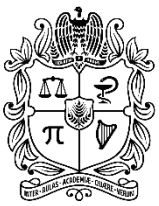


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Comparación de genes que codifican resistencia antibiótica en *Mycobacterium avium Complex* y *Mycobacterium tuberculosis Complex*.

Daniel Andrés Rojas Ospina

Universidad Nacional de Colombia

Facultad de Ingeniería, Ingeniería de sistemas e industrial

Bogotá, Colombia

2017

Comparación de genes que codifican resistencia antibiótica en *Mycobacterium avium Complex* y *Mycobacterium tuberculosis Complex*.

Daniel Andrés Rojas Ospina

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Magister en Bioinformática

Director (a):

PhD. Luis Fernando Niño V.

Línea de Investigación:

Tecnologías computacionales en bioinformática

Grupo de Investigación:

BioLISI

Universidad Nacional de Colombia

Facultad de Ingeniería, Ingeniería de sistemas e industrial

Bogotá, Colombia

2017

A mis padres y a ti Cristian P. mi hermano del alma.

“El futuro pertenece a aquellos que se preparan hoy para él”. Malcolm X.

Agradecimientos

Agradezco a mis padres, por su amor y lucha constante a lo largo de mi vida.

A mi director, PhD Luis Fernando Niño, por permitirme conocer un mundo de conocimiento y abrirme las puertas de su laboratorio de investigación.

A mis compañeros del grupo de investigación LISI.

A mis amigos y familiares, ya que siempre han sido un gran apoyo.

Resumen

En la actualidad, los complejos *M. avium* (MAC) y *M. tuberculosis* (MTBC) son considerados como agentes etiológicos en enfermedades como la tuberculosis humana, una de las 10 principales causas de mortalidad en el mundo. En 2016 10,4 millones de personas han enfermado de tuberculosis y 1,7 millones murieron por esta enfermedad, donde se asocia 0,4 millones de personas con VIH.

La resistencia antibiótica en *MTBC* se ha constituido en una crisis de salud pública y una amenaza para la seguridad sanitaria, según la OMS, hubo 600.000 nuevos casos de resistencia a rifampicina (fármaco de primera línea más eficaz).

Por tal razón, la implementación de nuevas estrategias para el tratamiento de este tipo de enfermedades ha permitido el uso de tecnologías que permiten acelerar pruebas de susceptibilidad a medicamentos.

En este trabajo se presenta un flujo (*pipeline*) para el análisis y comparación de los complejos *Mycobacterium avium* y *Mycobacterium tuberculosis*, con el fin de identificar relaciones basadas en la resistencia a medicamentos que poseen estos dos complejos bacterianos, además de un flujo de ensamblaje y anotación de genomas por medio las herramientas ABySS, SOAPdenovo, SPADES y Velvet.

Para el análisis y comparación de genes de resistencia se utilizó un conjunto de cepas (genomas completos) del ENA asociados a *Mycobacterium tuberculosis complex*, además de un grupo de genomas de *M. avium complex*, estos genomas fueron analizados por la herramienta KvarQ, lo cual permitió el análisis y comparación de genes resistentes a antibióticos de primera línea, las mutaciones asociadas y sus regiones.

Palabras clave: *M. tuberculosis complex*, *M. avium complex,* secuenciación genómica, resistencia antibiótica, tuberculosis.

Abstract

At present, the species found in the M. avium and M. tuberculosis complexes are considered as etiological agents in diseases such as human tuberculosis, one of the 10 leading causes of mortality in the world. By 2016, 10.4 million people have fallen ill with tuberculosis and 1.7 million have died from this disease, where 0.4 million people with HIV are associated.

Antibiotic resistance in MTBC has become a public health crisis and a threat to health security, according to WHO, there was 600,000 new cases of resistance to rifampin (the most effective first-line drug).

For this reason, the implementation of new strategies for the treatment of this type of diseases has allowed the use of technologies that allow accelerating drug susceptibility tests.

This paper presents a pipeline for the analysis and comparison of the Mycobacterium avium and Mycobacterium tuberculosis complexes, in order to identify relationships based on the drug resistance of these two bacterial complexes, as well as an assembly flow and annotation of genomes through the tools ABySS, SOAPdenovo, SPADES and Velvet. For the analysis and comparison of resistance genes we used a set of strains (complete genomes) of the ENA associated with Mycobacterium tuberculosis complex, as well as a group of genomes of *M. avium complex*, these genomes were analyzed by the KvarQ tool, which permited the analysis and comparison of genes resistant to first-line antibiotics, the associated mutations and their regions.

Keywords: *M. tuberculosis complex*, *M. avium complex,* genomic sequencing, antibiotic resistance, tuberculosis.

Tabla de contenido

1. Estado del arte 5
1.1 Género *Mycobacterium* 5
1.1.1 Complejo *Mycobacterium avium* (MAC) 5
1.1.1.1 Características genómicas 6
1.1.2 Complejo *Mycobacterium tuberculosis* (MTBC) 6
1.1.3 Características genómicas 7
1.2 Resistencia antibiótica 7
1.3 Genómica comparativa 9
1.4 Identificación y predicción de genes 10

2. Estudios previos 12
2.1 Bases de datos de resistencia a medicamentos en TB 12
2.2 Predicción y análisis de resistencia antibiótica en micobacterias 14

3. Materiales y métodos 15
3.1 Objetivos 15
3.1.1 Objetivo general 15
3.1.2 Objetivos específicos 15
3.2 Análisis computacional 15
3.2.1 Descarga información genómica 17
3.2.2 Planteamiento del pipeline 17
3.2.2.1 Preprocesamiento 18
3.2.2.2 Ensamblaje de genomas 18
3.2.2.3 Evaluación del ensamblaje 19
3.2.2.4 Anotación 19
3.2.2.5 Estimación de la resistencia 20
3.3 Modificación del pipeline 20
3.3.1 Análisis de la herramienta 20
3.3.2 Funcionamiento de KvarQ 21
3.3.3 KvarQ frente a otras herramientas 22
3.4 Comparar genes 24

4. Resultados y discusión 25
4.1 Diseño pipeline 25
4.1.1 Diseño del pipeline inicial 25
4.1.2 Diseño del pipeline final 26
4.2 Análisis de calidad de las secuencias 27
4.3 Identificación de la resistencia 27
4.4 Comparación de genes de resistencia 29

5. Conclusiones y recomendaciones 30

Lista de figuras

		Pág.
Figura 1-1	Flujo de trabajo de un identificador de mutaciones. Adaptado [55].	11
Figura 2-1	Estructura de la base de datos MUBII-TB-DB, comprendida de la base de datos de mutaciones (recuadro naranja). En la sesión de consulta, la base de datos de mutaciones se utiliza para la interpretación de resultados y para la construcción de la base de datos de mutaciones BLAST. Los óvalos indican el uso de software externo (verde oscuro) o de rutinas desarrolladas por MUBII (verde pálido). El proceso de análisis de consultas (recuadro verde) combina el resultado BLAST y la experiencia del alineamiento utilizando la base de datos de mutaciones. Las salidas están representadas en azul [44].	13
Figura 2-2	Recuento de reporte de medicamentos de TB Drug Resistance Mutation Database asociados a resistencia en *M. tuberculosis*, se puede ver la proliferación de la resistencia a medicamentos de primera línea.	14
Figura 3-1	Descripción del proceso inicial, para el análisis computacional de las secuencias de MAC.	16
Figura 3-2	Descripción del proceso final, para el análisis computacional de las secuencias de MAC (color rojo) y MTBC (color azul), además del proceso de análisis de resistencia antibiótica (color verde).	16
Figura 4-1	Proceso de escaneo y preparación de los fastq [68].	22
Figura 4-2	Descripción del workflow, el cual identifica la resistencia en las cepas de MAC y las compara con las bases de datos de resistencia de MTB.	26
Figura 4-3	Calidad las lecturas por base en el grupo de genomas de MTBC	27
Figura 4-4	Calidad las lecturas por base en el grupo de genomas de MAC	27
Figura 4-5	Análisis de resistencia antibiótica en las cepas analizadas de MTBC, la gráfica de la izquierda muestra los porcentajes asociados a resistencia según el medicamento, mientras que la gráfica de la derecha muestra los porcentajes asociados a los genes de resistencia.	28
Figura 4-6	Gráfico de tendencia en la región asociada a resistencia en Porcentaje de resistencia antibiótica asociado multirresistencia en las cepas analizadas de MTBC.	28
Figura 4-7	Gráfico de tendencia en la región asociada a resistencia en rifampicina para genomas de MTBC	29

Lista de tablas

Pág.

Tabla 1-1: Drogas de primera línea y mecanismos moleculares de resistencia en *M. tuberculosis* [36]. 8
Tabla 1-2: Benchmark comparando herramientas para el análisis de resistencia antibiótica en MTBC [59]……..23

Lista de Símbolos y abreviaturas

Abreviatura	Término
Blast	Herramienta de búsqueda de alineamientos locales (Basic Alignment Search Tool, por sus siglas en inglés)
VIH	Virus de la inmunodeficiencia humana
MAC	Mycobacterium avium complex
MTBC	Mycobacterium tuberculosis complex
SRA	Archivo de lectura de secuencia (Sequence read archive)
NCBI	Centro Nacional de Información Biotecnológica (National Center for Biotechnology Information)
TB	tuberculosis
OMS	Organización mundial de la salud
SIDA	síndrome de la inmunodeficiencia adquirida
GC	Guanina citosina
MDR	Resistencia a múltiples fármacos (multi drug resistance)
RMP	Rifampicina
INH	Isoniacida
XDR	Extremadamente resistente
HMM	Modelos ocultos de Markov (Hidden Markov Models)
EMBL	The European Bioinformatics Institute
OLC	Overlap-Layout-Consensus
SNPs	polimorfismo de un solo nucleótido (Single Nucleotide Polymorphism)
ORFs	Marco abierto de lectura (Open Reading Frame)

Introducción

Las micobacterias son microorganismos distribuidos ampliamente en la naturaleza, el género *Mycobacterium* está formado por bacilos aerobios largos de 1.0 a 10 µm [1], inmóviles y no esporulados, alguno de estos son patógenos causantes de graves enfermedades en mamíferos, como la tuberculosis y la lepra [1]. El género *Mycobacterium* se encuentra ubicado en el *phylum Actinobacteria*, clase *Actinobacteria*, orden *Actinomycetales*, suborden *Corynebacterineae*, familia *Mycobaceriaceae*.

El genoma de las micobacterias comprende aproximadamente 4’411.529 pares de bases, contiene alrededor de 4.000 genes y posee un alto contenido de GC, lo cual denota la ausencia de transferencia horizontal [2].

