropyltransferase* | nuevo | 0.635 | Síntesis de homocisteína |
| *Rv2246* | *3-oxoacyl-[acyl-carrier-protein] synthase 2* | nuevo | 0.709 | Biosíntesis de micolatos |
| *Rv2217* | *Octanoyltransferase* | nuevo | 0.703 | Biosíntesis e incorporación de lipoatos I |
| *Rv1018c* | *Bifunctional protein GlmU* | nuevo | 0.911 | Biosíntesis de UDP-N-acetil-D-glucosamina I |
| *Rv1465* | *Rv1465* | nuevo | 0.926 | biosíntesis de grupos de hierro y azufre |
| *RV1170* | *1D-myo-inositol 2-acetamido-2-deoxy-alpha-Dglucopyranoside deacetylase* | revalidado | 0.781 | Biosíntesis de micotiol |
| *Rv1285* | *Sulfate adenylyltransferase subunit 2* | revalidado | 0.891 | reducción de selenato |
| *Rv3464* | *dTDP-glucose 4,6-dehydratase* | revalidado | 0.676 | Biosíntesis de dTDP-L-ramnosa |
| *Rv1484* | *Enoyl-[acyl-carrier-protein] reductase [NADH]* | revalidado | 0.919 | Biosíntesis de 8-amino-7-oxononanoato I |
| *Rv2225* | *3-methyl-2-oxobutanoate hydroxymethyltransferase* | revalidado | 0.937 | Biosíntesis de fosfopantotenato I |
| *Rv2276* | *Mycocyclosin synthase* | revalidado | 0.887 | Biosíntesis de micociclosina |

## Aplicaciones de Target-Pathogen a otros organismos

### Aplicación en Kp13

El género *Klebsiella* engloba una variedad de enterobacterias Gram negativas. Dentro de las especies de este género se encuentra *Klebsiella pneumoniae*, un bacilo, inmóvil, anaerobio facultativo que presenta una prominente cápsula de lipopolisacáridos que envuelve enteramente su superficie. Se encuentra mayoritariamente en el intestino contribuyendo a la flora normal de los seres humanos, pero cuando se encuentra fuera del tracto intestinal puede causar severas patologías.

Diversas cepas pueden poseer una gran cantidad de genes de evasión y resistencia que le permiten evadir el sistema inmune del hospedador y sobrevivir a la acción de los antimicrobianos [215,216]. Cuando las infecciones son producidas por cepas resistentes a antibióticos, los cuadros empeoran gravemente y el tratamiento clínico se complica. A la fecha se han descrito cepas resistentes a la mayoría de las drogas utilizadas actualmente en la práctica clínica. Distintas cepas de esta especie son productoras de enzimas betalactamasas de espectro extendido (ESBLs). Las ESBLs les confieren resistencia a los antibióticos betalactámicos, como penicilinas, y cefamicinas incluyendo los que contienen el grupo oximino, como las cefalosporinas de tercera y cuarta generación, y el aztreonam [217]. Una de las alternativas para combatir estas cepas resistentes eran los carbapenémicos, situación que ha cambiado debido a la aparición de resistencia asociada a las carbapenemasas de *Klebsiella pneumoniae* (KPC) [218].

A continuación se presentan los resultados de la búsqueda de blancos en *Klebsiella pneumoniae* cepa Kp13 aplicando Target-Pathogen. Kp13 fue aislada durante un brote nosocomial en una unidad de cuidados intensivos ocurrido en 2009 en el sur de Brasil. Esta cepa es resistente a muchos antibióticos, incluida la polimixina B (medicamento de "último recurso"). A parte de los datos de anotación, metabolismo, estructura y drogabilidad, se agregó información de transcriptómica en condiciones de exposición a polimixina y datos de conservación en distintas especies patogénicas (de *Klebsiella pneumoniae*). Para la priorización de blancos, primero se aplicó un filtro que nos permitiera descartar proteínas no drogables (DS<0.5) y con homólogos cercanos en el genoma humano (Figura 31). Sobre esos blancos se utilizaron 3 SF distintas (Ecuación 3), para luego analizar el consenso de los resultados (Figura 32). La primera utiliza información de esencialidad, conservación y contexto metabólico. La segunda incorpora información de expresión en polimixina y la tercera un criterio *off-target* a proteínas de la microbiota (con la intención de reducir el impacto sobre la misma).

$$SF_1 = \frac{E_{mgh} + E_{kpn}}{2} + C_v + C_y + chk,$$

$$SF_2 = \frac{\frac{E_{mgh}+E_{kpn}}{2} + C_v + C_y + chk}{4} + P_b,$$

$$SF_3 = \frac{\frac{E_{mgh}+E_{kpn}}{2} + C_v + C_y + chk}{4} - G_M,$$

Ecuación 3: SFs para KP13. $E_{mgh}$: esencial en *K. pneumoniae* MGH 78578 [219] $E_{kpn}$: Esencial en *K. pneumoniae* Kp13. $C_V$: Conservación en genomas de *K. pneumoniae* patogénicos. $C_y$: centralidad de la reacción. Chk: si el gen es chokepoint. Pb: sobreexpresión en polimixina (1 si 0 no). $G_M$: 1 - proporción de organismos de la microbiota que poseen un homólogo.

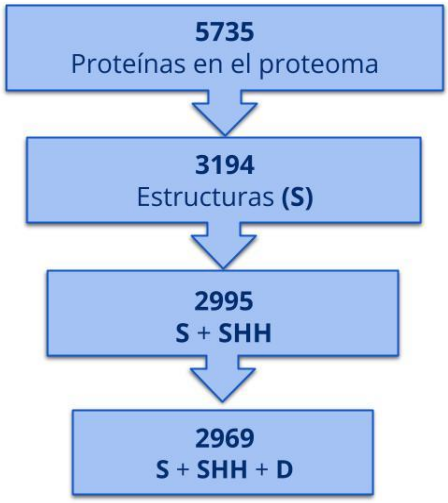



Figura 31: Filtro sobre proteínas de KP13. S: filtro con estructura. SHH: Filtro sin homólogo en humanos. D: drogabilidad > 0.5

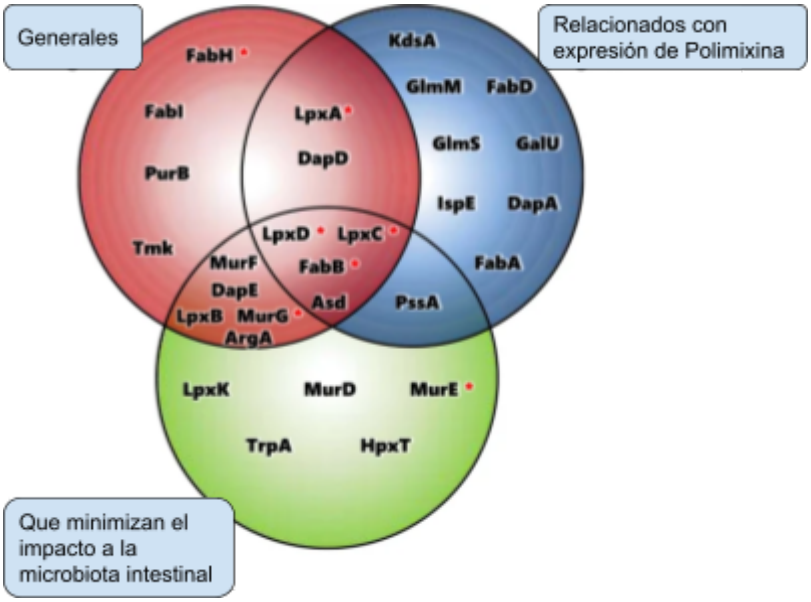



Figura 32: Resultados de las 3 SFs de la Ecuación 3.

Tabla 13: Se omite el sufijo del locus Kp13 de K. pneumoniae 'KP13_'; *drogabilidad de la proteína considerando el bolsillo de mayor puntuación. # B: biosíntesis. & Datos recopilados de http://www.drugbank.ca para ortólogos de cada proteína con relevancia estudiada por blanco. ND, no determinado. Drogabilidad: DS calculado por FPocket del *pocket* más drogable. Esencialidad: Hit tiene un homólogo en la base de datos DEG o en el genoma de KP MGH marcado como tal. *Chokepoint*: si el gen cataliza una reacción *chokepoint*. Centradilidad: *betweenness centrality* de la reacción de la proteína. Kps patológicas: proporción en la cual está conservada la proteína en los distintos genomas de *K. pneumoniae* evaluados.

| Locus | Gen | Producto | Tamaño (aa) | drogabilidad * | Esencialidad | Choke point | Centralidad de la red | En Kps pato-genicas (%) | Vias metabolicas # | Inhibidores en DrugBank & |
|---|---|---|---|---|---|---|---|---|---|---|
| 01032 | fabB | 3-oxoacyl-[acyl-carrier-protein] synthase 1 | 407 | 0.74 | Si | Si | 0.64 | 100.0 | Biotin [B], Fatty acids [B] | Aprobado (DB01034) |
| 01899 | lpxC | UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase | 306 | 0.78 | Si | Si | 0.29 | 100.0 | Lipopolysaccharide [B] | Experimental (DB07861) |
| 01800 | lpxD | UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase | 341 | 0.82 | Si | Si | 0.07 | 100.0 | Lipopolysaccharide [B] | ND |
| 00662 | asd | Aspartate-semialdehyde dehydrogenase | 368 | 0.58 | Si | Si | 0.05 | 100.0 | L-lysine [B], L-threonine [B], L-methionine [B], L-homoserine [B] | Experimental (DB03502) |

En la Tabla 13 se listan las proteínas que fueron priorizadas entre las primeras 15 utilizando las 3 SFs. Es decir que estas proteínas pueden ser utilizadas como blancos generales o en sinergia con PB y que drogas diseñadas para estos blancos tendrían poco impacto en la microbiota. Dentro de los blancos que podemos destacar encontramos proteínas involucradas en rutas metabólicas asociadas a la biosíntesis de ácidos grasos y lipopolisacáridos (LPS) y proteínas de peptidoglicano involucradas en los componentes de la biosíntesis de ácidos grasos, las cuales permiten la homeostasis de la membrana bacteriana [220,221]. En particular, la enoil-[acil-carrier-protein] reductasa [NADH] (*FabI*) ha sido el blanco del desarrollo de nuevos agentes antibacterianos [149,222]. La síntesis de lipopolisacárido un componente esencial de la membrana externa Gram-negativa, con las enzimas citoplasmáticas *LpxA*, *LpxC* y *LpxD* involucradas en los pasos iniciales de la producción de lípido A a través de la vía metabólica de Raetz [223]. También, dichas proteínas cumplen la mayoría de los criterios definidos anteriormente, que hacen que las mismas sean atractivas para el direccionamiento de fármacos (Figura 33) . En las últimas dos décadas, se han desarrollado numerosos inhibidores de *LpxC* como agentes bactericidas contra organismos gram negativos patógenos, incluido *K. pneumoniae*, con revisiones exhaustivas recientes que detallan estos desarrollos [224,225]. Sin embargo, solo una molécula (ACHN-975) entró en ensayos clínicos en humanos, y luego se suspendió durante la Fase I debido a efectos inflamatorios no deseados en el lugar de la inyección [226]. Sin embargo, la investigación de los inhibidores de *LpxC* no se ha interrumpido y recientemente se propuso un nuevo inhibidor que promete ser valioso para el desarrollo clínico (LPC-069, un inhibidor de *LpxC* a base de bi-fenilacetileno) para combatir un amplio panel de aislados clínicos Gram-negativos, incluidas varias cepas multirresistentes y extremadamente resistentes a los medicamentos sin efectos adversos conocidos en ratones [227]. En consecuencia, nuestros resultados mostraron que el gen que codifica la proteína *LpxC* se conserva en todas las cepas patógenas estudiadas de *K. pneumoniae* y no presenta homólogos cercanos dentro del proteoma humano.

Es importante remarcar que los blancos identificados podrían ser de interés en una perspectiva de terapia combinada cuando se trata de infecciones por Kp resistente, posiblemente actuando sinérgicamente con otros fármacos, de manera que involucre un compuesto no antimicrobiano con un bactericida. Como prueba de este concepto, en otras enfermedades infecciosas, como la bacteriemia por *Pseudomonas aeruginosa*, se ha propuesto la combinación de inhibidores de proteínas de eflujo (como la fenil-arginina-β-naftilamida) y

quelantes de hierro para controlar el proceso de infección en vista de la sobreexpresión del sistema de eflujo MexAB-OprM durante la privación de hierro [228] .

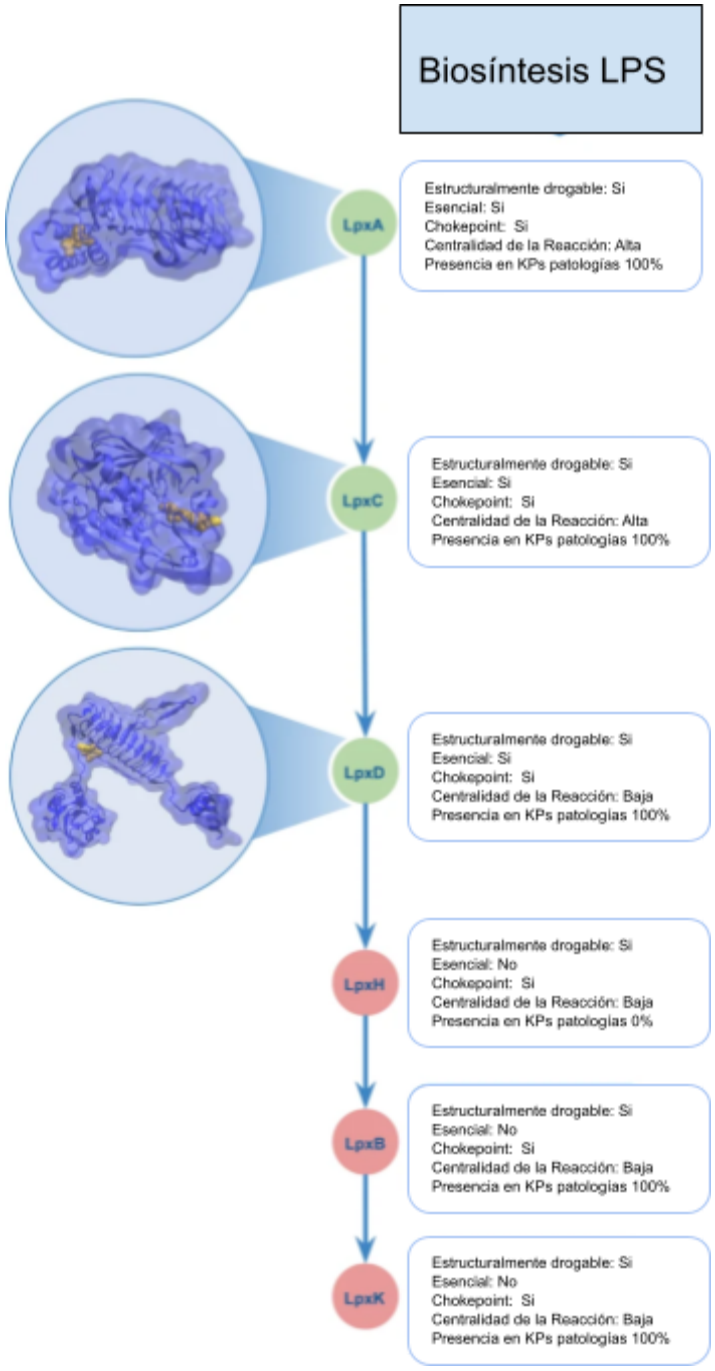



Figura 33: Proteínas anotadas de la vía metabólica de la síntesis de Lipopolisacáridos. En verde figuran las proteínas que aparecen en nuestro ranking. La presencia en genomas de K. pneumoniae patógenos se hizo en base a 38 secuencias cerradas a nivel cromosoma obtenidas de NCBI

## Aplicación en Bartonella bacilliformis

Bartonella bacilliformis (Bb) es una bacteria patógena gram negativa responsable de la enfermedad de Carrión (EC) que causa anemia hemolítica y lesiones cutáneas [229] . Esta infección es endémica en algunas zonas de Perú, Colombia y Ecuador. Se asocia principalmente con la pobreza, los cambios climáticos y al poco apoyo financiero que recibe [230] . Esta enfermedad tiene dos fases clínicas diferentes. La primera, denominada fase temprana o aguda (fiebre de Oroya), cuyos síntomas incluyen fiebre y anemia profunda y generalmente aparecen alrededor de 60 días después de la infección. La segunda fase, crónica o verruga peruana, se caracteriza por la aparición de erupciones dérmicas conocidas como verrugas.

Si bien se ha mostrado que Bb es altamente susceptible a una amplia gama de fármacos in vitro, como los betalactámicos (incluidas las penicilinas y las cefalosporinas), los aminoglucósidos y las quinolonas [231], se observado que existen pacientes en los que el tratamiento con antibióticos no tiene éxito. Aunque el ciprofloxacino, es el fármaco de elección para adultos en la fase aguda de la enfermedad de Carrión, se han descrito resistencias a quinolonas producidas por mutaciones o sustituciones de aminoácidos en blancos moleculares [232]. Por otro lado, mutaciones que confieren resistencia a ciprofloxacino, rifampicina y eritromicina se han caracterizado molecularmente en Biswas, Raoult & Rolain et al [232]. El fármaco de elección para el tratamiento de la fase eruptiva de la EC es la rifampicina; sin embargo, con eficacia variable [233]. Cabe destacar que la fase crónica no responde ni al tratamiento con cloranfenicol ni con penicilina [234].

Utilizando Target-Pathogen se realizó la priorización de proteínas blanco, utilizando como organismo modelo *B. bacilliformis* USM-LMMB07 [235], recolectada en 2011 en el norte de Perú (Huancabamba, Piura) e identificada como el agente causal de la fiebre de Oroya.

De las 1143 proteínas predichas que forman el proteoma de Bb USM-LMMB07, obtuvimos 882 modelos estructurales, y pudimos predecir bolsillos altamente susceptibles a fármacos para 235 de ellos (Figura 34). La reconstrucción de la red metabólica de esta cepa permitió la identificación del complemento metabólico y las actividades enzimáticas realizadas por Bb, así como importantes métricas topológicas en la red metabólica. Aplicando una serie de filtros en Target-Pathogen (Figura 34) se pudo obtener el conjunto final de seis posibles blancos para el desarrollo de nuevos fármacos (tabla 14). Todas estas proteínas son altamente drogables, no presentan homólogos cercanos en humanos, son relevantes desde el punto de vista metabólico (están asociados a reacciones *chokepoint* con alta centralidad) y el molde con

el que se modeló la estructura proteica, tiene un pocket capaz de unir a compuestos tipo fármacos, que coincide con el bolsillo predicho para el modelo, todas características que las convierten en excelentes blancos farmacológicos.

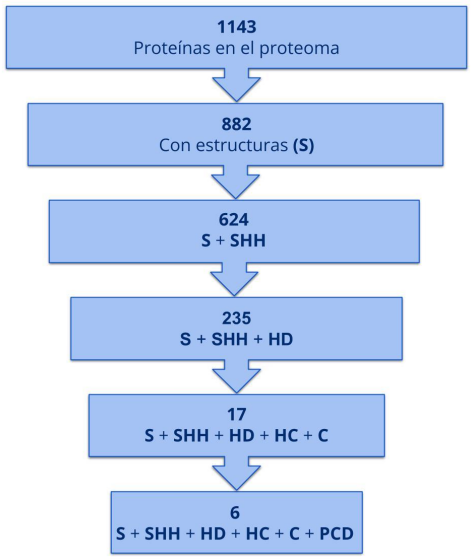



Figura 34: Esquema general del proceso de priorización de blancos moleculares en *B. bacilliformis*. Se aplicaron los siguientes filtros de manera sucesiva. SHH: Sin homólogo en Humano. HD: Altamente Drogable. HC: Alta Centralidad. C: Asociado a Reacción Chokepoint. PDC: El Pocket predicho une a una Droga en la estructura experimental que se utilizó como molde para el modelado.

Uno de ellos, *FabI*, participa en los procesos de síntesis de ácidos grasos (FAS). Las proteínas de esta vía, como *FabA*, *FabB*, *FabD*, *FabI* y *FabH*, tienen un papel esencial durante la biosíntesis de membranas de fosfolípidos, lipoproteínas y LPS y representan blancos atractivos debido a las diferencias estructurales entre las proteínas humanas y bacterianas [236]. Con dos inhibidores disponibles comercialmente, triclosán e isoniazida, *FabI* era el blanco favorito de los esfuerzos antimicrobianos dirigidos a síntesis de ácidos grasos [237] y fue propuesto como un blanco atractivo para combatir *K. pneumoniae* [165]. Otra proteína priorizada es *FolA* que proporciona la principal actividad de dihidrofolato reductasa en la ruta del tetrahidrofolato o vitamina B9. La enzima cataliza la reducción de dihidrofolato a tetrahidrofolato a través de la transferencia de hidruro de NADPH al C6 del anillo de pteridina. El tetrahidrofolato es un intermediario importante en la biosíntesis de proteínas y ácidos nucleicos, es biosintetizado *de novo* por bacterias y está involucrado en importantes vías de biosíntesis

como la metionina, las purinas, el timidilato y el pantotenato. Los seres humanos dependen por completo de las fuentes nutricionales de folato, transformándolo en una vitamina esencial. Debido a que el dihidrofolato reductasa es esencial para la división y el crecimiento celular, es un blanco atractivo para el desarrollo de fármacos. Los agentes antimicrobianos contra *FolA*, como 5-substituted-2, 4-diaminopyrimidine (TMP), se usan ampliamente para tratar infecciones por *Streptococcus pyogenes, S. pneumoniae, Escherichia coli* y *K. pneumoniae* [238]. El hallazgo de las proteínas *FabI* y *FolA* en nuestro trabajo contribuye a validar nuestro enfoque metodológico.

El producto proteico de *aroA*, 3-fosfoshikimato-1-carboxiviniltransferasa (EPSP sintasa), es otro blanco identificado en este trabajo. Esta proteína participa en la biosíntesis de aminoácidos aromáticos, sideróforos y metabolitos como el folato, la ubiquinona y la vitamina K y en el sexto paso de la vía del corismato [239] . Además, varios patógenos bacterianos utilizan sideróforos derivados del corismato como factores de virulencia. La inhibición de la EPSP sintasa es la base del ampliamente utilizado herbicida, el glifosato. También fue reportado en otros trabajos, que las EPSP sintasas de Clase I, que están presentes en plantas y algunas bacterias (p. ej., *Salmonella typhimurium* y *E. coli*), y se inhiben a bajas concentraciones de glifosato [240]. Varias propiedades del producto del gen *aroA* lo hacen atractivo como blanco antimicrobiano, incluyendo la esencialidad, la drogabilidad y la falta de un homólogo humano.