Las micobacterias se caracterizan por ser microorganismos aerobios o microaerófilos, inmóviles [3], las llamadas micobacterias no tuberculosas (MNT), están distribuidas en diferentes hábitats como el agua, suelo y el tracto gastrointestinal de los animales. Dentro de sus condiciones de supervivencia, a diferencia de otras bacterias patógenas, las micobacterias tienen una gran capacidad de resistencia en el ambiente, ya que poseen una pared celular muy fuerte, lo que las hace resistentes a desinfectantes químicos, ácidos, alcalinos, detergentes e incluso crean resistencia antibiótica.

Dentro del género *Mycobacterium* los dos complejos de gran importancia son el complejo *Mycobacterium avium* (MAC) y complejo *Mycobacterium tuberculosis* (MTBC), debido a su capacidad para producir enfermedades en animales y humanos (especialmente en pacientes inmunodeprimidos), la incidencia de este tipo de bacterias ha estado presente durante toda la historia de la humanidad, siendo agentes importantes de enfermedades infecciosas.

En las últimas décadas, la proliferación de enfermedades asociadas a micobacterias ha aumentado, tomado como ejemplo de la tuberculosis (TB). Un tercio de la población

mundial padece esta enfermedad, produce ocho millones de casos nuevos y mata a más de dos millones de personas al año (según cifras de la OMS) [71]; además de la coinfección entre TB y VIH, problemática que ha sido consecuencia del aumento de índice de pobreza, falla en los sistemas de control de salud pública, escasa investigación en salud, la emergencia de bacilos multirresistente a drogas, entre otros [9].

La resistencia a medicamentos, se ha vuelto el principal problema que poseen las enfermedades asociadas a las micobacterias, ya que recurrir a un tratamiento médico ideal se ha vuelto una tarea compleja, esto se debe al resurgimiento de enfermedades tuberculosas con características epidémicas ligadas a cepas multirresistentes.

En esencia, la resistencia a medicamentos se debe predominantemente a alteraciones en genes que codifican blancos de antibióticos, teniendo como resultado la identificación de múltiples mutaciones asociadas al desarrollo de resistencia a antibióticos de primera línea, como se evidencia en el caso de tratamientos a enfermedades tuberculosas.

Por tal razón, es de suma necesidad implementar estrategias de diagnóstico y estudios de susceptibilidad a drogas que permitan obtener resultados inmediatos. Este trabajo propone validar una estrategia para el análisis y comparación de los genes asociados a resistencia antibiótica de los genomas completos de los complejos *M. avium* y *M. tuberculosis*.

1. Estado del arte

1.1 Género *Mycobacterium*

El género *Mycobacterium* está comprendido por bacterias aerobias y no móviles, las cuales se clasifican como ácido-alcohol resistentes, debido a su pared celular única entre los procariotas. Dicho género pertenece a la familia *Mycobacteriaceae,* la cual comprende alrededor de 150 especies y 11 subespecies. Las micobacterias se caracterizan por ser bacilos rectos o ligeramente curvos [39].

Las micobacterias poseen una serie de lípidos en su envoltura, tales como ácidos micólicos, arabinogalactano, lipoarabinomanano y glicolípidos, que presentan cerca del 60% en peso seco de la pared celular [40]. Estos compuestos en su mayoría hidrofóbicos confieren una serie de características a las micobacterias como la resistencia a antibióticos.

Para clasificar las micobacterias se tiene en cuenta diferente tipo de características como su velocidad de crecimiento y la presencia de pigmentos. A nivel clínico las micobacterias se clasifican según su patogenicidad de la siguiente forma [41]:

- Patógenos estrictos, con alto riesgo de transmisión por el aire
- Patógenos oportunistas, con riesgo individual moderado
    - Oportunistas mayores
    - Oportunistas menores
- Saprófitas, no causantes de enfermedad

Dentro de la clasificación de micobacterias con importancia a nivel clínico, se encuentran los dos grupos de bacterias llamados complejo *Mycobacterium avium* (MAC) y complejo *Mycobacterium tuberculosis* (MTBC) [40].

1.1.1 Complejo *Mycobacterium avium* (MAC)

Este tipo de micobacterias es conocido como no tuberculosas o ambientales, a diferencia del complejo MTBC, MAC posee especies con un mayor grado de divergencia genética y

con importantes diferencias fenotípicas entre ellas, resaltando su ácido-alcohol resistencia, crecimiento lento y producción de pigmento amarillo en ausencia de luz. Este complejo incluye las especies *M. avium* ssp. *avium, M.avium* ssp. paratuberculosis, *M. avium* ssp. *silvaticum, M. avium* ssp. *hominissuis* y *M. intracellulare* [41].

Las bacterias ubicadas en este complejo son ubicuas en el ambiente (aislados en el agua, polvo, alimentos, animales, entre otros), además son capaces de causar enfermedades en los animales y el hombre, en el hombre se pueden generar patologías en personas inmunocompetentes e inmunodeprimidos, sobre todos en infectados con el virus de la inmunodeficiencia humana (VIH). En pacientes con SIDA, la frecuencia de infección causada por MAC aumenta, a lo que se asocia la producción de una infección diseminada, asociada con una gran mortalidad, sobre todo en personas que presentan recuentos de linfocitos CD4 menores a 100 células/mm^3[40].

Dentro del complejo, *M. avium* es la especie con más importancia clínica, ya que aproximadamente el 90% de los aislados de MAC en pacientes con SIDA corresponden a esta especie.

Las infecciones por MAC afectan principalmente al tracto gastrointestinal y pulmón, su sintomatología está asociada con fiebre, tos, dolor abdominal, diarrea y pérdida de peso.

1.1.1.1 Características genómicas

En la actualidad, dentro del complejo *M. avium,* se conocen los genomas de *M. avium* 104 (cepa aislada a partir de sangre de un paciente con SIDA), representante de *M. avium* ssp. *hominissuis* y *M.avium* ssp. *paratuberculosis* cepa K-10 (cepa aislada a partir de una vaca con la enfermedad de Johne) y *M. colombiense* CECT 3035 (secuenciado recientemente) [39] [42].

Los genomas del complejo *M. avium* se caracterizan por su rico contenido en GC, por su tamaño que oscila entre 4,83 y 5,6 Mb dependiendo de la cepa. En la anotación de los genes de *M. avium* se incluyen 5313 genes, de los cuales 5120 codifican para proteínas. La cepa K-10 de *M. avium* se caracteriza porque aproximadamente el 75% del genoma posee homología con la cepa H37Rv de *M. tuberculosis* [40].

1.1.2 Complejo *Mycobacterium tuberculosis* (MTBC)

Este complejo posee organismos altamente relacionados, ya que comparten más del 99% de identidad, está compuesto por las especies *M. tuberculosis, M. africanum, M. canetti,*

M. bovis, M. pinnipedi, M. caprae y *M. microti.* Estas especies poseen diferencias a nivel morfológico, en su bioquímica, tipo de hospederos y patrones de enfermedades en animales.

Las micobacterias de este complejo se identifican por su capacidad de producir tuberculosis (TB), tipo de infección que puede ser TB pulmonar o extrapulmonar. El caso más importante a nivel epidemiológico es la TB pulmonar, ya que es una enfermedad de gran importancia para la salud pública, debido a su constante incremento fruto de problemas de pobreza, tratamientos mal realizados, el virus de inmunodeficiencia humana (VIH), entre otros.

1.1.3 Características genómicas

Este complejo se caracteriza por tener un alto grado de GC de aproximadamente un 65%, el cual está distribuido de forma homogénea, lo que indica que el genoma de estas bacterias no ha recibido información por transferencia horizontal de islas de patogenia. Dentro de su genoma se encuentran alrededor de 4000 genes identificados.

1.2 Resistencia antibiótica

Con la reaparición de la *M. tuberculosis,* la resistencia antibiótica es un problema que ha estado latente durante los últimos años, por tal razón la resistencia en diferentes cepas se ha clasificado como:

- Monoresistencia, clasificada como resistencia a un tipo de antibiótico de primera línea.
- Multiresistencia (MDR), bacterias que desarrollan resistencia a rifampicina (RMP) e isoniazida (INH).
- Ultrarresistencia (XDR), bacterias resistentes a medicamentos de primera línea y a miembros de la familia de las fluroquinolonas, y por lo menos resistencia a alguno de los medicamentos de segunda línea.

En la actualidad no se tiene información acerca de resistencia adquirida a antibióticos en micobacterias como *M. tuberculosis* y *M. avium* por transferencia horizontal de información genética, ya que este tipo de bacilo no posee elementos móviles como plásmidos o transposones [12] [36].

Las micobacterias poseen una resistencia natural a varias drogas ya que poseen una pared compleja e hidrófoba, con una permeabilidad reducida para un gran número de compuestos [13], por esto, las micobacterias son un gran ejemplo de la presión selectiva, mostrando su capacidad de adaptación a medios adversos.

La resistencia a drogas en este tipo de bacilos se debe principalmente a alteraciones en la secuencia de nucleótidos de genes que codifican blancos de antibióticos, ya que en

poblaciones de células sensibles puede haber un remanente de mutantes resistentes por generación, estos mutantes, varían según la droga a la cual presenten resistencia; por esto al formular una sola droga para el tratamiento de enfermedades conferidas a micobacterias, se favorece la selección de mutantes resistentes al tipo de medicamento implementado [14] [15].

En el caso de *M. tuberculosis* la resistencia a drogas se debe a la acumulación de mutaciones individuales en varios genes, siendo responsables de inhibir la función de un antibiótico especifico (Tabla1-1).

Tabla 1-1: Drogas de primera línea y mecanismos moleculares de resistencia en *M. tuberculosis* [36].

Agente antimicrobiano	**Mecanismo de acción de la droga**	**Genes afectados (producto génico) y mecanismo de resistencia**	**Frecuencia de mutaciones**
Isoniazida	Inhibición de la síntesis de ácidos micólicos	KatG: Las mutaciones de este gen impiden la activación de isoniazida. inhA: Las mutaciones de este gen inducen sobreexpresión del gen y niveles elevados de la enzima enoil reductasa en cantidades que superan el poder inhibitorio de INH	50-68% 21-34%
Rifampicina	Inhibición de la transcripción de genes por interacción con ARN polimerasa	rpoB: Las mutaciones en este gen impiden la interacción de la rifampicina con el ARN polimerasa	96-98%
Pirazinamida	No dilucidado	pncA: Las mutaciones en este gen determinan una disminución en la producción del ácido pirazinoico, el metabolito activo	72-97%
Estreptomicina	Inhibición de la síntesis de proteínas	rpsl y rrs, las mutaciones en estos genes impiden la interacción con estreptomicina	64-67% 8-21%
Etambutol	Inhibición de la síntesis de componentes de la pared celular de micobacterias (arabino galactanos)	embCAB: Las mutaciones en este gen o su sobreexpresión permiten la síntesis continua de arabinogalactanos	47-65%

1.3 Genómica comparativa

Se asevera que la bioinformática es “la ciencia de las comparaciones”, reafirmando la sentencia de Champollion “Todo es el resultado de comparaciones” [16].