Nuestro análisis también reveló que el etilentetrahidrofolato-ARNt-(uracil-5-)-metiltransferasa (*TrmFO* o *GidA*) es un blanco prometedor para combatir la Bb. *TrmFO* es una proteína de unión a dinucleótido de flavina (FAD), que se ha identificado como una enzima modificadora de ARNt, responsable de la biosíntesis de m5U-54 dependiente de folato. En las bacterias, se ha informado que las enzimas de modificación del ARN exhiben una función de pleiotrópica asociada con muchos procesos celulares como la virulencia, la respuesta al estrés, la morfología, el crecimiento, la susceptibilidad a los antibióticos y otros [241] . Esta reportado previamente, que la deleción del gen deleción ha producido un crecimiento celular defectuoso, disminuyendo el crecimiento en aproximadamente un 10-30% en varias bacterias como *S. enterica*, *P. aeruginosa* y *E. coli* [242,243,244] Dado que la mutación *TrmFO* tiene un efecto pleiotrópico, que afecta diversos rasgos fenotípicos, incluido el crecimiento de células bacterianas, *TrmFO* podría considerarse un blanco nuevo y potencial en B. bacilliformis.

El producto proteico de *murE*, es una enzima bacteriana que pertenece a las ligasas *Mur* dependientes de ATP, que son esenciales para la producción de peptidoglicano (PG). PG

es un componente clave de la pared celular de casi todas las eubacterias, es responsable de la rigidez y la forma de las células bacterianas, sirve como plataforma para anclar otros componentes de la envoltura celular y es esencial para el crecimiento y la supervivencia de las bacterias. Las enzimas que catalizan PG se consideran una de las mejores fuentes de blancos antibacterianos. Al igual que las otras enzimas de la clase Mur, *MurE* ha sido objeto de esfuerzos para producir nuevos antimicrobianos. Se están estudiando diferentes compuestos químicos teniendo en cuenta sus estudios SAR (relación estructura-actividad) y su potencial inhibitorio. Tal es el caso de la hiperenona A que se obtiene a partir de extractos de las diferentes partes de la planta *Hypericum acmosepalum* y derivados de quinolonas como la novedosa N-metil-2-alquenil-4-quinolona, son capaces de inhibir la *MurE* ligasa de *M. tuberculosis* y *Staphylococcus aureus* metilino resistentes (MRSA) [245]

Tabla 14: Blancos obtenidos para Bb. Pi: punto isoelectrico; Mw: peso molecular. *** Término de GO más específico de la rama Función Molecular *** Término de GO más específico de la rama Proceso Biológico

| Gen | Descripción | Pi* | *Mw*** | Localización celular | Función Molecular *** | Proceso Biologico **** |
|---|---|---|---|---|---|---|
| *fabI* | *enoyl-(acyl carrier-protein) reductase* | 5.78 | 29 368 | Membrana citoplasmática | Actividad catalítica | Síntesis de Ácidos Grasos |
| *folA* | *dihydrofolate reductase* | 6.59 | 18 799 | Citoplasma | Unión a NADP | Proceso metabólico de un carbono<br>biosíntesis de nucleótidos<br>biosíntesis de amidas |
| *aroA* | *3-phosphoshikimate carboxyvinyltransferase* | 8.50 | 47 559 | Citoplasma | Actividad de Transferasa | Forma parte de los procesos de la familia de los aminoácidos aromáticos, del proceso biosintético del corismato. |
| *trmFO* | *FADH(2)-oxidising methylenetetrahydrofolate--tRNA-(uracil(54)-C(5))-methyltransferase* | 7.21 | 51 693 | Citoplasma | Actividad de metiltransferasa de unión a dinucleótido de flavina y adenina | Procesamiento de tRNA |
| *uppP* | *undecaprenyl-diphosphatase* | 9.2 | 22 650 | Membrana citoplasmática | Actividad catalítica | Componente integral de Membrana |
| *murE* | *UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2, 6-diaminopimelate ligase* | 6.51 | 52 537 | Citoplasma | Unión a ligando | Ciclo celular, regulación de la forma celular, proceso biosintético de peptidoglicano, división celular y organización de la pared celular |

## Aplicación en Listeria

*Listeria monocytogenes* (Lm) es un bacilo grampositivo corto, móvil, no esporulado y responsable de la listeriosis en humanos. Las infecciones por Lm se han convertido en los últimos años en una de las principales enfermedades transmitidas por los alimentos. Se asocian principalmente al consumo de alimentos contaminados como carne, pescado y verduras y los denominados productos listos para el consumo [246].

La terapia de elección para tratar la Listeriosis se basa en $\beta$-lactámicos, penicilina G o ampicilina combinados con o sin un aminoglucósido clásico como la gentamicina. Se considera un tratamiento alternativo a base de trimetoprima y sulfametoxazol para pacientes alérgicos a la penicilina. Además, la tetraciclina, la eritromicina y la vancomicina también se han utilizado para combatir la listeriosis humana [247,248]. En las últimas décadas, Lm ha desarrollado resistencias a una amplia gama de agentes antimicrobianos, incluso a los utilizados en tratamientos de referencia. La primera cepa Lm multirresistente se descubrió en Francia en 1988 [249]. Desde entonces, se han detectado cepas multirresistentes en aislados clínicos y plantas de producción de procesamiento de alimentos [250,251].

En el trabajo Palumbo et al. 2022 [252], del cual el tesista fue coautor, se hizo una priorización de proteínas blanco, utilizando como organismo modelo *L. monocytogenes* cepa *Lm-EGD*.

La información estructural, metabólica y de *off-target* obtenida en el *pipeline* de TP, se enriqueció con información de expresión durante la etapa de infección de la bacteria. La misma se obtuvo de trabajos previos que analizaron de microarrays utilizados para la expresión génica de condiciones fisiológicas relevantes que imitan el entorno patógeno durante la infección: replicación intracelular en macrófagos [253] y lumen intestinal y sangre [254] (436 en intestino, 541 en sangre y 809 en macrófagos).

Utilizando Target-Pathogen, se filtraron proteínas no esenciales, no drogables (DS<0,5) y con homólogos humanos cercanos. Posteriormente, se asignó una puntuación a cada proteína siguiendo la SF de la Ecuación 4.

$$SF = \frac{CP+C+Cv-GM}{4} + \frac{Intestine+Blood+Intracellular}{3}$$

Ecuación 4**:** CP y C reflejan información metabólica. Si la proteína está asociada con una reacción de cuello de botella, CP toma el valor de 1 (de lo contrario, CP = 0). C muestra la relación entre la centralidad de intermediación del nodo de la reacción asociada y el nodo con la centralidad más alta dentro de la red. Finalmente, $C_v$ es la proporción de aciertos de la proteína en Lm patógena y GM es la proporción de organismos del microbioma intestinal que tienen al menos una proteína homóloga con Lm. Luego Intestino + Sangre + Intracelular toma el valor de 1 si la proteína se sobreexpresa en esas condiciones, respectivamente.

Se sabe que al ingerir alimentos contaminados con Lm, el patógeno puede proliferar en el intestino y, luego de atravesar la barrera intestinal, diseminarse a la sangre para llegar al cerebro, bazo, hígado o incluso cruzar la barrera sangre-placentaria. En nuestro análisis, los datos de expresión [254,253] se tuvieron en cuenta para priorizar aquellos genes que se regularon positivamente durante las condiciones que imitan la infección. Las condiciones seleccionadas, que agrupan diferentes informes, comprenden supervivencia y replicación en ambientes intestinal, sanguíneo y macrófago.

Como se muestra en la Tabla 15, la proteína mejor clasificada, la trans-aldolasa *Tal2*, participa en la rama no oxidativa de la vía de las pentosas fosfato. Esta proteína se regula transcripcionalmente en el intestino, la sangre y los contextos de macrófagos y cumple con los requisitos de un blanco atractivo. Esta proteína es esencial, drogable (DS = 0,92), metabólicamente relevante y ampliamente conservada dentro de los patógenos Lm. Otro gen regulado al alza implicado en esta vía es *lmo2661*, que codifica una ribulosa-5-fosfato 3-epimerasa. Aunque no pudimos predecir la epimerasa como un gen esencial, nuestros resultados revelaron su asociación con una reacción chokepoint. Además, *lmo2661* es altamente susceptible a fármacos (DS: 0,82), se conserva en todas las cepas analizadas y no presenta homólogos cercanos en humanos. La importancia de esta vía radica en la producción de xilulosa y ribosa-5-fosfato como precursores de la biosíntesis de nucleótidos durante el crecimiento intracelular como se demostró en otro modelo de infección por macrófagos. [255].

La segunda proteína mejor clasificada está implicada en la vía de utilización de la ramnosa: proteína aldolasa ramnulosa-1-fosfato *RhaD*. *RhaD* es un excelente candidato para estudios posteriores, ya que encontramos que es esencial y altamente drogable, se conserva en todas las cepas patógenas analizadas y no presenta homólogos cercanos en humanos y su microbiota. Particularmente, aparece sobreexpresado bajo diferentes contextos que imitan la infección. Aunque solo *RhaD* es un blanco esencial dentro de la vía de utilización de la

ramnosa, otras proteínas como la ramnulo quinasa (*RhaB*) y la *rhamnose mutarotase* (*RhaM*), también albergan muchas características que las convierten en blancos atractivos. Curiosamente, los datos experimentales han demostrado que los mutantes *ΔRhaB* de *Listeria monocytogenes* crecen de manera menos eficiente en las células macrófagas[253].

Tabla 15: Blancos de de listeria de acuerdo con la SF de la Ecuación 4

| Orden | Locus tag | Gen | Producto | Pathways | DS | C | CP | P(%) | GM(%) | Intestino | Sangre | Intra-Celular |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | *lmo0343* | *tal2* | *transaldolase* | Pentosas fosfato | 0,92 | 0,65 | Si | 100 | 22 | Si | Si | Si |
| 2 | *lmo2847* | *rhaD* | *rhamnulose-1-phosphate aldolase* | Degradación de L-rhamnose | 0,89 | 0,09 | Si | 100 | 6 | Si | Si | Si |
| 3 | *lmo0794* | | *hypothetical protein* | - | 0,77 | | | 100 | 2 | Si | Si | Si |
| 4 | *lmo2335* | *fruA* | *PTS fructose transporter subunit IIABC* | - | 0.77 | | | 100 | 7 | Si | Si | Si |
| 5 | *lmo1838* | *pyrB* | *aspartate carbamoyltransferase* | Biosíntesis de UMP | 0,87 | 0,16 | Si | 100 | 13 | No | Si | Si |
| 6 | *lmo1293* | *glpD* | *glycerol-3-phosphate dehydrogenase* | Degradación de glicerol y glicofosfodiester | 0,52 | | Si | 100 | 3 | Si | No | Si |
| 7 | *lmo0317* | | *phosphomethylpyrimidine kinase* | - | 0,87 | | Si | 100 | 6 | Si | Si | No |
| 8 | *lmo2681* | *kdpB* | *potassium-transporting ATPase subunit B* | - | 0,88 | | | 100 | 40 | Si | Si | Si |
| 9 | *lmo2564* | | *4-oxalocrotonate isomerase* | - | 0,58 | | Si | 100 | 7 | Si | Si | No |
| 10 | *lmo2824* | | *D-3-phosphoglycerate dehydrogenase* | Biosíntesis de Serina | 0,93 | 0,01 | Si | 100 | 11 | No | Si | Si |

DS: Score de drogabilidad en el pocket más drogable; C: centralidad; CP: *chokepoint*; P(%): % de cepas patogénicas con un homólogo; GM(%): % de organismos de la microbiota intestinal con homólogo

# Conclusiones

- La búsqueda de blancos es una etapa crítica para el desarrollo de nuevas drogas, que puede ser potenciada integrando información de anotación, estructural, metabólica y de expresión.
- Se desarrolló la herramienta y base de datos Target-Pathogen, la cual a través de filtros y funciones de puntuación, permite obtener potenciales blancos de interés farmacológico, tanto novedosos como ya conocidos, donde lo primero es el foco del trabajo y lo segundo aporta a la validación de la metodología.
- Utilizando Target-Pathogen se pudo obtener una priorización de blancos para atacar *Mycobacterium tuberculosis* en fase latente, haciendo uso de la actividad micobactericida dependiente de la concentración de especies ERON. Para ello se integró información de múltiples fuentes: estructura, drogabilidad, información de expresión simulando condiciones de infección y de vías metabólicas. En el resultado de esto, por un lado se pudo validar el enfoque metodológico, ya que en las primeras posiciones del ranking se obtuvieron varias proteínas actualmente estudiadas como blancos, como ser *mshB / Rv1170* y *InhA / Rv1484*, pero también se propusieron nuevos, donde se destaca *ino1 / Rv0046c*, por afectar una vía clave para la supervivencia del patógeno, ser esencial, drogable y sobreexpresar en 3 de las 4 condiciones de infección en fase latente analizadas.
- La función de puntuación es clave para combinar el conocimiento de un experto en el patógeno con la información importante para la determinación de un buen blanco, lo cual permitió publicar trabajos tanto de manera local como en conjunto con otros equipos de trabajo de otros países, por ejemplo:
    - *Klebsiella pneumoniae* - Kp13: se destacaron *lpxC* y *lpxD*, donde ambas son necesarias para la biosíntesis de lipopolisacáridos. Si bien el primero tiene una droga que lo tiene como blanco, la misma no ha avanzado en etapas clínicas. Consideramos el caso de Kp13 particularmente interesante, ya que el proyecto plantea la búsqueda de un blanco para una terapia complementaria, intentando lograr una sinergia entre drogas para atacar los casos más graves de la enfermedad.
    - *Bartonella bacilliformis*: entre los blancos interesantes obtenidos en el análisis, tenemos a *aroA* participa de la vía del shikimato, no presente en humanos y que

sumado a su esencialidad y alta drogabilidad, parece estar relacionada con la virulencia en algunas bacterias. En la lista también se destaca *TrmFO*, relacionada con la virulencia, crecimiento de la bacteria y ya otros estudios fue reconocido como un potencial blanco terapéutico.

- *Listeria monocytogenes* - Lm-EGD: los blancos mejor ranqueados fueron *Tal2* y *RhaD*, donde ambas se sobreexpresan en condiciones de infección y poseen alta drogabilidad. En particular, se observó que en *RhaD*, está implicada en vías que utilizan rhamnose, componente el cual se incorpora a la pared celular WTA de la bacteria, promoviendo su resistencia a antibióticos.

# Conclusiones generales

La resistencia a fármacos antibacterianos es un problema creciente en los sistemas de salud de todo el mundo. En el caso de la Tuberculosis, la Argentina tiene una carga media en el sistema de salud y la circulación de cepas Mtb MDR y XDR representan una porción importante de ese problema. En este contexto, tanto local como global, es importante generar nuevos blancos para el desarrollo de nuevas drogas capaces de combatir la enfermedad.

En el primer capítulo de esta tesis se estudiaron los mecanismos moleculares de la resistencia en los aislamientos XDR secuenciados en Argentina entre los años 2003-2011. Primero se analizaron las variantes que tienen un fenotipo de resistencia asociado a una o más drogas. Para RIF, INH, aminoglucósidos y ETH, se encontraron variantes ya reportadas localmente y con preponderancia de pocas variantes, *katG* Ser315Thr / fabG1-inhA para INH, *rpoB* Ser450Trp y Asp435Val para RIF, *embB* Met306Ile y Gly406Asp para EMB, *rrs* A1401G para aminoglucósidos y *gyrA* Asp94Ala para las FQLs. Por otro lado, si bien no se encontraron 3 o más mutaciones para un mismo antibiótico, en un ~50% de las muestras se encontraron 2 para el mismo antibiótico, más de lo esperado teniendo en cuenta el impacto negativo de cada una de ellas en el *fitness* de las cepas. Para la gran mayoría de los casos, fue posible encontrar las distintas mutaciones que justifican un fenotipo XDR. En el caso particular de 2 aislamientos, no se pudo encontrar mutaciones asociadas a resistencia para INH. Tampoco se pudieron encontrar variantes de resistencia para aminoglucósidos y para fluoroquinolonas en 4 y 6 aislamientos respectivamente. En todos estos casos, repartidos en varias ramas del árbol filogenético, no fue posible dilucidar el mecanismo de resistencia.

Seguidamente se realizó un análisis filogenético de las muestras XDR en el contexto de otros brotes de gran relevancia en nuestro país (cepas M y Ra) y de referencia de todos los linajes conocidos a nivel global. A partir del análisis del árbol, se pudo inferir que los aislamientos XDR locales podían agruparse mayoritariamente en 5 grupos monofiléticos de interés: el grupo LAM3, H, LAM5, T1 y T2. Los primeros 2, están asociados con el brote de la cepas MDR Ra y M respectivamente, donde las muestras de este trabajo consisten en sus versiones XDR. En particular, dentro del grupo H, todas las variantes de resistencia están conservadas, excepto en FQLs. Algo para remarcar, es que dentro de cada grupo, generalmente se tienen las mismas variantes de resistencia. En INH, cada grupo tiene la variante *katG* Ser315Thr o *fabG1* C-15T respectivamente, excepto el grupo LAM3, cuyos

aislamientos portan ambos tipos de variantes. En el caso particular de *fabG1* C-15T, esa misma mutación explica la resistencia de los aislamientos al antibiótico ETH. Para RIF, hubo 2 mutaciones preponderantes, ambas en *rpoB*, Ser450Trp para el grupo LAM3, T1, T2 y H, y Asp435Val para el grupo LAM5. Para aminoglucósidos, la principal mutación fue *rrs* A1401G, portándola los grupos LAM5, T2, la mayor parte del T1 y el H. Contrariamente, dentro del grupo de los aislamientos LAM3 el mecanismo de resistencia parece estar provocado por mutaciones en el promotor del gen *eis* para KAN y rpsL para STR. Los grupos LAM5 y H, son los únicos que portan la misma variante de resistencia para PZA. Las FQLs solo siguen esta lógica en el grupo LAM5 y T2 con *gyrA* Asp94Ala, en el resto de los casos, no parece haber relación entre variantes de resistencia y los distintos grupos filogenéticos.

En el segundo capítulo presentamos Target-Pathogen, una base de datos que permite la consulta y priorización de targets de fármacos en patógenos. Es una aplicación web fácil de usar que permite la consolidación de datos basados en todo el genoma de diversas fuentes de una manera fácil de usar, con una interfaz gráfica para visualización y manipulación estructural. Hay pocas bases de datos existentes que estén diseñadas para la selección de blancos en un conjunto de patógenos relevantes, pero la mayoría de estas bases de datos se enfocan en un análisis de una sola proteína o se enfocan en características de proteínas específicas. Por ejemplo, Drug Target Database [256] es un recurso útil para seleccionar targets potenciales utilizando docking inverso. La base de datos de blancos terapéuticos [256,257] proporciona un gran volumen de datos de blancos terapéuticos previamente conocidos. Otra base de datos que incluye datos de blancos bien conocidos es TargetDB [257], pero solo se centra en la información estructural. TDR Targets [258] está en su versión 6 y es un recurso centrado en las enfermedades tropicales desatendidas. Tiene como plus interesante, incorporada la funcionalidad de buscar blancos por droga, ya que utilizando un esquema de capas / grafo multipartito, traslada la información de afinidad de drogas calculada en distintos ensayos, a homólogos o proteínas que comparten dominios, facilitando el reposicionamiento de compuestos.

Con información de estructura y drogabilidad, Target-Pathogen es un recurso valioso para analizar genomas completos de patógenos relevantes. Utilizándolo, los investigadores pueden priorizar rápidamente los genes de interés de una manera rápida e intuitiva, ejecutando consultas simples (como buscar proteínas con un puntaje alto de drogabilidad o asociadas con reacciones alta centralidad), filtrando por diferentes datos, asignando pesos numéricos para diferentes propiedades y combinando estos resultados para producir una lista de blancos

rankeados según su interès para el desarrollo de fármacos. Una vez que se cargan los datos, se pueden incluir para obtener una clasificación personalizada de posibles blancos de fármacos en patógenos. Otra característica clave en la que Target-Pathogen se destaca es su capacidad para clasificar no solo las proteínas sino también vías metabólicas completas. Esta característica permite priorizar rutas prometedoras para desarrollar nuevos fármacos con el fin de atacar sinérgicamente a varias proteínas de la misma vía lo que disminuye las posibilidades de generación de nuevas resistencias. Actualmente la base cuenta con 25 genomas de alta relevancia clínica, pero la base de datos se puede actualizar fácilmente con otros patógenos de interés a medida que se han automatizado los procesos bioinformáticos.

Utilizando Target-Pathogen se realizó una priorización y análisis de blancos moleculares para atacar *Mycobacterium tuberculosis* en fase latente, basado en la hipótesis de que si conocemos aquellas proteìnas que son blanco de las especies reactivas del nitrògeno y el oxìgeno a las que se ve expuesta el bacilo dentro del macrófago durante la fase latente, podemos desarrollar una droga capaz de hacer sinergia con la respuesta inmunitaria del hospedador para combatir a *Mtb* màs eficientemente. Para cumplir con este objetivo se integraron los datos de expresiòn provenientes de experimentos que simulan la infección con datos de drogabilidad estructural de cada proteìna codificada por el genoma de *Mtb* y datos provenientes de la reconstrucciòn metabòlica que nos permitiò determinar la relevancia de cada blanco en el contexto metabólico. Utilizando Target-Pathogen se pudo revalidar blancos propuestos con anterioridad y, fundamentalmente proponer blancos novedosos entre los que se destaca *ino1*, por su rol en la síntesis de micotiol, particularmente expresado en condiciones de infección.

Por último se utilizó Target-Pathogen para determinar blancos de relevancia en otros patògenos de interés clìnico en Argentina y la región como *Klebsiella pneumoniae*, *Bartonella bacilliformis* y *Listeria monocytogenes*. En todos los casos, los resultados obtenidos fueron de utilidad tanto para validar la metodología, al encontrar proteínas ya estudiados en bibliografía o ya blancos de compuestos existentes, como para predecir blancos novedosos, en el sentido que cumplian con los requerimientos que justifiquen mayor análisis y planificación de ensayos experimentales. Para *K. pneumoniae* se destacaron *lpxC* y *lpxD*, ambas proteínas pertenecientes a la vía de biosentìsis de polisacáridos, por lo que ademàs de sus caracterìsticas atractivas como blanco terapèutico (altamente drogables, centrales en el contexto metabòlico, altamente conservadas y sin homólogos cercanos en el genoma humano y su microbiota intestinal), pueden ser utilizadas para el diseño de un tratamiento sinèrgico y

complementario a Pb. Para el caso de *B. bacilliformis y L. monocytogenes* se propusieron como blancos de interés, entre otros, *aroA* y *TrmFO y RhaD* respectivamente relevantes en la síntesis de la pared celular. La estructura de todos los blancos seleccionados presentan una cavidad con alto score de drogabilidad.