La genómica comparativa tuvo sus inicios cuando Margaret Dayhoff compilo por primera vez todas las secuencias de aminoácidos de las proteínas conocidas hasta entonces y las comparo mediante el estudio sistemático de los cambios de aminoácidos que se producen en proteínas con funciones similares, lo que llevo a la construcción de matrices de sustitución evolutivas, cuantificando el parecido entro dos secuencias [16].

Otro evento que propicio el surgimiento de la genómica comparativa, fue la secuenciación del ADN, el cual se centra en establecer correspondencias con genes o diferentes características genómicas entre organismos, por ejemplo, encontrar relaciones con genes que perteneces a un grupo de microorganismos relacionados, que pueden afectar la salud [17].

Por tal razón, para analizar el genoma es necesario entender la secuencia del mismo, lo que conlleva a saber que genes están contenidos, su ubicación y que funciones realizan las proteínas que están codificadas. A lo largo de la evolución, los cambios que sufren los genomas y sus genes están sometidos a la presión natural, por esto, la comparación de genomas, genes y proteínas, es la aproximación más intuitiva y directa para interpretar los genomas.

Muchos de los algoritmos actualmente empleados para determinar similitudes entre genes se basan en el análisis comparativo de genomas, los métodos que emplean se basan en la obtención de un alineamiento entre las secuencias, herramienta útil que ayuda a detectar orígenes evolutivos y estructuras comunes, entre otras cosas. La función principal de un alineamiento es establecer un numero de coincidencias entre la comparación de una o más secuencias de cadenas de ADN, ARN o estructuras primarias. Las coincidencias pueden indicar relaciones funcionales o evolutivas entre genes o proteínas. Estos métodos difieren en la relación entre el alineamiento y la predicción de la estructura exónica entre ellas. Las regiones a comparar se dividen en codificantes y no codificantes, los métodos más populares para realizar estas comparaciones son los modelos ocultos de Markov [18].

En la actualidad los alineamientos realizan búsquedas en bases de datos con el fin de encontrar soluciones a problemas de investigación como la búsqueda de secuencias homólogas. Estos sistemas de búsqueda en bases de datos varían según el tipo de

algoritmo implementado, influenciando el tiempo de ejecución y la sensibilidad al realizar el alineamiento [19].

1.4 Identificación y predicción de genes

La predicción e identificación de un gen en procariotas se logra mediante la detección de regiones promotoras y la búsqueda de ORFs. Para realizar este proceso de identificación a nivel computacional, existen diferentes estrategias [32] como los algoritmos basados en modelos probabilísticos, como es el caso de los modelos ocultos de Markov (HMM) ya que un HMM mejora la búsqueda de secuencias relacionadas de forma distante, al convertir un alineamiento múltiple de secuencias en un sistema de puntaje de posiciones específicas, basado en probabilidad [33].

De esta manera, un perfil HMM, contiene estados para coincidencia, inserción o eliminación, los cuales son usados para modelar una familia de secuencias, a partir de un estado se indica con qué probabilidad puede producirse un salto al siguiente estado (por ejemplo, una inserción o deleción). Por lo anterior, los modelos estadísticos como HMM ayudan a resolver problemas como homología lejana [33].

Por lo anterior, al momento de identificar mutaciones y variaciones genómicas, se realiza un mapeo de las lecturas de las secuencias sobre un genoma de referencia, con el fin de detectar alelos mutados basados en la comparación con la secuencia de referencia. Encontrar mutaciones es de gran importancia ya que permite dimensionar un efecto o función de la mutación [55]. Para encontrar dichas mutaciones es necesario usar alineadores, los cuales implementan algoritmos como Burrows-Wheeler y tablas hash [56].

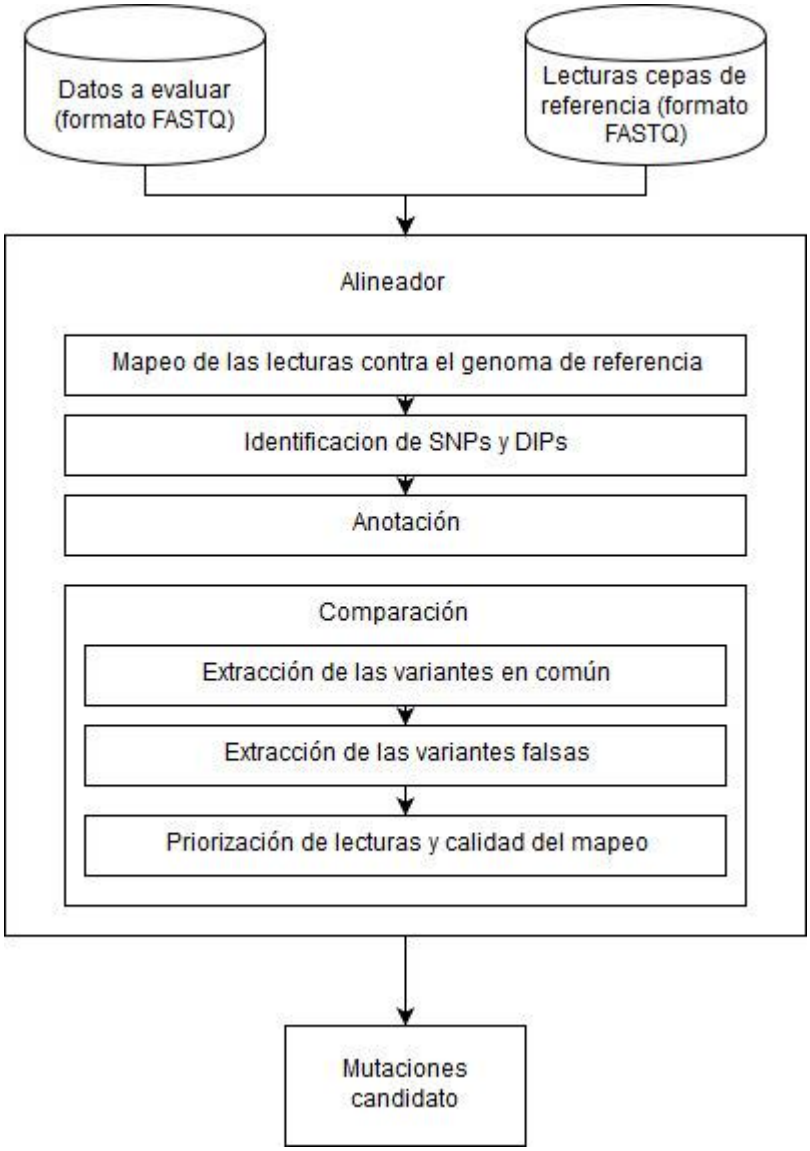


Figura 1-1: Flujo de un identificador de mutaciones. Adaptado [55].

2. Estudios previos

Se cree que la resistencia a fármacos en TB está asociada exclusivamente por mutaciones cromosómicas, que afectan al objetivo del fármaco y a las enzimas bacterianas que activan los profármacos. Desde principio de los años 90, varios estudios han descrito los mecanismos genéticos de resistencia a los fármacos en *M. tuberculosis*, por tal razón se ha almacenado gran cantidad de datos sobre mutaciones asociados provenientes de aislados resistentes a fármacos. Medicamentos como la isoniacida y rifampicina están relacionados con la identificación de un gran número de mutaciones que confieren resistencia, sobre todo con mayor presencia en aislados clínicos. Por otro lado, en los fármacos como la estreptomicina (tabla 1-1) y medicamentos de segunda línea, las mutaciones asociadas a resistencia se producen con menos frecuencia.

2.1 Bases de datos de resistencia a medicamentos en TB

Las bases de datos centralizadas, con información curada de la literatura y de experimentos de alto rendimiento, han ayudado a incursionar en nuevas metodologías para la biología de sistemas, las cuales han ayudado a los investigadores a identificar características importantes que solo podrían haber sido evidentes con el uso del análisis exhaustivo de datos combinados, por tal razón se hizo necesario crear bases de datos que proporcionen un recurso integral y único [43].

Las bases de datos de resistencia permiten acelerar y fomentar nuevos descubrimientos para aplicaciones que pueden ayudar desde el diagnostico hasta el descubrimiento de fármacos. Estas herramientas tienen la ventaja de ser rápidas, de alto rendimiento y fácilmente comparables, sin embargo, su desarrollo para diagnostico se basa en información detalla sobre las mutaciones que conducen a la resistencia a los fármacos y su frecuencia es relativa.

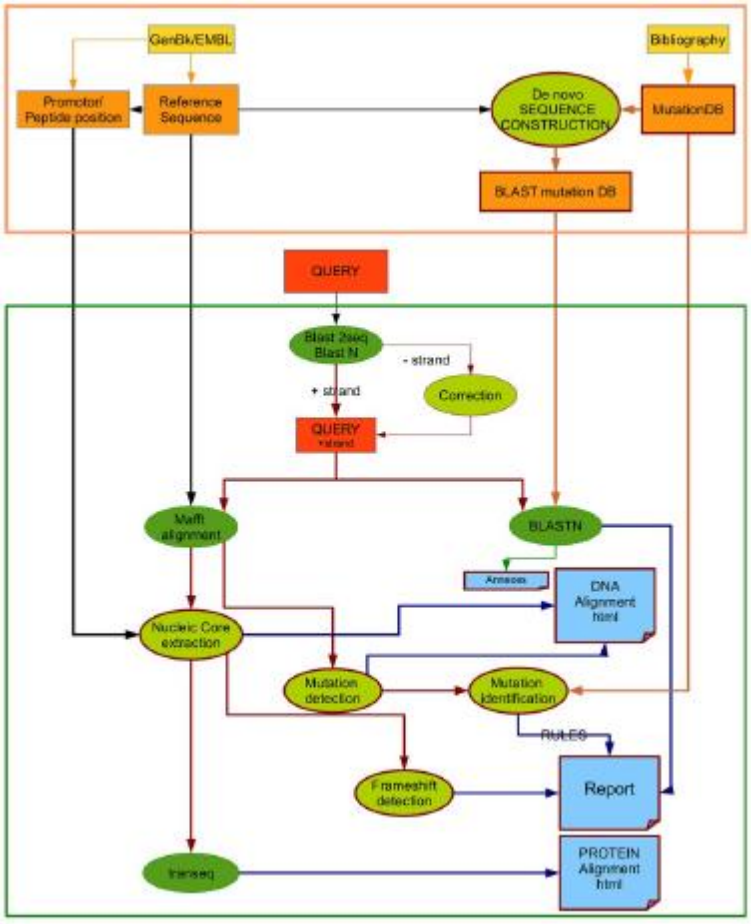


Figura 2-1: Estructura de la base de datos MUBII-TB-DB, comprendida de la base de datos de mutaciones (recuadro naranja). En la sesión de consulta, la base de datos de mutaciones se utiliza para la interpretación de resultados y para la construcción de la base de datos de mutaciones BLAST. Los óvalos indican el uso de software externo (verde oscuro) o de rutinas desarrolladas por MUBII (verde pálido). El proceso de análisis de consultas (recuadro verde) combina el resultado BLAST y la experiencia del alineamiento utilizando la base de datos de mutaciones. Las salidas están representadas en azul [44].