El objetivo de Target-Pathogen no es reemplazar las estrategias de laboratorio para la identificación de blanco, sino colaborar a abrir nuevas líneas de investigación sobre los mismos y facilitar la planificación de ensayos de mesada, basada en la mejor información disponible. Su meta es convertirse en un recurso de referencia para los investigadores que trabajan en el campo de la identificación de blancos y/o el descubrimiento de fármacos, que desean traducir preguntas biológicas de una manera computacionalmente manejable al explorar, filtrar y ponderar la gran cantidad de conjuntos de datos a escala genómica disponibles en la actualidad. Target-Pathogen se actualizará continuamente como parte de una iniciativa nacional para desarrollar herramientas para patógenos.

## Papers Relacionados con la Tesis

Serral, Castello, **Sosa**. et al. From Genome to Drugs: New Approaches in Antimicrobial Discovery. Front. Pharmacol. 0, (2021).

**Sosa**, E. J. et al. Target-Pathogen: a structural bioinformatic approach to prioritize drug targets in pathogens. Nucleic Acids Res. 46, D413–D418 (2018).

Farfán-López M et al. Prioritization of potential drug targets against Bartonella bacilliformis by an integrative in-silico approach. Liebertpub.com https://www.liebertpub.com/doi/abs/.

Rondón, L. et al. Fluoromycobacteriophages Can Detect Viable Mycobacterium tuberculosis and Determine Phenotypic Rifampicin Resistance in 3–5 Days From Sputum Collection. Front. Microbiol. 0, (2018).

Bigi MM et al. Single nucleotide polymorphisms may explain the contrasting phenotypes of two variants of a multidrug-resistant Mycobacterium tuberculosis strain. Kekkaku 103, 28–36 (2017).

Defelipe et al. A whole genome bioinformatic approach to determine potential latent phase specific targets in Mycobacterium tuberculosis. Tuberculosis,Volume 97,2016,Pages 81-192

## Papers No Relacionados con la Tesis

Carolina, T. et al. Surveillance of SARS-CoV-2 variants in Argentina: detection of Alpha, Gamma, Lambda, Epsilon and Zeta in locally transmitted and imported cases. medRxiv 2021.07.19.21260779 (2021).

Barcudi D. Et Al. MRSA dynamic circulation between the community and the hospital setting: New insights from a cohort study. J. Infect. 80, 24–37 (2020).

# Anexo 1: aislamientos del proyecto

| | Cepa | ***Espoligotipo exp*** | ***Espoligotipo in silico*** | **Linaje in silico** | |
|---|---|---|---|---|---|
| 1 | 1074 | - | LAM9 | 4.1.2.1 | |
| 2 | 1940 | - | LAM5 | 4.3.3 | |
| 3 | 2003 | - | Unknown | 4.3.3 | |
| 4 | 10010 | T1 Ghana | T1 | 4.8 | |
| 5 | 10900 | H3 | T1 | 4.1.2.1 | |
| 6 | 11164 | H2 | H2 | 4.1.2.1.1 | |
| 7 | 11401 | LAM3 | LAM3 | 4.3.2 | |
| 8 | 11880 | LAM5 | LAM3 | 4.3.2 | |
| 9 | 12223 | H2 | - | - | Descartada |
| 10 | 12694 | H2 | H2 | 4.1.2.1.1 | |
| 11 | 12715 | H3 | H2 | 4.1.2.1.1 | |
| 12 | 12821 | - | T1 | 4.8;4.3.2 | |
| 13 | 12876 | H2 | - | 4.1.2.1.1 | |
| 14 | 13196 | H3 | T | - | Descartada |
| 15 | 14579 | T3-T2 ? | T1 | 4 | |
| 16 | 14698 | T1 Ghana | Beijing | 4.9 | |
| 17 | 15213 | Beijing | - | 2.2.1 | |
| 18 | 16561 | T5 Madrid | T1 | 4.7 | |
| 19 | 16785 | LAM3 | LAM5 | - | Descartada |
| 20 | 17270 | - | - | 4.1.2.1 | |
| 21 | 17488 | LAM5 | H2 | 4.3.3 | |
| 22 | 17591 | - | H2 | 4.8 | |
| 23 | 17817 | T1 Ghana | T-tuscany | 4.8 | |
| 24 | 18984 | - | T | - | Descartada |
| 25 | 19623 | H2 | T1 | 4.1.2.1.1 | |
| 26 | 19905 | H2 | H2 | 4.1.2.1.1 | |
| 27 | 20246 | Tuscany | T1 | 4.3.3 | |
| 28 | 20394 | T3-T2 ? | T-tuscany | 4.1.2 | |
| 29 | 20483 | - | T2 | 4.8 | |
| 30 | 20753 | H2 | - | 4.1.2.1.1 | |

| | | | | | |
|---|---|---|---|---|---|
| 31 | 20811 | LAM5 | X1 | 4.8;4.3.3 | |
| 32 | 22372 | T1 Tuscany | T3 | 4.3.3 | |
| 33 | 22468 | T2 | - | 4.8 | |
| 34 | 24300 | LAM9 | LAM3 | - | Descartada |
| 35 | 24821 | - | H2 | 4.3.4.2 | |
| 36 | 24830 | X1 | H2 | 4.1.1 | |
| 37 | 25024 | - | LAM5 | 4 | |
| 38 | 25144 | T3 | LAM5 | - | Descartada |
| 39 | 25203 | - | LAM5 | 4.3.2 | |
| 40 | 156324 | - | LAM5 | 4.1.2.1.1 | |
| 41 | 161216138 | H2 | LAM5 | 4.1.2.1.1 | |
| 42 | 13429 | LAM5 | T-tuscany | 4.3.3 | |
| 43 | 13431 | LAM5 | LAM5 | 4.3.3 | |
| 44 | 15058 | LAM5 | - | 4.3.3 | |
| 45 | 15059 | LAM5 | | 4.3.3 | |
| 46 | 15271 | LAM5 | | 4.3.3 | |
| 47 | 16306 | T1 Tuscany | | 4.3.3 | |
| 48 | 18712 | LAM5 | | 4.3.3 | |
| 49 | 23100 | T1 Ghana | | - | |

# Bibliografia


1. *Global tuberculosis report 2021*. (World Health Organization, 2021).
2. World Health Organization. *Global Tuberculosis Report 2015*. (World Health Organization, 2015).
3. Eldholm, V. *et al.* Evolution of extensively drug-resistant Mycobacterium tuberculosis from a susceptible ancestor in a single patient. *Genome Biol.* **15**, 1–11 (2014).
4. Programme, G. T. Global tuberculosis report 2021. https://www.who.int/publications/i/item/9789240037021 (2021).
5. Togun, T., Kampmann, B., Stoker, N. G. & Lipman, M. Anticipating the impact of the COVID-19 pandemic on TB patients and TB control programmes. *Ann. Clin. Microbiol. Antimicrob.* **19**, 21 (2020).
6. Mirzayev, F. *et al.* World Health Organization recommendations on the treatment of drug-resistant tuberculosis, 2020 update. *Eur. Respir. J.* **57**, (2021).
7. Harding, E. WHO global progress report on tuberculosis elimination. *Lancet Respir Med* **8**, 19 (2020).
8. Woodman, M., Haeusler, I. L. & Grandjean, L. Tuberculosis Genetic Epidemiology: A Latin American Perspective. *Genes* **10**, (2019).
9. Porvaznik, I., Solovič, I. & Mokrý, J. Non-Tuberculous Mycobacteria: Classification, Diagnostics, and Therapy. *Adv. Exp. Med. Biol.* **944**, 19–25 (2017).
10. Website. https://doi.org/10.1093/ajcp/30.3.267 doi:10.1093/ajcp/30.3.267.
11. Niemann, S. & Supply, P. Diversity and Evolution of Mycobacterium tuberculosis: Moving to Whole-Genome-Based Approaches. *Cold Spring Harb. Perspect. Med.* **4**, (2014).
12. Achtman, M. Evolution, population structure, and phylogeography of genetically

monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* **62**, 53–70 (2008).

13. Kato-Maeda, M. *et al.* Strain classification of Mycobacterium tuberculosis: congruence between large sequence polymorphisms and spoligotypes. *Int. J. Tuberc. Lung Dis.* **15**, 131–133 (2011).
14. Martens, E. & Demain, A. L. The antibiotic resistance crisis, with a focus on the United States. *The Journal of Antibiotics* vol. 70 520–526 (2017).
15. Clatworthy, A. E., Pierson, E. & Hung, D. T. Targeting virulence: a new paradigm for antimicrobial therapy. *Nat. Chem. Biol.* **3**, 541 (2007).
16. TÉLAM. Diputados aprobó proyecto de Prevención y Control de la Resistencia a Antimicrobianos. https://www.telam.com.ar/notas/202207/597623-resistencia-a-antimicrobianos-ley.html.
17. Nasiri, M. J. *et al.* Drug resistance pattern of Mycobacterium tuberculosis isolates from patients of five provinces of Iran. *Asian Pac. J. Trop. Med.* **7**, 193–196 (2014).
18. Andrews, J. R. *et al.* Exogenous reinfection as a cause of multidrug-resistant and extensively drug-resistant tuberculosis in rural South Africa. *J. Infect. Dis.* **198**, 1582–1589 (2008).
19. Gandhi, N. R. *et al.* Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *Lancet* **375**, 1830–1843 (2010).
20. WHO | Antimicrobial resistance: global report on surveillance 2014. (2016).
21. *Antibacterial Agents in Clinical Development: An Analysis of the Antibacterial Clinical Development Pipeline, Including Tuberculosis*. (2017).
22. Datos recientes revelan los altos niveles de resistencia a los antibióticos en todo el mundo. https://www.who.int/es/news/item/29-01-2018-high-levels-of-antibiotic-resistance-found-worldwide-new-data-shows.

23. World Health Organization. *Global Tuberculosis Report 2018*. (World Health Organization, 2018).
24. Zumla, A. *et al.* Drug-Resistant Tuberculosis—Current Dilemmas, Unanswered Questions, Challenges, and Priority Needs. *The Journal of Infectious Diseases* vol. 205 S228–S240 (2012).
25. Dheda, K. *et al.* Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis. *The Lancet Respiratory Medicine* **2**, 321–338 (2014).
26. Spellberg, B., Powers, J. H., Brass, E. P., Miller, L. G. & Edwards, J. E., Jr. Trends in antimicrobial drug development: implications for the future. *Clin. Infect. Dis.* **38**, 1279–1286 (2004).
27. Rosamond, J. & Allsop, A. Harnessing the power of the genome in the search for new antibiotics. *Science* **287**, 1973–1976 (2000).
28. Ventola, C. L. The antibiotic resistance crisis: part 1: causes and threats. *P T* **40**, 277–283 (2015).
29. Hutchinson, L. & Kirk, R. High drug attrition rates—where are we going wrong? *Nat. Rev. Clin. Oncol.* **8**, 189–190 (2011).
30. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 711–715 (2004).
31. Del Giacco, L. & Cattaneo, C. Introduction to genomics. *Methods Mol. Biol.* **823**, 79–88 (2012).
32. Setubal, J. C., Almeida, N. F. & Wattam, A. R. Comparative Genomics for Prokaryotes. *Methods Mol. Biol.* **1704**, 55–78 (2018).
33. Weissenbach, J. The rise of genomics. *C. R. Biol.* **339**, 231–239 (2016).
34. Burguener, G. F. Desarrollo de herramientas bioinformáticas para la anotación de

genomas. (Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales, 2016).

35. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
36. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
37. Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155**, 335–350 (1987).
38. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 1–11 (2016).
39. Rosenberg, M. S. *Sequence alignment: Methods, models, concepts, and strategies*. (University of California Press, 2009). doi:10.1525/9780520943742.
40. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
41. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
42. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science* vol. 27 135–145 (2018).
43. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
44. Katoh, K. & Toh, H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* vol. 26 1899–1900 (2010).
45. Radusky, L. G. Herramientas bioinformáticas para el análisis estructural de proteínas a escala genómica. (Universidad de Buenos Aires. Facultad de Ciencias Exactas y

Naturales, 2017).

46. Lanzarotti, E. Herramientas bioinformáticas para la predicción y análisis de estructuras de proteínas : de genomas a motivos estructurales. (Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales, 2016).
47. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
48. Consortium, T. G. O. & The Gene Ontology Consortium. Gene Ontology Annotations and Resources. *Nucleic Acids Research* vol. 41 D530–D535 (2012).
49. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Research* vol. 28 304–305 (2000).
50. Falquet, L. *et al.* The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30**, 235–238 (2002).
51. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
52. Sosa, E. J. *et al.* Target-Pathogen: a structural bioinformatic approach to prioritize drug targets in pathogens. *Nucleic Acids Res.* **46**, D413–D418 (2018).
53. Eldholm, V. *et al.* Impact of HIV co-infection on the evolution and transmission of multidrug-resistant tuberculosis. *Elife* **5**, (2016).
54. Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W. & Small, P. M. Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4871–4876 (2004).
55. Gagneux, S. *et al.* Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2869–2873 (2006).
56. Reed, M. B. *et al.* Major Mycobacterium tuberculosis lineages associate with patient country

of origin. *J. Clin. Microbiol.* **47**, 1119–1128 (2009).

57. Trauner, A., Borrell, S., Reither, K. & Gagneux, S. Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. *Drugs* **74**, 1063–1072 (2014).
58. Seifert, M., Catanzaro, D., Catanzaro, A. & Rodwell, T. C. Genetic mutations associated with isoniazid resistance in Mycobacterium tuberculosis: a systematic review. *PLoS One* **10**, e0119628 (2015).
59. Borrell, S. & Gagneux, S. Strain diversity, epistasis and the evolution of drug resistance in Mycobacterium tuberculosis. *Clin. Microbiol. Infect.* **17**, 815–820 (2011).
60. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
61. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).
62. Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* **11**, 41 (2019).
63. Salamon, H. *et al.* Integration of published information into a resistance-associated mutation database for Mycobacterium tuberculosis. *J. Infect. Dis.* **211 Suppl 2**, S50–7 (2015).
64. Nasiri, M. J. *et al.* New Insights in to the Intrinsic and Acquired Drug Resistance Mechanisms in Mycobacteria. *Front. Microbiol.* **8**, 681 (2017).
65. 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.* **73**, 1138–1151 (2018).

66. Martínez, N. L. & Santonja, J. T. *Paleontología: conceptos y métodos*. (Síntesis, 1994).

67. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).

68. Arenas, M. Advances in computer simulation of genome evolution: toward more realistic evolutionary genomics analysis by approximate bayesian computation. *J. Mol. Evol.* **80**, 189–192 (2015).

69. Sorek, Kunin & Hugenholtz. CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.*

70. He, L., Fan, X. & Xie, J. Comparative genomic structures of Mycobacterium CRISPR-Cas. *J. Cell. Biochem.* **113**, 2464–2473 (2012).

71. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907–914 (1997).

72. Dale, J. W. *et al.* Spacer oligonucleotide typing of bacteria of the Mycobacterium tuberculosis complex: recommendations for standardised nomenclature [The Language of Our Science]. *Int. J. Tuberc. Lung Dis.* **5**, 216–219 (2001).

73. Jagielski, T. *et al.* Methodological and Clinical Aspects of the Molecular Epidemiology of Mycobacterium tuberculosis and Other Mycobacteria. *Clin. Microbiol. Rev.* **29**, 239–290 (2016).

74. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5**, 4812 (2014).

75. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res.* **7**, 1338 (2018).

76. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

77. Li, X., Li, H. & Li, L. Probability forecasting of burst transmission for IEEE 802.16 BWA systems. in *2010 5th International Conference on Computer Science & Education* (2010). doi:10.1109/iccse.2010.5593642.

78. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

79. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

80. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

81. Flandrois, J.-P., Lina, G. & Dumitrescu, O. MUBII-TB-DB: a database of mutations associated with antibiotic resistance in Mycobacterium tuberculosis. *BMC Bioinformatics* **15**, 107 (2014).

82. Posada, D. jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* **25**, 1253–1256 (2008).

83. Sequencing Coverage. https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html.

84. Fernandez Do Porto, D. A. *et al.* Five-year microevolution of a multidrug-resistant Mycobacterium tuberculosis strain within a patient with inadequate compliance to treatment. *BMC Infect. Dis.* **21**, 394 (2021).

85. Carolina, T. *et al.* Surveillance of SARS-CoV-2 variants in Argentina: detection of Alpha, Gamma, Lambda, Epsilon and Zeta in locally transmitted and imported cases. *medRxiv* 2021.07.19.21260779 (2021).

86. Monteserin, J. *et al.* Relation of Mycobacterium tuberculosis mutations at katG315 and inhA-15 with drug resistance profile, genetic background, and clustering in Argentina. *Diagn. Microbiol. Infect. Dis.* **89**, 197–201 (2017).

87. Ezekiel, D. H. & Hutchins, J. E. Mutations affecting RNA polymerase associated with rifampicin resistance in Escherichia coli. *Nature* **220**, 276–277 (1968).

88. Imperiale, B. R., Di Giulio, Á. B., Mancino, M. B., Zumárraga, M. J. & Morcillo, N. S. Surveillance and characterization of drug-resistant *Mycobacterium tuberculosis* isolated in a reference hospital from Argentina during 8 years' period. *Int J Mycobacteriol* **8**, 223–228 (2019).

89. Jagielski, T. *et al.* Characterization of Mutations Conferring Resistance to Rifampin in Mycobacterium tuberculosis Clinical Strains. *Antimicrob. Agents Chemother.* **62**, (2018).

90. Huitric, E., Werngren, J., Juréen, P. & Hoffner, S. Resistance levels and rpoB gene mutations among in vitro-selected rifampin-resistant Mycobacterium tuberculosis mutants. *Antimicrob. Agents Chemother.* **50**, 2860–2862 (2006).

91. Monteserin, J. *et al.* Genotypic diversity of Mycobacterium tuberculosis in Buenos Aires, Argentina. *Infect. Genet. Evol.* **62**, 1–7 (2018).

92. Zaw, M. T., Emran, N. A. & Lin, Z. Mutations inside rifampicin-resistance determining region of rpoB gene associated with rifampicin-resistance in Mycobacterium tuberculosis. *J. Infect. Public Health* **11**, 605–610 (2018).

93. Mariam, D. H., Mengistu, Y., Hoffner, S. E. & Andersson, D. I. Effect of rpoB mutations conferring rifampin resistance on fitness of Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **48**, 1289–1294 (2004).

94. Eldholm, V. *et al.* Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nat. Commun.* **6**,.

95. Coll, F. *et al.* Author Correction: Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **50**, 764 (2018).
96. Sun, Q. *et al.* Mutations within Are Associated with Variable Level of Ethambutol Resistance in Mycobacterium tuberculosis Isolates from China. *Antimicrob. Agents Chemother.* **62**, (2018).
97. Sreevatsan, S. *et al.* Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations. *Antimicrob. Agents Chemother.* **41**, 1677–1681 (1997).
98. Almeida Da Silva, P. E. A. & Palomino, J. C. Molecular basis and mechanisms of drug resistance in Mycobacterium tuberculosis: classical and new drugs. *J. Antimicrob. Chemother.* **66**, 1417–1430 (2011).
99. Alangaden, G. J. *et al.* Mechanism of resistance to amikacin and kanamycin in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **42**, 1295–1297 (1998).
100. Wang, Y. *et al.* The roles of rpsL, rrs, and gidB mutations in predicting streptomycin-resistant drugs used on clinical Mycobacterium tuberculosis isolates from Hebei Province, China. *Int. J. Clin. Exp. Pathol.* **12**, 2713–2721 (2019).
101. Minnick, M. F., Wilson, Z. R., Smitherman, L. S. & Samuels, D. S. gyrA mutations in ciprofloxacin-resistant Bartonella bacilliformis strains obtained in vitro. *Antimicrob. Agents Chemother.* **47**, 383–386 (2003).
102. Giannoni, F. *et al.* Evaluation of a new line probe assay for rapid identification of gyrA mutations in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **49**, 2928–2933 (2005).
103. Imperiale, B. R., Zumárraga, M. J., Di Giulio, A. B., Cataldi, A. A. & Morcillo, N. S. Molecular and phenotypic characterisation of Mycobacterium tuberculosis resistant to anti-tuberculosis drugs. *Int. J. Tuberc. Lung Dis.* **17**, 1088–1093 (2013).

104. Maus, C. E., Plikaytis, B. B. & Shinnick, T. M. Molecular analysis of cross-resistance to capreomycin, kanamycin, amikacin, and viomycin in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **49**, 3192–3197 (2005).

105. Kambli, P. *et al.* Correlating rrs and eis promoter mutations in clinical isolates of Mycobacterium tuberculosis with phenotypic susceptibility levels to the second-line injectables. *International Journal of Mycobacteriology* vol. 5 1–6 (2016).

106. Prasad, M. S., Bhole, R. P., Khedekar, P. B. & Chikhale, R. V. Mycobacterium enoyl acyl carrier protein reductase (InhA): A key target for antitubercular drug discovery. *Bioorg. Chem.* **115**, 105242 (2021).

107. Ushtanit, A. *et al.* Molecular Determinants of Ethionamide Resistance in Clinical Isolates of Mycobacterium tuberculosis. *Antibiotics (Basel)* **11**, (2022).

108. Monteserin, J. *et al.* Trends of Two Epidemic Multidrug-Resistant Strains of Mycobacterium tuberculosis in Argentina Disclosed by Tailored Molecular Strategy. *Am. J. Trop. Med. Hyg.* **101**, 1308–1311 (2019).

109. Melnyk, A. H., Wong, A. & Kassen, R. The fitness costs of antibiotic resistance mutations. *Evolutionary Applications* vol. 8 273–283 (2015).

110. Gagneux, S. Fitness cost of drug resistance inMycobacterium tuberculosis. *Clinical Microbiology and Infection* vol. 15 66–68 (2009).

111. Wenzel, R. P. The antibiotic pipeline--challenges, costs, and values. *N. Engl. J. Med.* **351**, 523–526 (2004).

112. Fernandes, P. The global challenge of new classes of antibacterial agents: an industry perspective. *Curr. Opin. Pharmacol.* **24**, 7–11 (2015).

113. Chiba, H. & Uchiyama, I. Improvement of domain-level ortholog clustering by optimizing domain-specific sum-of-pairs score. *BMC Bioinformatics* **15**, 148 (2014).

114. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr Protoc* **1**, e323 (2021).
115. Hallgren, J. *et al.* DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv* 2022.04.08.487609 (2022) doi:10.1101/2022.04.08.487609.
116. Mariotti, M., Lobanov, A. V., Guigo, R. & Gladyshev, V. N. SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res.* **41**, (2013).
117. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).
118. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* vol. 29 2933–2935 (2013).
119. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
120. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
121. Rombel, I. T., Sykes, K. F., Rayner, S. & Johnston, S. A. ORF-FINDER: a vector for high-throughput gene identification. *Gene* **282**, 33–41 (2002).
122. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115 (1998).
123. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
124. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 1–11 (2010).
125. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in

KEGG. *Nucleic Acids Res.* **42**, D199–205 (2014).

126. Krieger, C. J. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32**, D438–42 (2004).

127. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–80 (2016).