Las bases de datos de resistencia a medicamentos en tuberculosis (Figura 2-1) están centradas en almacenar información relacionada a mutaciones de las diferentes bases de datos biológicas (GeneBank, EMBL) [44]. La base de datos TB Drug Resistance Mutation Database tiene 1178 reportes asociados a 9 medicamentos y 49 genes (Figura 2-2).

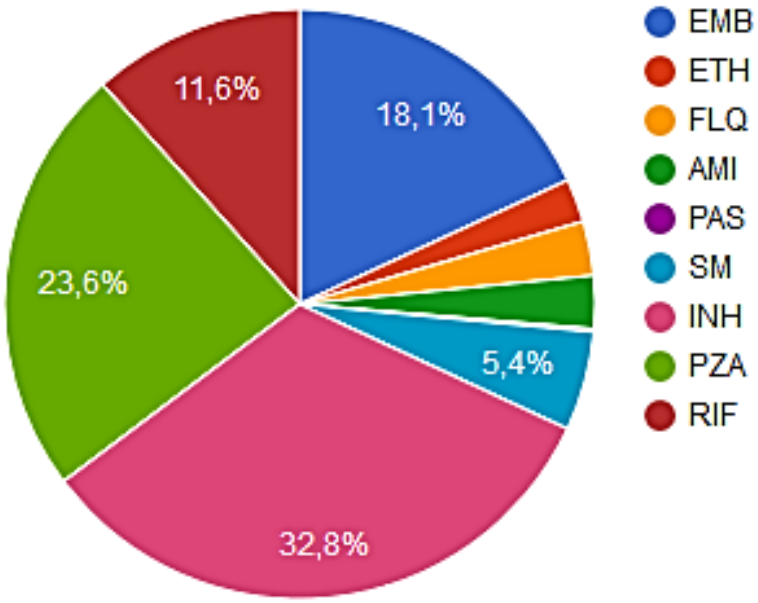


Figura 2-2: Recuento de reporte de medicamentos de TB Drug Resistance Mutation Database asociados a resistencia en *M. tuberculosis*, se puede ver la proliferación de la resistencia a medicamentos de primera línea.

2.2 Predicción y análisis de resistencia antibiótica en micobacterias

La secuenciación del genoma completo ha generado la posibilidad de realizar análisis de resistencias en tuberculosis (TB), por tal razón, herramientas que permiten automatizar y estandarizar el análisis e interpretación de datos han visualizado un futuro promisorio para este tipo de enfermedades.

Las cepas de *M. tuberculosis* se caracterizan por varios medios de tipado, como la detección de genes específicos y el análisis de su posición dentro del genoma de *M. tuberculosis*.

3. Materiales y métodos

3.1 Objetivos

3.1.1 Objetivo general

Comparar los genes de resistencia antibiótica de los genomas completos de miembros de los complejos *Mycobacterium avium* y *Mycobacterium tuberculosis.*

3.1.2 Objetivos específicos

- Proponer una estrategia para la identificación de genes de resistencia en los complejos *Mycobacterium avium* y *Mycobacterium tuberculosis*.
- Identificar los genes que codifican resistencia antibiótica en los complejos *Mycobacterium tuberculosis* y *Mycobacterium avium.*

3.2 Análisis computacional

En el presente trabajo se planteó diferentes flujos computacionales, orientados a resolver los objetivos planteados, teniendo en cuenta que a partir de diferentes metodologías se debe presentar la mejor estrategia para el análisis y comparación de resistencia antibiótica entre los complejos MAC y MTBC, para ello, se presentan dos pipeline que sobresalieron, el primero que se representa de manera general en la figura 3-1, allí sobresalen las etapas de preprocesamiento, ensamblaje de novo y asignación de genes. Posteriormente, fue encontrado un flujo de trabajo que permitió realizar un mejor análisis de la resistencia antibiótica en ambos complejos, por lo cual se decidió continuar y finalizar el proceso con este (figura 3-2). En los numerales posteriores se describirá de forma más detallada las herramientas y métodos.

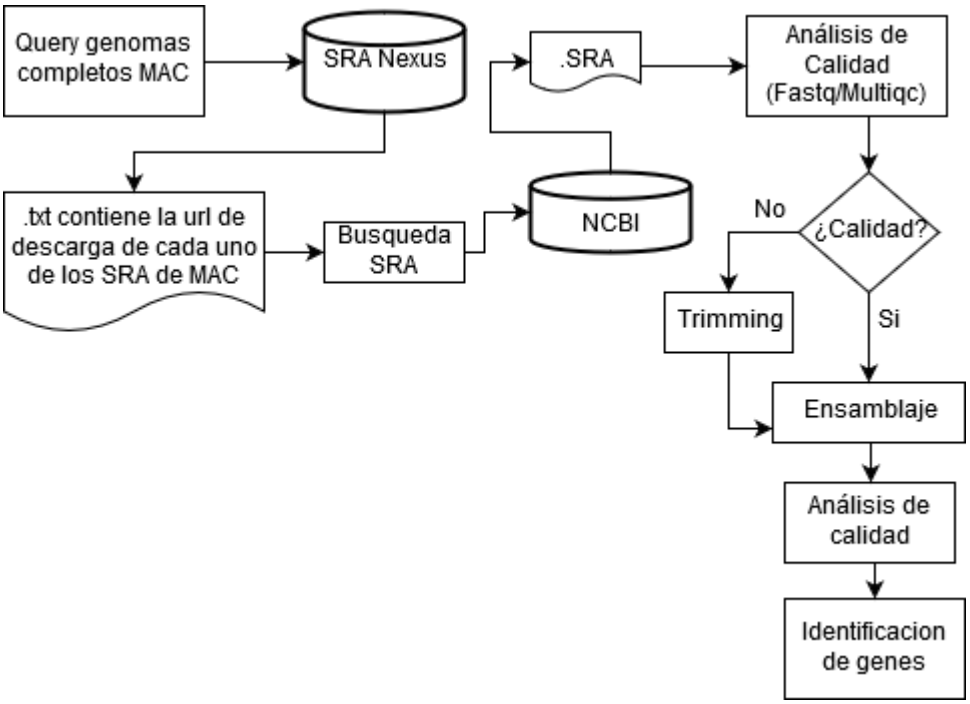


Figura 3-1: Descripción del proceso inicial, para el análisis computacional de las secuencias de MAC.

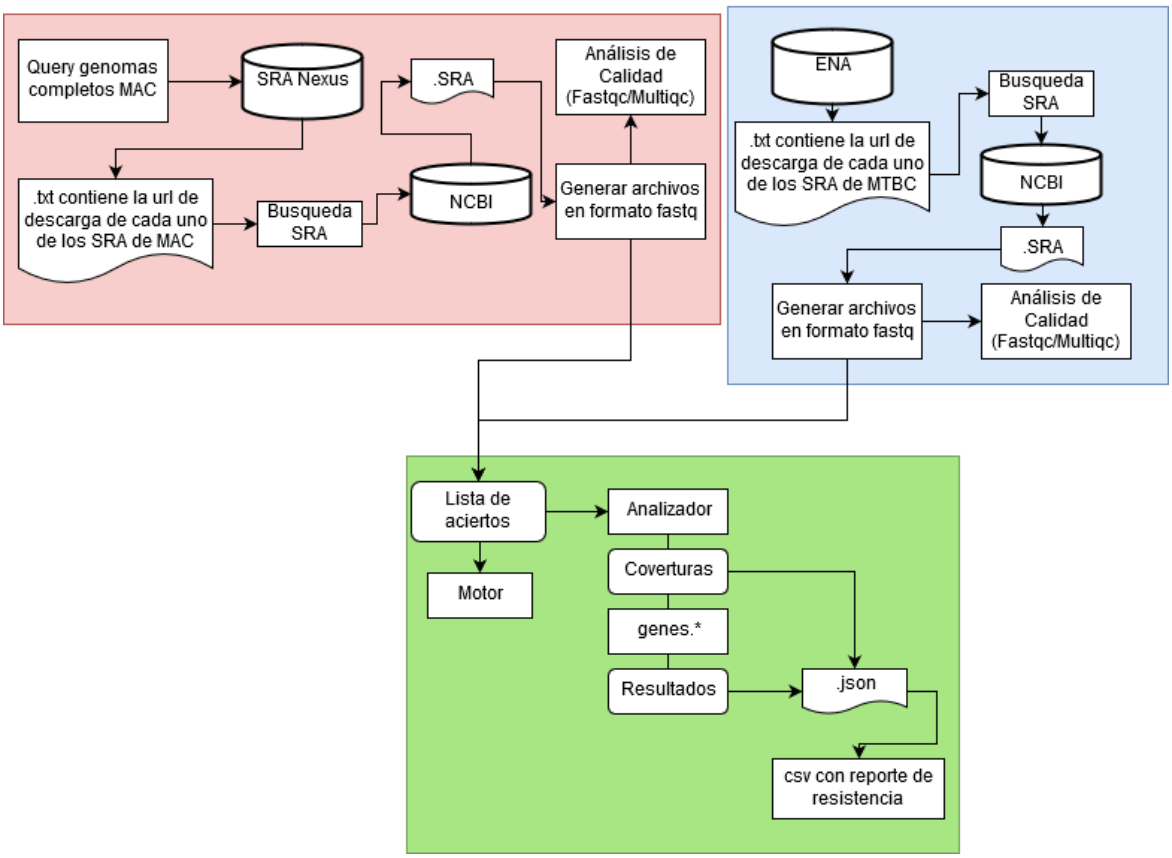


Figura 3-2: Descripción del proceso final, para el análisis computacional de las secuencias de MAC (color rojo) y MTBC (color azul), además del proceso de análisis de resistencia antibiótica (color verde).

3.2.1 Descarga información genómica

Para la descarga de la información genómica de los complejos *M. avium* se usaron los datos alojados en el sitio DNA SRAnexus, el cual permite buscar y filtrar todos los datos accesibles públicamente del archivo de lectura de secuencia disponibles en el NCBI SRA, esta plataforma arrojó un archivo en formato txt, el cual poseía la información de los ftp para la descarga de los datos en formato SRA, por tal razón, desde esta plataforma se descargaron 36 archivos en dicho formato. SRAnexus es una plataforma no disponible desde el 2017, por lo cual se reemplazó este proceso para la descarga de los genomas del complejo *M. tuberculosis*. Para la descarga de cepas del complejo *M. tuberculosis* se usó un conjunto de cepas del ENA, con numero de acceso PRJEB7727, por lo cual se descargaron 264 formatos SRA para el caso de MTBC, este grupo de genomas completos están relacionadas con la secuenciación completa del genoma de aislados del complejo *M. tuberculosis* de Sierra Leona [57] [58]. Este grupo de cepas ha sido descrito anteriormente, teniendo en cuenta que han sido reportadas como resistentes a algún antibiótico. Las cepas son altamente diversas filogenéticamente y comprenden genotipos de *M. africanum* con linajes de *M. tuberculosis* (*Beijing, Camerún, EAI, Ghana, Haarlem,* tipo S, *Sierra Leona* y tipo X) [59].