128. Fabregat, A. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**, 142 (2017).

129. Karp, P. D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics* **18**, S225–S232 (2002).

130. Karp, P. D., Latendresse, M. & Caspi, R. The Pathway Tools Pathway Prediction Algorithm. *Stand. Genomic Sci.* **5**, 424–429 (2011).

131. A review of computational tools for design and reconstruction of metabolic pathways. *Synthetic and Systems Biotechnology* **2**, 243–252 (2017).

132. Lacroix, V., Cottret, L., Thébault, P. & Sagot, M.-F. An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **5**, 594–617 (2008).

133. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).

134. He, X. & Zhang, J. Why do hubs tend to be essential in protein networks? *PLoS Genet.* **2**, e88 (2006).

135. Zotenko, E., Mestre, J., O'Leary, D. P. & Przytycka, T. M. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.* **4**, e1000140 (2008).

136. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* vol. 450 964–972 (2007).

137. Sánchez, R. & Sali, A. Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.* **7**, 206–214 (1997).

138. Eisenhaber, F., Persson, B. & Argos, P. Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.* **30**, 1–94 (1995).

139. Laurents, D. V., Subbiah, S. & Levitt, M. Different protein sequences can give rise to highly similar folds through different stabilizing interactions. *Protein Sci.* **3**, (1994).

140. Dumas, V. G. Decodificando la diversidad catalítica de los citocromos p450 bacterianos mediante métodos bioinformáticos. (Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales, 2016).

141. Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* vol. 27 343–350 (2011).

142. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).

143. Schmidtke, P. & Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **53**, 5858–5867 (2010).

144. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).

145. Ligon, B. L. Penicillin: its discovery and early development. *Semin. Pediatr. Infect. Dis.* **15**, 52–57 (2004).

146. Kolter, R. & van Wezel, G. P. Goodbye to brute force in antibiotic discovery? *Nat Microbiol* **1**, 15020 (2016).

147. Reymond, J.-L., van Deursen, R., Blum, L. C. & Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* vol. 1 30 (2010).

148. Selzer, P. M., Brutsche, S., Wiesner, P., Schmid, P. & Müllner, H. Target-based drug discovery for the development of novel antiinfectives. *International Journal of Medical Microbiology* vol. 290 191–201 (2000).

149. Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **6**, 29–40 (2007).

150. Projan, S. J. Why is big Pharma getting out of antibacterial drug discovery? *Curr. Opin. Microbiol.* **6**, 427–430 (2003).

151. Hackbarth, C. J. *et al.* N-alkyl urea hydroxamic acids as a new class of peptide deformylase inhibitors with antibacterial activity. *Antimicrob. Agents Chemother.* **46**, 2752–2764 (2002).

152. Flores, A. & Quesada, E. Entry inhibitors directed towards glycoprotein gp120: an overview on a promising target for HIV-1 therapy. *Curr. Med. Chem.* **20**, 751–771 (2013).

153. Farha, M. A. *et al.* Inhibition of WTA synthesis blocks the cooperative action of PBPs and sensitizes MRSA to β-lactams. *ACS Chem. Biol.* **8**, 226–233 (2013).

154. Starkey, M. *et al.* Identification of Anti-virulence Compounds That Disrupt Quorum-Sensing Regulated Acute and Persistent Pathogenicity. *PLoS Pathogens* vol. 10 e1004321 (2014).

155. Storek, K. M. *et al.* Monoclonal antibody targeting the β-barrel assembly machine of Escherichia coli is bactericidal. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 3692–3697 (2018).

156. da Silva, F. A. B., Carels, N., dos Santos, M. T. & Lopes, F. J. P. *Networks in Systems Biology: Applications for Disease Modeling*. (Springer Nature, 2020).

157. Pundir, S., Magrane, M., Martin, M. J., O'Donovan, C. & The UniProt Consortium. Searching and Navigating UniProt Databases. *Current Protocols in Bioinformatics* vol. 50 (2015).

158. Hancock, J. M. & Bishop, M. J. HMMer. *Dictionary of Bioinformatics and Computational*

*Biology* (2004) doi:10.1002/9780471650126.dob0323.pub2.

159. Attwood, T. PFAM. *Dictionary of Bioinformatics and Computational Biology* (2004) doi:10.1002/0471650129.dob0526.

160. Furnham, N. *et al.* The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **42**, D485–9 (2014).

161. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery* vol. 1 727–730 (2002).

162. Cheng, A. C. *et al.* Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **25**, 71–75 (2007).

163. Xie, L. & Bourne, P. E. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics* **8 Suppl 4**, S9 (2007).

164. Schmidt, F., Matter, H., Hessler, G. & Czich, A. Predictive in silico off-target profiling in drug discovery. *Future Med. Chem.* **6**, 295–317 (2014).

165. Ramos, P. I. P. *et al.* An integrative, multi-omics approach towards the prioritization of Klebsiella pneumoniae drug targets. *Sci. Rep.* **8**, 10755 (2018).

166. Zhang, R. DEG: a database of essential genes. *Nucleic Acids Research* vol. 32 271D–272 (2004).

167. Luo, H., Lin, Y., Gao, F., Zhang, C.-T. & Zhang, R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements: Table 1. *Nucleic Acids Research* vol. 42 D574–D580 (2014).

168. Barh, D. *et al.* Conserved host-pathogen PPIs. Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis, and E. coli

targeted by Piper betel compounds. *Integr. Biol.* **5**, 495–509 (2013).

169. Barh, D. *et al.* A novel comparative genomics analysis for common drug and vaccine targets in Corynebacterium pseudotuberculosis and other CMN group of human pathogens. *Chem. Biol. Drug Des.* **78**, 73–84 (2011).

170. Radusky, L. *et al.* TuberQ: a Mycobacterium tuberculosis protein druggability database. *Database* **2014**, bau035 (2014).

171. Defelipe, L. A. *et al.* A whole genome bioinformatic approach to determine potential latent phase specific targets in Mycobacterium tuberculosis. *Tuberculosis* **97**, 181–192 (2016).

172. Albert, R. Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947–4957 (2005).

173. Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D. & Altman, R. B. Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome Res.* **14**, 917–924 (2004).

174. Karp, P. D. *et al.* Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics* vol. 17 877–890 (2016).

175. Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J. & Holmes, I. H. JBrowse: a next-generation genome browser. *Genome Res.* **19**, 1630–1638 (2009).

176. Consortium, G. O. & Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* vol. 32 258D–261 (2004).

177. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Krona: Interactive Metagenomic Visualization in a Web Browser. *Encyclopedia of Metagenomics* 339–346 (2015) doi:10.1007/978-1-4899-7478-5_802.

178. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–8, 27–8 (1996).

179. Guruprasad, K., Savitha, S. & Babu, A. V. N. Computational tools for the analysis of

heteroatom groups and their neighbours in protein tertiary structure. *Int. J. Biol. Macromol.* **37**, 35–41 (2005).

180. Tiwari, N., Gedda, M. R., Tiwari, V. K., Singh, S. P. & Singh, R. K. Limitations of Current Therapeutic Options, Possible Drug Targets and Scope of Natural Products in Control of Leishmaniasis. *Mini Rev. Med. Chem.* **18**, 26–41 (2018).

181. Chawla, B. & Madhubala, R. Drug targets in Leishmania. *Journal of Parasitic Diseases* vol. 34 1–13 (2010).

182. Naula, C., Parsons, M. & Mottram, J. C. Protein kinases as drug targets in trypanosomes and Leishmania. *Biochim. Biophys. Acta* **1754**, 151–159 (2005).

183. Mann, P. A. *et al.* Murgocil is a Highly Bioactive Staphylococcal-Specific Inhibitor of the Peptidoglycan Glycosyltransferase Enzyme MurG. *ACS Chem. Biol.* **8**, 2442–2451 (2013).

184. Rajendran, V., Kalita, P., Shukla, H., Kumar, A. & Tripathi, T. Aminoacyl-tRNA synthetases: Structure, function, and drug discovery. *Int. J. Biol. Macromol.* **111**, 400–414 (2018).

185. Wayne, L. G. In Vitro Model of Hypoxically Induced Nonreplicating Persistence of Mycobacterium tuberculosis. *Mycobacterium Tuberculosis Protocols* 247–269 doi:10.1385/1-59259-147-7:247.

186. Murphy, D. J. & Brown, J. R. Identification of gene targets against dormant phase Mycobacterium tuberculosis infections. *BMC Infect. Dis.* **7**, 84 (2007).

187. Koul, A., Arnoult, E., Lounis, N., Guillemont, J. & Andries, K. The challenge of new drug discovery for tuberculosis. *Nature* **469**, 483–490 (2011).

188. Wormser, G. P., Belknap, R. W. & Cohn, D. L. Tuberculosis, Second Edition Edited by William N. Rom and Stuart M. Garay Philadelphia: Lippincott Williams & Wilkins, 2004. 944 pp., illustrated. $159.95 (cloth). *Clin. Infect. Dis.* **38**, 1509–1510 (2004).

189. Matsumoto, M. *et al.* OPC-67683, a nitro-dihydro-imidazooxazole derivative with promising

action against tuberculosis in vitro and in mice. *PLoS Med.* **3**, e466 (2006).

190. Singh, R. *et al.* PA-824 kills nonreplicating Mycobacterium tuberculosis by intracellular NO release. *Science* **322**, 1392–1395 (2008).

191. Cellitti, S. E. *et al.* Structure of Ddn, the Deazaflavin-dependent nitroreductase from Mycobacterium tuberculosis involved in bioreductive activation of PA-824, with co-factor F420. (2012) doi:10.2210/pdb3r5r/pdb.

192. Manjunatha, U., Boshoff, H. I. & Barry, C. E. The mechanism of action of PA-824: Novel insights from transcriptional profiling. *Commun. Integr. Biol.* **2**, 215–218 (2009).

193. Rhee, K. Y., Erdjument-Bromage, H., Tempst, P. & Nathan, C. F. S-nitroso proteome of Mycobacterium tuberculosis: Enzymes of intermediary metabolism and antioxidant defense. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 467–472 (2005).

194. Voskuil, M. I., Bartek, I. L., Visconti, K. & Schoolnik, G. K. The response of mycobacterium tuberculosis to reactive oxygen and nitrogen species. *Front. Microbiol.* **2**, 105 (2011).

195. Voskuil, M. I., Visconti, K. C. & Schoolnik, G. K. Mycobacterium tuberculosis gene expression during adaptation to stationary phase and low-oxygen dormancy. *Tuberculosis* **84**, 218–227 (2004).

196. Voskuil, M. I. *et al.* Inhibition of respiration by nitric oxide induces a Mycobacterium tuberculosis dormancy program. *J. Exp. Med.* **198**, 705–713 (2003).

197. Muttucumaru, D. G. N., Roberts, G., Hinds, J., Stabler, R. A. & Parish, T. Gene expression profile of Mycobacterium tuberculosis in a non-replicating state. *Tuberculosis* **84**, 239–246 (2004).

198. Park, H.-D. *et al.* Rv3133c/dosR is a transcription factor that mediates the hypoxic response of Mycobacterium tuberculosis. *Molecular Microbiology* vol. 48 833–843 (2003).

199. Sherman, D. R. *et al.* Regulation of the Mycobacterium tuberculosis hypoxic response gene

encoding alpha -crystallin. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 7534–7539 (2001).

200. Betts, J. C., Lukey, P. T., Robb, L. C., McAdam, R. A. & Duncan, K. Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. *Mol. Microbiol.* **43**, 717–731 (2002).