3.2.2 Planteamiento del pipeline

En el presente trabajo se tiene dentro de sus objetivos plantear la mejor estrategia para identificar la resistencia antibiótica en los complejos *mycobacterium avium* y *mycobacterium tuberculosis*, por ende, en el periodo de exploración inicial se planteó un flujo computacional que se representa de manera general en la figura 3-1, donde se destacan las etapas de procesamiento, ensamblaje, análisis de calidad e identificación de genes. Luego de ello se pretendía hacer una estimación de la resistencia antibiótica, para realizar la comparación de generes asociados a resistencia antibiótica entre los dos complejos.

En este pipeline inicial se planteó realizar un ensamblaje de novo, para encontrar nuevas variantes asociadas con la resistencia antibiótica, ya que implementar un flujo de ensamblaje de genomas de referencia se asocia con la resecuenciación de los genomas a estudiar. Para el desarrollo de este flujo, se incorporaron los siguientes pasos:

3.2.2.1 Preprocesamiento

En la etapa de preprocesamiento de los datos se implementó el SRA Toolkit [45], herramienta mediante la cual se puede acceder a los datos descargados en la etapa de "Descarga de información genómica" y convertirlos a otro tipo de formato, en este caso implementar el formato fastq, para analizar la calidad de los datos crudos.

Para el análisis de calidad se implementó FastQC, herramienta que genera un reporte que puede identificar problemas en el material secuenciado que puede haber sido originado en el secuenciador o en la librería. FastQC examina el rango de valores de calidad para las lecturas en cada posición.

Al tener un alto número de genomas es necesario realizar un reporte general que agrupe en un solo formato los archivos FastQC, para lo cual se implementó MultiQC [61].

Los errores encontrados en los archivos son eliminados con Trimmomatic [46], herramienta que realiza cortes a datos de tecnología Illumina, usando puntuaciones de calidad phred 33 y adaptadores universales [46].

3.2.2.2 Ensamblaje de genomas

Todos los métodos de ensamblaje se basan en la unión de fragmentos de ADN altamente relacionados dentro de la posición del genoma. Por esto, la similitud entre secuencias de ADN se usa para conectar fragmentos de secuencias contiguas o *contigs* (secuencias consenso obtenidas a partir del ensamblaje de las lecturas).

Las estrategias empleadas por los ensambladores de secuencias se pueden agrupar en tres paradigmas principales: *Greedy, Overlap-Layout-Consensus* (OLC) y grafos de Bruijin [47].

Greedy, es un algoritmo que conecta las lecturas que mejor se solapan, de manera iterativa, siempre y cuando no contradiga el ensamblaje construido. Este algoritmo no es muy empleado ya que no usa información global y no resuelve de manera clara largas regiones repetitivas de los genomas [48].

OLC, es un método que primero identifica todas las lecturas pares que se solapan muy bien y organiza esta información en un grafo, el cual permite el desarrollo de complejos algoritmos de ensamblaje, que tienen en cuenta la relación global entre las lecturas. El genoma es construido por la búsqueda de un único camino que atraviese todos los nodos sólo una vez [49].

Grafos de Bruijin, este algoritmo modela la relación entre subcadenas de longitud k, conocidos como k-meros, exactas dentro de las lecturas. De forma parecida a OLC, los nodos en el grafo representan k-meros adyacentes que se solapan por *k-1* letras, teniendo que la longitud del *k-mero* correlaciona con la longitud del solapamiento que el ensamblador es capaz de detectar [47] [50].

Para el proceso de ensamblaje de novo, se implementaron 4 herramientas Abyss [62], SOAPdenovo [66], SPADES [64], Velvet [65], con k-meros de 55 y 77, los otros parámetros se usaron por defecto para lecturas finales pareadas y corrección de errores, allí se evaluaron el número de *contigs*, la longitud del *contig* más largo, el N50 y N75.

3.2.2.3 Evaluación del ensamblaje

Hay diferentes métricas que permiten evaluar la calidad del ensamblaje, como el análisis de medida de los *contigs*, así como el total del ensamblaje que debe coincidir con el tamaño esperado del genoma. A nivel estadístico el N50 es el índice de la continuidad del genoma, ya que corresponde con el menor de los mayores *contigs* que cubren la mitad del genoma ensamblado [47] [51].

Para este proceso se utiliza QUAST [67].

3.2.2.4 Anotación

El proceso de anotación permite interpretar la información contenida en el genoma ensamblado, por tal razón es necesario identificar las principales características. El proceso de anotación está comprendido por dos procesos: la anotación estructural (predicción de genes codificantes) y la anotación funcional (asignación de información biológica a los genes previamente predichos) [47]. Para este caso, se implementó el método de novo de la herramienta RAST, el cual utiliza algoritmos estadísticos para determinar si la secuencia es codificante, por medio de la detección de patrones o motivos específicos de la secuencia [52].

3.2.2.5 Estimación de la resistencia

Para determinar la resistencia se planteó el uso de herramientas como ARG-ANNOT y ResFinder, estas fueron creadas con el fin de encontrar genes de resistencia a antibióticos en genomas bacterianos por medio del uso de BLAST [53].

ARG-ANNOT, posee información del NCBI GeneBank y su base de datos incluye alrededor de 1.689 genes de resistencia a antibióticos [54].

3.3 Modificación del pipeline

Para estimar de la resistencia antibiótica en MAC y MTBC, se determinó que herramientas anteriormente mencionadas como ARG-ANNOT y ResFinder no poseían un sistema de bases de datos abierta, además de no contar con un API para la conexión con *bash*, por lo cual se recurrió a un flujo alterno para determinar la resistencia antibiótica en los genomas como se muestra en el recuadro verde de la figura 4-3, por tal razón, se implementó la herramienta KvarQ (Version 0.12.2) [68], la cual escanea los archivos fastq directamente, omitiendo la necesidad de mapear todas las lecturas de secuenciación a un genoma de referencia o realizar un proceso de ensamblaje de novo. KvarQ carga "testsuites" que definen SNPs específicos o regiones cortas de interés en un genoma de referencia, lo que permite sintetizar directamente los resultados relevantes en función de la aparición de marcadores en los fastq, teniendo un 99% de congruencia [68], por tal razón, se decidió tener esta herramienta como solución final para el desarrollo de este proyecto de tesis.

3.3.1 Análisis de la herramienta

KvarQ permite realizar el análisis sobre archivos en formato fastq, detectando el estado alélico de polimorfismos conocidos. Esta herramienta permite extraer información relevante directamente de las lecturas de secuenciación, sin necesidad de asignar un genoma de referencia sobre las lecturas [68].

Antes del proceso de escaneo, las secuencias marcador se generan a partir de mutaciones conocidas o regiones de interés. Dichas secuencias marcador se alojan en bancos de pruebas, los cuales definen los polimorfismos o SNPs relevantes. Luego, durante el proceso de escaneo, las lecturas en formato fastq son recortadas dependiendo de su puntuación en la escala PHRED [68], la cual se define con un valor de corte Q=13.

Después del proceso anterior, las lecturas que coinciden con la secuencia objetivo y excedieron el puntaje de calidad mínimo, se ensamblan a una cobertura que indica la profundidad de cobertura general, así como las mutaciones detectadas.
una de las principales características y principales diferencias de KvarQ con las herramientas de análisis genético actual, es que este software extrae información relevante de las lecturas de secuenciación, sin necesidad de asignar cada lectura a un genoma de referencia.

3.3.2 Funcionamiento de KvarQ

Antes de realizar un escaneo de los fastq a identificar, hay que tener en cuenta que las secuencias marcador se generan a partir de mutaciones conocidas o regiones de interés. por ello, KvarQ define unos polimorfismos (SNP) o regiones de interés que poseen mutaciones potenciales, estos módulos se definen como “testsuites” parametrizadas en Python.
Las secuencias marcador se extraen de genomas de referencias, teniendo en cuenta que para estas secuencias marcador también se genera su correspondiente cadena complementaria.
el algoritmo de escaneo se implementa en C, lo que permite es dividir el archivo fastq en partes pequeñas, que se distribuyen en múltiples hilos y se escanean simultáneamente. inicialmente cada lectura se recorta mediante un puntaje de calidad dado por el usuario. la parte recortada se desplaza a lo largo de cada una de las secuencias objetivo y cada coincidencia se registra en la lista de aciertos si se garantiza una superposición mínima. el proceso de escaneo continua hasta que se escanea todo el archivo o se alcanza una cobertura mínima específica.
Luego de ello, los datos recopilados de los fastq se traducen en estructuras de datos intermedias. Estas estructuras de datos también llamadas coberturas, combinan las lecturas de la cadena original y complementaria, las ubican dentro del genoma y resumen las bases que no coinciden en las lecturas, como se muestra en la figura 4-1.
KvarQ omite las bases que no satisfacen los requisitos mínimos de calidad de la lectura determinado por el usuario.

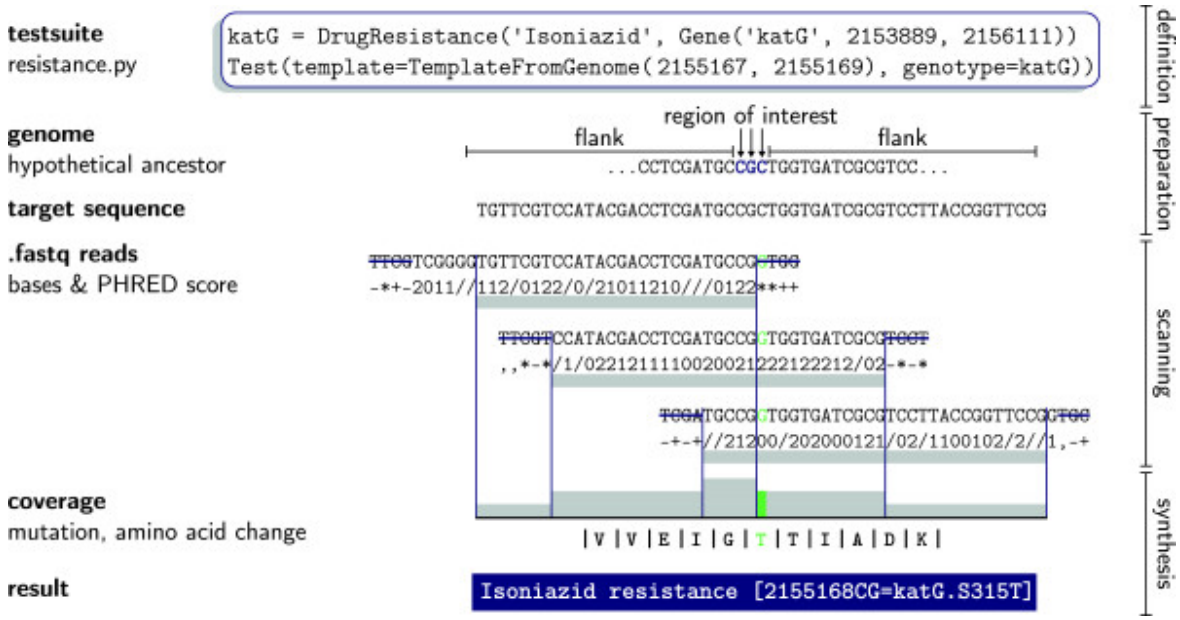


Figura 4-2: proceso de escaneo y preparación de los fastq [68].

3.3.3 KvarQ frente a otras herramientas

Kvar es una herramienta que presenta características que son bastante útiles a la hora de realizar un proceso de automatización, teniendo en cuenta que posee un proceso de carga de procesos por lotes, además de presentar la facilidad de escanear archivos fastq por medio de la comparación de variantes conocidas, teniendo en cuenta que esas variantes se pueden ir actualizando dentro de las bases de KvarQ, también se puede modificar esas bases para encontrar resistencia sobre otro tipo de bacterias.

A continuación una tabla comparativa con otras herramientas que realizan un análisis de resistencia antibiótica sobre MTBC.

Tabla 1-2: Benchmark comparando herramientas para el análisis de resistencia antibiótica en MTBC [59].

Característica	CASTB (Versión 1.1)	KvarQ (Versión 0.12.2)	Mykrobe Predictor TB (Versión 0.1.3)	PhyResSE (Versión 1.0)	TBProfiler
Basado en la web	Sí (es necesario registrarse)	No	No	Sí	Sí
Proceso por lotes	No	Sí	Sólo en la versión de línea de comandos	Sí	Sólo en la versión de línea de comandos
Lecturas de extremo emparejado	Sí	Sólo archivos combinados	Sólo archivos combinados	Sí	Sí
Tubería	Montaje de terciopelo de-novo;BIGSdb ; Mapeo MUMmer; scr ipts personalizad os	Módulos / paquetes de Python con extensiones C (sin asignación)	Mapeo Stampy (H37Rv Versión 2); variante llamada SAMtools	Mapeo BWA (H37Rv Versión 3), preprocesamiento y variante de llamada con GATK	mapeo instantáneo (regiones seleccionadas de H37Rv versión 3);variante llamada SAMtools
Informe de lenguaje sencillo	Sí	Conversión de .json a .csv	Sí	Si detallada	Sí
Exportación de datos	Sí	Sí	Sí	Sí	No
Bases de la predicción de linaje.	Virtual LSP, spoligot yping *in silico* , RFLP o MIRU-VNTR	Comas *et al* . 2009, Stucki*et al* . 2012 y no publicados.	Stucki et al. 2012	Homolka *et al* . 2012, Coll *et al* .2014 y otros publicados	Coll *et al* . 2014
Detección de NTM	Sí	Por defecto solo MTBC, sin embargo es posible alimentar su librería	Sí	No (solo MTBC)	No (solo MTBC)
Variantes reportadas	Ninguna	Solo mutaciones de resistencia.	Solo mutaciones de resistencia.	Todas las mutaciones	Todas las mutaciones en los genes candidatos.

Detección de infecciones mixtas	Sólo Beijing-no Beijing	Sí	No	Sí	Sí
Hetero Resistencia	No	Visual	Sí, sin puntajes de calidad.	> 10%	Sí
Posible modificación	No	Sí	Sí	Sí	Sí

3.4 Comparar genes

Para comparar los genes resultantes de las herramientas anteriormente mencionadas, se implementaron los resultados arrojados por la herramienta KvarQ, por lo cual se comparan las regiones asociadas a la resistencia, los SNPs y los genes asociados a la misma.

4. Resultados y discusión

Cabe aclarar que el principal inconveniente para el desarrollo de este flujo de trabajo es que no hay mucha información a nivel genómico relacionada con el complejo *M. avium*, teniendo en cuenta que las bases de datos de referencia como el NCBI en su reporte de ensamblaje y anotación posee alrededor de 17 genomas completos reportados, a diferencia de *M. tuberculosis*, el cual posee en la anotación de genomas completos alrededor de 141 genomas reportados.

Otro de los retos en este proceso fue encontrar bases de datos de resistencia antibiótica de fácil consulta, ya que la mayoría de accesos y reportes, están ligados a información que no se encuentra completamente publica en la web.

Teniendo en cuenta lo anterior, el desarrollo de este proyecto de tesis tiene como primera etapa proponer una estrategia para identificar genes de resistencia en los complejos *M. avium y M. tuberculosis*, para generar la identificación de estos genes, los cuales se explicarán en los numerales 4.1 a 4.4, con el fin de desarrollar el objetivo principal.

4.1 Diseño pipeline

De acuerdo con la revisión bibliográfica realizada, se encontró diferentes herramientas y bases de datos para la identificación de resistencia antibiótica en bacterias, las cuales usan varios tipos de algoritmos que comparan regiones y mutaciones asociadas. Por tal razón, fue necesario determinar la información genómica a rastrear sobre los dos complejos de *mycobacterium*.

Al identificar los genes de resistencia asociados con el complejo *M. tuberculosis*, con sus respectivos porcentajes de mutación asociados a resistencia antibiótica tabla 1-1, fue necesario identificar el grupo de antibióticos de primera línea que tienen los dos complejos en común en el tratamiento médico, para este caso, los medicamentos implementados para el tratamiento de pacientes con alguna cepa del complejo *M. avium* son Rifampicina y Etambutol [69]. Por tal razón la prioridad es encontrar un flujo de trabajo idóneo frente a la identificación de resistencia asociada a este tipo de medicamentos.

4.1.1 Diseño del pipeline inicial

Para el diseño del pipeline se planteó un flujo de trabajo inicial, el cual se observa en la figura 4-2, donde se realizó un proceso de ensamblaje de genomas (tabla 1-1) para implementar la identificación de genes asociados a resistencia, y por medio del uso de un identificador de genes, predecir resistencia a partir de datos basados en modelos de

homología y SNPs, además de la implementación de herramientas como ResFinder y ARG-ANNOT, los cuales permiten encontrar resistencia antibiótica asociada a genes.

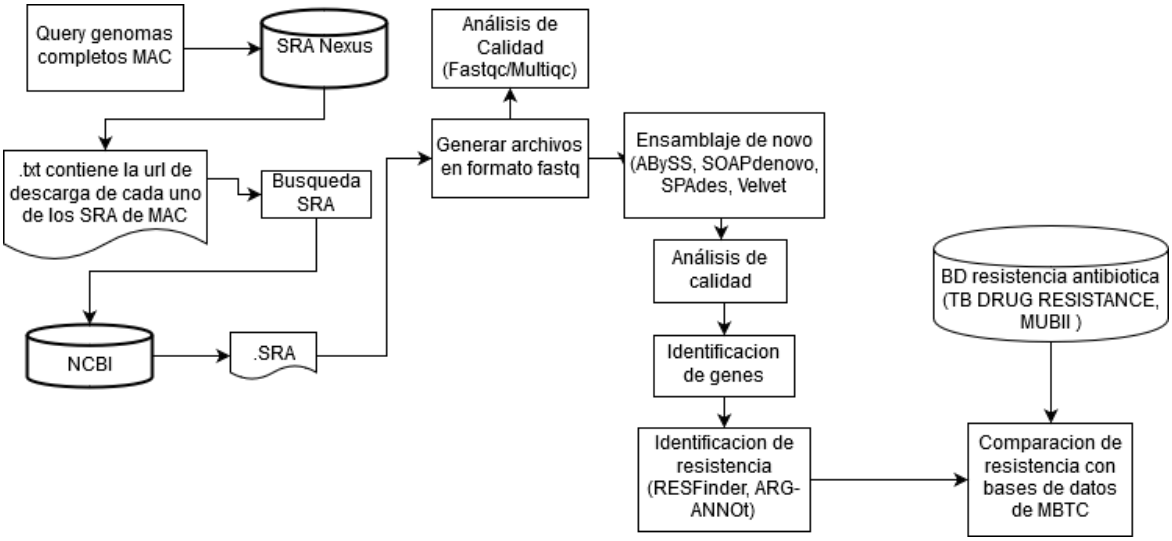


Figura 4-2: Descripción del flujo que identifica la resistencia en las cepas de MAC y compara con las bases de datos de resistencia de MTB.

4.1.2 Diseño del pipeline final

Para el diseño del pipeline final se tuvo en cuenta el análisis de genomas en crudo, el cual se observa en la figura 3-2. Ya que KvarQ (Version 0.12.2) [68] es una herramienta que permite escanear directamente archivos fastq, analizando variantes o SNPs para secuencias de genoma bacteriano, lo cual genera una gran ventaja al omitir la necesidad de mapear todas las lecturas de secuenciación a un genoma de referencia o realizar un proceso de ensamblaje de novo.

Para el análisis de información, KvarQ carga "*testsuites*" útiles, que definen SNPs específicos o regiones cortas de interés en un genoma de referencia, lo cual permite sintetizar los resultados relevantes en función de aparición de marcadores en los fastq [68].

Otra de las ventajas por las cuales se eligió KvarQ es el uso de carga de datos por lotes (versión de línea de comandos), lo que permite el análisis en paralelo de múltiples genomas, a diferencia de otras herramientas que poseen versiones de procesamiento en línea.

4.2 Análisis de calidad de las secuencias

Dentro de los genomas descargados del complejo *M. tuberculosis* y *M. avium* se observa que la calidad, posee buenos valores, a pesar de la dispersión de secuencias finales, siendo algo intrínseco producto de las tecnologías de secuenciación.

Figura 4-3: Calidad de las lecturas por base en el grupo de genomas de MTBC

Figura 4-4: Calidad de las lecturas por base en el grupo de genomas de MAC

4.3 Identificación de la resistencia

Dentro de los genomas descargados del complejo *M. tuberculosis*, al implementar KvarQ se encontró resistencia a medicamentos de primera línea como isoniacida, estreptomicina, entre otros, como se evidencia en la figura 4-5; cabe aclarar que el reporte que genera la herramienta indica baja cobertura para algunas secuencias.

El análisis de resistencia está asociada a los genes encontrados en el reporte arrojado por KvarQ simplificado en la figura 4-5, cabe resaltar que los mayores picos de resistencia antibiótica se caracterizan en los genes katG, rpsL y rpoB, a su vez, los medicamentos menos eficaces para el tratamiento son la isoniacida, estreptomicina y rifampicina.