201. Hampshire, T. *et al.* Stationary phase gene expression of Mycobacterium tuberculosis following a progressive nutrient depletion: a model for persistent organisms? *Tuberculosis* **84**, 228–238 (2004).

202. Karakousis, P. C. *et al.* Dormancy phenotype displayed by extracellular Mycobacterium tuberculosis within artificial granulomas in mice. *J. Exp. Med.* **200**, 647–657 (2004).

203. Ohno, H. *et al.* The effects of reactive nitrogen intermediates on gene expression in Mycobacterium tuberculosis. *Cell. Microbiol.* **5**, 637–648 (2003).

204. Schnappinger, D. *et al.* Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages: Insights into the Phagosomal Environment. *J. Exp. Med.* **198**, 693–704 (2003).

205. Talaat, A. M., Lyons, R., Howard, S. T. & Johnston, S. A. The temporal expression profile of Mycobacterium tuberculosis infection in mice. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4602–4607 (2004).

206. Raman, K., Yeturu, K. & Chandra, N. targetTB: A target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Systems Biology* vol. 2 (2008).

207. Agüero, F. *et al.* Genomic-scale prioritization of drug targets: the TDR Targets database. *Nat. Rev. Drug Discov.* **7**, 900–907 (2008).

208. Anand, P. & Chandra, N. Characterizing the pocketome of Mycobacterium tuberculosis and application in rationalizing polypharmacological target selection. *Sci. Rep.* **4**, 6356 (2014).

209. Sassetti, C. M., Boyd, D. H. & Rubin, E. J. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48**, 77–84 (2003).

210. Griffin, J. E. *et al.* High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog.* **7**, e1002251 (2011).

211. Movahedzadeh, F. *et al.* The Mycobacterium tuberculosis ino1 gene is essential for growth and virulence. *Mol. Microbiol.* **51**, 1003–1014 (2004).

212. Cordillot, M. *et al.* In vitro cross-linking of Mycobacterium tuberculosis peptidoglycan by L,D-transpeptidases and inactivation of these enzymes by carbapenems. *Antimicrob. Agents Chemother.* **57**, 5940–5945 (2013).

213. Feng, Z. & Barletta, R. G. Roles of Mycobacterium smegmatis D-alanine:D-alanine ligase and D-alanine racemase in the mechanisms of action of and resistance to the peptidoglycan inhibitor D-cycloserine. *Antimicrob. Agents Chemother.* **47**, 283–291 (2003).

214. Belanger, A. E. *et al.* The embAB genes of Mycobacterium avium encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 11919–11924 (1996).

215. Ramos, P. I. P. *et al.* Pyrosequencing-based analysis reveals a novel capsular gene cluster in a KPC-producing Klebsiella pneumoniae clinical isolate identified in Brazil. *BMC Microbiol.* **12**, 173 (2012).

216. Ramos, P. I. P. *et al.* Comparative analysis of the complete genome of KPC-2-producing Klebsiella pneumoniae Kp13 reveals remarkable genome plasticity and a wide repertoire of virulence and resistance mechanisms. *BMC Genomics* **15**, 54 (2014).

217. Bush, K., Jacoby, G. A. & Medeiros, A. A. A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrob. Agents Chemother.* **39**, 1211–1233 (1995).

218. Campos, A. C. *et al.* Outbreak of Klebsiella pneumoniae carbapenemase-producing K pneumoniae: A systematic review. *Am. J. Infect. Control* **44**, 1374–1380 (2016).

219. Liao, Y.-C. *et al.* An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. *J. Bacteriol.* **193**, 1710–1717 (2011).

220. Polyak, S. W., Abell, A. D., Wilce, M. C. J., Zhang, L. & Booker, G. W. Structure, function and selective inhibition of bacterial acetyl-coa carboxylase. *Applied Microbiology and Biotechnology* vol. 93 983–992 (2012).

221. Cheng, C. C. *et al.* Discovery and optimization of antibacterial AccC inhibitors. *Bioorg. Med. Chem. Lett.* **19**, 6507–6514 (2009).

222. Payne, D. J. *et al.* Discovery of a novel and potent class of FabI-directed antibacterial agents. *Antimicrob. Agents Chemother.* **46**, 3118–3124 (2002).

223. Joo, S. H. Lipid A as a Drug Target and Therapeutic Molecule. *Biomolecules & Therapeutics* vol. 23 510–516 (2015).

224. Erwin, A. L. Antibacterial Drug Discovery Targeting the Lipopolysaccharide Biosynthetic Enzyme LpxC. *Cold Spring Harb. Perspect. Med.* **6**, (2016).

225. Kalinin, D. V. & Holl, R. LpxC inhibitors: a patent review (2010-2016). *Expert Opin. Ther. Pat.* **27**, 1227–1250 (2017).

226. Kalinin, D. V. & Holl, R. Insights into the Zinc-Dependent Deacetylase LpxC: Biochemical Properties and Inhibitor Design. *Curr. Top. Med. Chem.* **16**, 2379–2430 (2016).

227. Lemaître, N. *et al.* Curative Treatment of Severe Gram-Negative Bacterial Infections by a New Class of Antibiotics Targeting LpxC. *MBio* **8**, (2017).

228. Liu, Y., Yang, L. & Molin, S. Synergistic activities of an efflux pump inhibitor and iron chelators against Pseudomonas aeruginosa growth and biofilm formation. *Antimicrob. Agents Chemother.* **54**, 3960–3963 (2010).

229. Sanchez Clemente, N. *et al.* Bartonella bacilliformis: a systematic review of the literature to guide the research agenda for elimination. *PLoS Negl. Trop. Dis.* **6**, e1819 (2012).

230. Gomes, C. & Ruiz, J. Carrion's Disease: the Sound of Silence. *Clin. Microbiol. Rev.* **31**, (2018).

231. Battisti, J. M., Smitherman, L. S., Samuels, D. S. & Minnick, M. F. Mutations in Bartonella bacilliformis gyrB confer resistance to coumermycin A1. *Antimicrob. Agents Chemother.* **42**, 2906–2913 (1998).

232. Biswas, S., Raoult, D. & Rolain, J.-M. Molecular mechanisms of resistance to antibiotics in Bartonella bacilliformis. *J. Antimicrob. Chemother.* **59**, 1065–1070 (2007).

233. Rolain, J. M. *et al.* Recommendations for treatment of human infections caused by Bartonella species. *Antimicrob. Agents Chemother.* **48**, 1921–1933 (2004).

234. Maguiña, C., Guerra, H. & Ventosilla, P. Bartonellosis. *Clin. Dermatol.* **27**, 271–280 (2009).

235. Farfán-López, M. *et al.* Prioritisation of potential drug targets against *Bartonella bacilliformis* by an integrative *in-silico* approach. *Mem. Inst. Oswaldo Cruz* **115**, (2020).

236. Heath, R. J., Yu, Y. T., Shapiro, M. A., Olson, E. & Rock, C. O. Broad spectrum antimicrobial biocides target the FabI component of fatty acid synthesis. *J. Biol. Chem.* **273**, 30316–30320 (1998).

237. Leibundgut, M., Maier, T., Jenni, S. & Ban, N. The multienzyme architecture of eukaryotic fatty acid synthases. *Curr. Opin. Struct. Biol.* **18**, 714–725 (2008).

238. Hawser, S., Lociuro, S. & Islam, K. Dihydrofolate reductase inhibitors as antibacterial agents. *Biochem. Pharmacol.* **71**, 941–948 (2006).

239. Anderson, K. S., Sikorski, J. A. & Johnson, K. A. Evaluation of 5-enolpyruvoylshikimate-3-phosphate synthase substrate and inhibitor binding by stopped-flow and equilibrium fluorescence measurements. *Biochemistry* **27**, 1604–1610

(1988).

240. Funke, T., Han, H., Healy-Fried, M. L., Fischer, M. & Schönbrunn, E. Molecular basis for the herbicide resistance of Roundup Ready crops. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 13010–13015 (2006).

241. Shippy, D. C. & Fadl, A. A. RNA modification enzymes encoded by the gid operon: Implications in biology and virulence of bacteria. *Microb. Pathog.* **89**, 100–107 (2015).

242. Shippy, D. C., Eakley, N. M., Bochsler, P. N., Chopra, A. K. & Fadl, A. A. Biological and virulence characteristics of Salmonella enterica serovar Typhimurium following deletion of glucose-inhibited division (gidA) gene. *Microb. Pathog.* **50**, 303–313 (2011).

243. Gupta, R., Gobble, T. R. & Schuster, M. GidA posttranscriptionally regulates rhl quorum sensing in Pseudomonas aeruginosa. *J. Bacteriol.* **191**, 5785–5792 (2009).

244. Yim, L., Moukadiri, I., Björk, G. R. & Armengod, M.-E. Further insights into the tRNA modification process controlled by proteins MnmE and GidA of Escherichia coli. *Nucleic Acids Res.* **34**, 5892–5905 (2006).

245. Sangshetti, J. N., Joshi, S. S., Patil, R. H., Moloney, M. G. & Shinde, D. B. Mur Ligase Inhibitors as Anti-bacterials: A Comprehensive Review. *Curr. Pharm. Des.* **23**, 3164–3196 (2017).

246. Olaimat, A. N. *et al.* Emergence of Antibiotic Resistance in Listeria monocytogenes Isolated from Food Products: A Comprehensive Review. *Compr. Rev. Food Sci. Food Saf.* **17**, 1277–1292 (2018).

247. Temple, M. E. & Nahata, M. C. Treatment of listeriosis. *Ann. Pharmacother.* **34**, 656–661 (2000).

248. Pagliano, P., Arslan, F. & Ascione, T. Epidemiology and treatment of the commonest form of listeriosis: meningitis and bacteraemia. *Infez. Med.* **25**, 210–216 (2017).

249. Poyart-Salmeron, C., Carlier, C., Trieu-Cuot, P., Courtieu, A. L. & Courvalin, P. Transferable plasmid-mediated antibiotic resistance in Listeria monocytogenes. *Lancet* **335**, 1422–1426 (1990).

250. Pesavento, G., Ducci, B., Nieri, D., Comodo, N. & Lo Nostro, A. Prevalence and antibiotic susceptibility of Listeria spp. isolated from raw meat and retail foods. *Food Control* vol. 21 708–713 (2010).

251. Morvan, A. *et al.* Antimicrobial resistance of Listeria monocytogenes strains isolated from humans in France. *Antimicrob. Agents Chemother.* **54**, 2728–2731 (2010).

252. Palumbo, M. *et al.* Integrating diverse layers of omic data to identify novel drug targets in Listeria monocytogenes. *Front. Drug. Discov.* **0**, (2022).

253. Lobel, L., Sigal, N., Borovok, I., Ruppin, E. & Herskovits, A. A. Integrative genomic analysis identifies isoleucine and CodY as regulators of Listeria monocytogenes virulence. *PLoS Genet.* **8**, e1002887 (2012).

254. Toledo-Arana, A. *et al.* The Listeria transcriptional landscape from saprophytism to virulence. *Nature* **459**, 950–956 (2009).

255. Chatterjee, S. S. *et al.* Intracellular Gene Expression Profile of *Listeria monocytogenes*. *Infection and Immunity* vol. 74 1323–1338 (2006).

256. Gao, Z. *et al.* PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics* **9**, 104 (2008).

257. Chen, L., Oughtred, R., Berman, H. M. & Westbrook, J. TargetDB: a target registration database for structural genomics projects. *Bioinformatics* **20**, 2860–2862 (2004).

258. Urán Landaburu, L. *et al.* TDR Targets 6: driving drug discovery for human pathogens through intensive chemogenomic data integration. *Nucleic Acids Res.* **48**, D992–D1005 (2020).