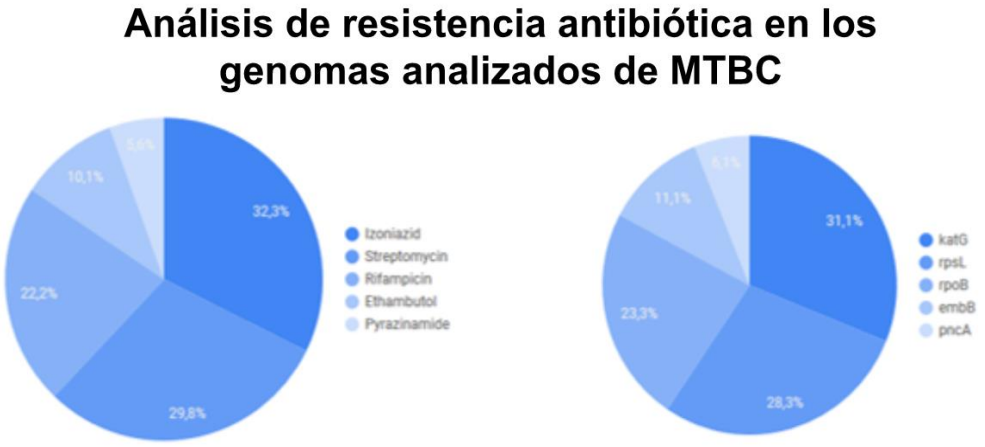


Figura 4-5: Análisis de resistencia antibiótica en las cepas analizadas de MTBC, la gráfica de la izquierda muestra los porcentajes asociados a resistencia según el medicamento, mientras que la gráfica de la derecha muestra los porcentajes asociados a los genes de resistencia.

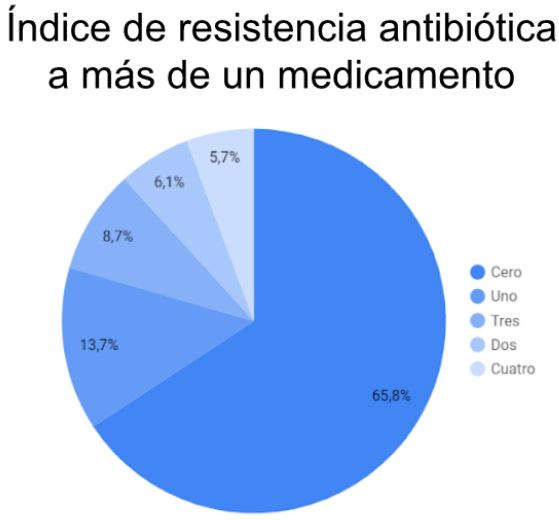


Figura 4-6: Porcentaje de resistencia antibiótica asociado multirresistencia en las cepas analizadas de MTBC.

Para el análisis de resistencia antibiótica en MAC, se encontraron múltiples secuencias con baja cobertura y sin información relacionada a resistencia antibiótica, por lo cual solo

se encontró la cepa ERR026811, la cual posee resistencia a Rifampicina por el gen rpoB con un cambio en la región 761135.
Dentro del proceso de análisis de la herramienta KvarQ se utilizó bacterias XDR (ultra resistentes) encontradas en el ENA, con numero de acceso PRJNA227148, donde se encuentra resistencia en múltiples antibióticos (Anexo E)

4.4 Comparación de genes de resistencia

Teniendo en cuenta que, solo se encontró un genoma con resistencia antibiótica en el complejo *M. avium*, es necesario realizar una comparación por referencias bibliográficas, ya que se han determinado mutaciones ligadas al gen rpoB del complejo *M. avium* y estas se encuentran asociadas a regiones de resistencia en rifampicina dentro del complejo *M. tuberculosis* [70].

Lo anterior indica que la cepa ERR026811 de *M. avium* concuerda con el gen rpoB, el cual es blanco de resistencia a rifampicina.

Dentro del análisis de los datos de MTBC, se indica que la media de las cepas asociadas a resistencia en rifampicina está alojada en la región 761.155, teniendo en cuenta que el valor máximo y mínimo de rangos asociados a mutaciones está ubicado entre la región 761.095 y 764.817 respectivamente.

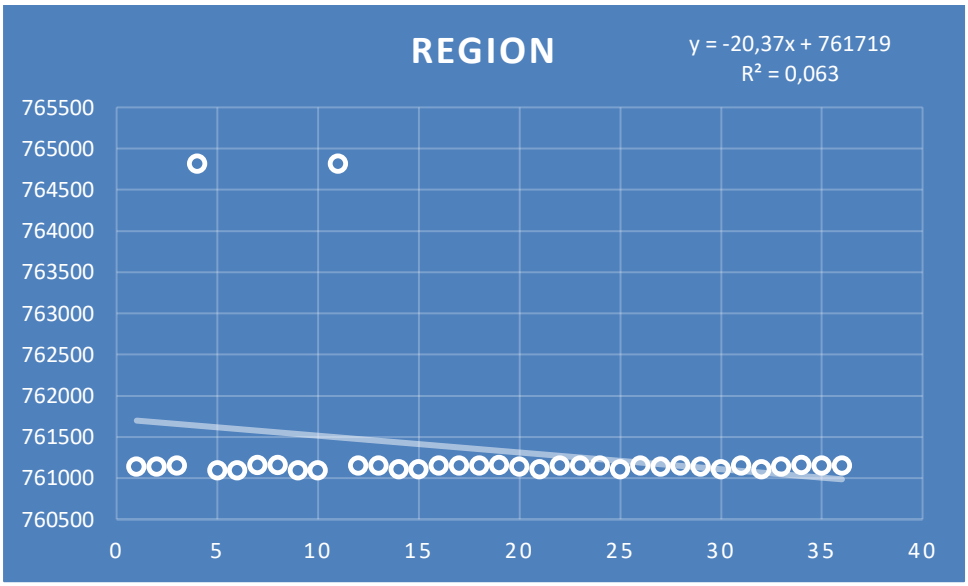


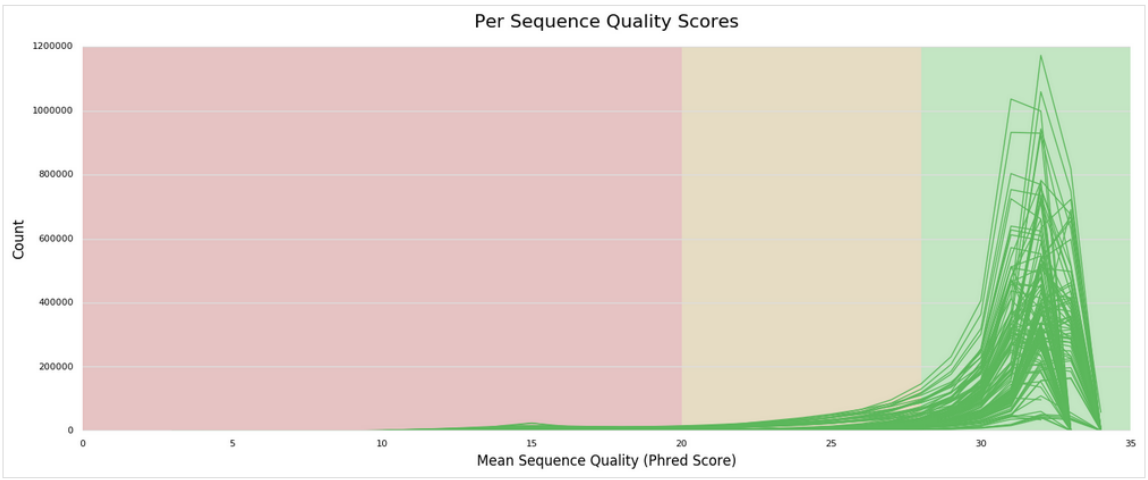
Figura 4-7: Gráfico de tendencia en la región asociada a resistencia en rifampicina para genomas de MTBC

5. Conclusiones y recomendaciones

- Las cepas de *M. tuberculosis complex* que poseen un mayor número de resistencia está asociada a los genes katG, rpsL y rpoB, los cuales están asociados a los antibióticos isoniacida, rifampicina y estreptomicina, lo cual valida la información relacionada en la tabla 1-1, donde las frecuencias de mutación asociadas a esos genes son las más altas.
- La única cepa de *M. avium complex*, que registro resistencia antibiótica, posee resistencia asociada al gen rpoB, por lo cual, se infiere que la resistencia coincide con la información bibliográfica, donde se reporta resistencia antibiótica en ese gen para ese complejo [70].
- La identificación de regiones asociadas a resistencia en el complejo *M. avium* es muy pobre debido a la falta de información que posea una mejor cobertura dentro de la tecnología de secuenciación.
- Se sugiere implementar este proceso con aislados propios, donde se pueda implementar cepas diferentes a las alojadas en las bases de datos públicas.

.

A. Anexo: Reporte de calidad luego del proceso de Trimming



Reporte de calidad luego de realizar el proceso de trimming con Trimmomatic a 64 genomas seleccionados de MAC para la construcción del preprocesamiento.

B. Genomas RefSeq NCBI

Especies MTBC	Cantidad	katG	inhA	rpoB	rpsl	rrs	embCAB	gyrA	gyrB
Africanum	2	1	0	0	0	0	0	0	1
Bovis	8	3	0	0	0	0	0	0	3
Canettii	5	5	0	0	0	0	0	0	5
Caprae	1	0	0	0	0	0	0	0	0
Colombiense	1	0	0	0	0	0	0	0	0
MTB	140	20	0	0	0	0	0	0	20

Análisis realizado a los genomas de RefSeq del NCBI, por medio de scripts hechos en Python, usando librería BIO para la conexión con el NCBI, en este proceso se seleccionó todos los documentos de tipo GBFF y analizando los genes implicados en resistencia antibiótica. Una de las características es que estos genomas se encuentran anotados por referencia con la cepa H37Rv, por lo cual arrojaron este volumen de coincidencias.

C. Anexo: Análisis de calidad del ensamblaje con QUAST

Report

	55	77
# contigs (>= 0 bp)	8061	6907
# contigs (>= 1000 bp)	122	122
# contigs (>= 5000 bp)	0	0
# contigs (>= 10000 bp)	0	0
# contigs (>= 25000 bp)	0	0
# contigs (>= 50000 bp)	0	0
Total length (>= 0 bp)	3046380	2797777
Total length (>= 1000 bp)	152663	152667
Total length (>= 5000 bp)	0	0
Total length (>= 10000 bp)	0	0
Total length (>= 25000 bp)	0	0
Total length (>= 50000 bp)	0	0
# contigs	1450	1450
Largest contig	3327	3327
Total length	1017639	1017643
GC (%)	66.44	66.44
N50	680	680
N75	576	576
L50	567	567
L75	975	975
# N's per 100 kbp	0.00	0.00

Report

	55	77
# contigs (>= 0 bp)	8061	6907
# contigs (>= 1000 bp)	122	122
# contigs (>= 5000 bp)	0	0
# contigs (>= 10000 bp)	0	0
# contigs (>= 25000 bp)	0	0
# contigs (>= 50000 bp)	0	0
Total length (>= 0 bp)	3046380	2797777
Total length (>= 1000 bp)	152663	152667
Total length (>= 5000 bp)	0	0
Total length (>= 10000 bp)	0	0
Total length (>= 25000 bp)	0	0
Total length (>= 50000 bp)	0	0
# contigs	1450	1450
Largest contig	3327	3327
Total length	1017639	1017643
GC (%)	66.44	66.44
N50	680	680
N75	576	576
L50	567	567
L75	975	975
# N's per 100 kbp	0.00	0.00

Reporte de calidad del ensamblaje realizado con el uso de k55 y k77 respectivamente, con los ensambladores ABYSS y SPADES a una misma cepa al azar. Las dos tablas presentan el mismo conjunto de datos, por lo cual se infiere que los dos ensambladores poseen el mismo comportamiento (número de contigs, N50 y N75).

D. Anexo: Secuencias de comandos para el flujo de descarga de archivos SRA y cambio de formato a fastq en MTBC

- #Esta línea permite descargar la informacion contenida en un archivo en #formato.txt, allí están contenidas todas las url a descargar
  cat sraFiles.txt | xargs -n3 -P3 -I{} wget -d {}
- #Esta línea descomprime todos los archivos descargados anteriormente
  ls *.sra.gz | xargs -n3 -P3 -I{} gzip -d {}
- #Esta línea cambia el formato sra a fastq
  fastq-dump ./*.sra

E. Anexo: Análisis de KvarQ sobre una bacteria de MTBC ultrarresistente

Dentro del proceso de análisis y validación de KvarQ, se realizó una prueba de análisis sobre una bacteria ultra resistentes, para este caso se usó *M. tuberculosis* XDR KZN 605 [72], teniendo como resultado resistencia a Amikacina, Kanamicina, Isoniacida, Rifampicina, Etambutol y Fluroquinolonas

filename	MTBC/resistance	MTBC/resistance	MTBC/resistance	MTBC/resistance	MTBC/resistance
XDR KZN 605	Kanamycin/Amikacin resistance::SNP1473246AG=rrsK.S467S	Isoniazid resistance::SNP1673432TA=inhA promoter mutation -8	Rifampicin resistance (RRDR) [761110AG=rpoB.D435G]	Fluoroquinolones resistance (QRDR) [7570CT=gyrA.A90V]	Ethambutol resistance::SNP4247429AG=embB.M306V

Bibliografía

[1] Rodriguez, G. (2016). *Micobacterias*. Retrieved 6 May 2016, from http://www.higiene.edu.uy/cefa/2008/micobacterias.pdf

[2] Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., & Harris, D. et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, *393*(6685), 537-544. http://dx.doi.org/10.1038/31159

[3] Torres-Ávila, J., Colorado, C., Gamboa, L., & Araujo, M. (2009). Genotipificación de Mycobacterium leprae colombiano para la Determinación de Patrones de Transmisión de la Enfermedad. *Salud Pública.*, *11*(1), 3-13. Retrieved from http://www.scielosp.org/pdf/rsap/v11n1/v11n1a02.pdf

[4] Uribarren Berrueta, T. (2015). *TUBERCULOSIS - Recursos en Bacteriología - UNAM*. *Facmed.unam.mx*. Retrieved 6 May 2016, from http://www.facmed.unam.mx/deptos/microbiologia/bacteriologia/tuberculosis.html

[5] Fountan FF. Tuberculosis update. J Infect Dis 2001; 94: 117

[6] Rodríguez Cuns, G. (2011). *Género Mycobacterium*. Retrieved 6 May 2016, from http://www.higiene.edu.uy/cefa/Libro2002/Cap%2024.pdf

[7] *List of Prokariotic names with standing in nomenclature*. (2016). *Bacterio.net*. Retrieved 12 May 2016, from http://www.bacterio.net/mycobacterium.html

[8] Nightingale, S., Byrd, L., Southern, P., Jockusch, J., Cal, S., & Wynne, B. (1992). Incidence of Mycobacterium avium-intracellulare Complex Bacteremia in Human Immunodeficiency Virus-Positive Patients. *Journal Of Infectious Diseases*, *165*(6), 1082-1085. http://dx.doi.org/10.1093/infdis/165.6.1082

[9] *Tuberculosis*. (2016). *Organización Mundial de la Salud*. Retrieved 6 May 2016, from http://www.who.int/mediacentre/factsheets/fs104/es/

[10] Kotilainen, H., Valtonen, V., Tukiainen, P., Poussa, T., Eskola, J., & Järvinen, A. (2015). Clinical findings in relation to mortality in non-tuberculous mycobacterial infections: patients with Mycobacterium avium complex have better survival than patients with other mycobacteria. *Eur J Clin Microbiol Infect Dis*, *34*(9), 1909-1918. http://dx.doi.org/10.1007/s10096-015-2432-8

[11] Brosch, R., Gordon, S., Marmiesse, M., Brodin, P., Buchrieser, C., & Eiglmeier, K. et al. (2002). A new evolutionary scenario for theMycobacterium tuberculosiscomplex. *Proceedings Of The National Academy Of Sciences*, *99*(6), 3684-3689. http://dx.doi.org/10.1073/pnas.052548299

[12] Hernández Neuta, J. (2011). *DESARROLLO DE UN MÉTODO MOLECULAR ALTERNATIVO PARA LA DETECCIÓN DE CEPAS DE Mycobacterium tuberculosis CON RESISTENCIA EXTENDIDA A DROGAS* (Maestría). Universidad Nacional de Colombia, Facultad de ciencias, Departamento de Química.

[13] Lemus Molina, D. (2007). *Métodos rápidos para la detección de resistencia en Mycobacterium tuberculosis* (Licenciatura). INSTITUTO DE MEDICINA TR OPICAL¨PEDRO KOURI¨SUBDIRECCION DE MICROBIOLOGIA DEPARTAMENTO DE BACTERIOLOGIA-MICOLOGIA.

[14] Culligan T. Drug-resistant tuberculosis: desperate measures?. J Infect Dis 2001; 357: 1124.

[15] David HL. Probability distribution of drug-resistant mutans in unselected populations of M. Tuberculosis. Appl Microbiol 1970; 20: 810-814.

[16] Bandyopadhyay, S., Maulik, U., & Roy, D. (2008). Gene Identification: Classical and Computational Intelligence Approaches. *IEEE Transactions On Systems, Man, And Cybernetics, Part C (Applications And Reviews)*, *38*(1), 55-68. http://dx.doi.org/10.1109/tsmcc.2007.906066

[17] Alfoldi, J. & Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. *Genome Research*, *23*(7), 1063-1068. http://dx.doi.org/10.1101/gr.157503.113

[18] Yoon, B. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *CG*, *10*(6), 402-415. http://dx.doi.org/10.2174/138920209789177575

[19] Jones, N. & Pevzner, P. (2004). *An introduction to bioinformatics algorithms*. Cambridge, MA: MIT Press.

[20] Măndoiu, I., Sunderraman, R., & Zelikovsky, A. (2008). *Bioinformatics research and applications*. Berlin: Springer.

[21] Roy, D. (2009). *Bioinformatics*. Oxford, U.K.: Alpha Science International.

[22] Giegerich, R. (2000). A systematic approach to dynamic programming in bioinformatics. *Bioinformatics*, *16*(8), 665-677. http://dx.doi.org/10.1093/bioinformatics/16.8.665

[23] *Dynamic programming and sequence alignment*. (2016). *Ibm.com*. Retrieved 6 May 2016, from http://www.ibm.com/developerworks/library/j-seqalign/

[24] Harris, R. (2007). *Improved pairwise alignment of genomic DNA*.

[25] Miyano, S. (2005). *Research in computational molecular biology*. Berlin [u.a.]: Springer.

[26] Leszczynski, J. (1999). *Computational molecular biology*. Amsterdam: Elsevier.

[27] Notredame, C. (2007). Recent Evolutions of Multiple Sequence Alignment Algorithms. *Plos Comput Biol*, *3*(8), e123. http://dx.doi.org/10.1371/journal.pcbi.0030123

[28] Notredame, C. (2007). Recent Evolutions of Multiple Sequence Alignment Algorithms. *Plos Comput Biol*, *3*(8), e123. http://dx.doi.org/10.1371/journal.pcbi.0030123

[29] Chenna, R. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, *31*(13), 3497-3500. http://dx.doi.org/10.1093/nar/gkg500

[30] Ari Löytynoja, A. & Goldman, N. (2004). An algorithm for progressive multiple alignment of sequences with insertions, *102*(30), 10557–10562. Retrieved from http://www.pnas.org/content/102/30/10557.short

[31] Chaves, D., Sandoval, A., Rodriguez, L., García, J., Restrepo, S., & Zambrano, M. (2010). Análisis comparativo de seis genomas del complejo Mycobacterium tuberculosis.

[32] Bandyopadhyay, S., Maulik, U., & Roy, D. (2008). Gene Identification: Classical and Computational Intelligence Approaches. *IEEE Transactions On Systems, Man, And Cybernetics, Part C (Applications And Reviews)*, *38*(1), 55-68. http://dx.doi.org/10.1109/tsmcc.2007.906066

[33] Baxevanis, A. (2016). An Overview of Gene Identification: Approaches, Strategies, and Considerations. Retrieved from http://sun-lab.med.nyu.edu/files_private/sun-lab/attachments/CPBI.Ch4.Findgenes.pdf

[34] Fontalvo D., Gómez D. Genes del *Mycobacterium tuberculosis* involucrados en la patogenicidad y resistencia a antibióticos durante la tuberculosis pulmonar y extrapulmonar. MÉD.UIS.2015;28(1):39-51.

[35] Vissa V., Sakamuri R., Li W., Brennan PJ. Defining mycobacteria: Shared and specific genome features for different lifestyles. Indian J Microbiol. 2009;49:11-47.

[36] Arráiz, N., Bermúdez, B., & Urdaneta, B. Resistencia a drogas en M. Tuberculosis : Bases moleculares. Retrieved from http://www.revistaavft.com/avft_24_1_2005/3.pdf

[37] Colombia, M. (2016). *¿Qué es tuberculosis (TB)?*. *Minsalud.gov.co*. Retrieved 6 May 2016, from https://www.minsalud.gov.co/salud/publica/PET/Paginas/Tuberculosis.aspx

[38] *About RefSeq : NCBI Reference Sequence Database*. (2016). *Ncbi.nlm.nih.gov*. Retrieved 6 May 2016, from http://www.ncbi.nlm.nih.gov/refseq/about/

[39] Leguizamón Guerrero, John Emerson (2012) *Búsqueda de marcadores genómicos y lipídicos de Mycobacterium colombiense, una micobacteria potencialmente virulenta perteneciente al complejo MAC.*Maestría thesis, Universidad Nacional de Colombia.

[40] Sánchez, J. (2008). *COMPLEJO MYCOBACTERIUM AVIUM: DIAGNÓSTICO, CARACTERIZACIÓN MOLECULAR E INTERFERENCIA CON EL DIAGNÓSTICO DE LA TUBERCULOSIS*. Doctorado. UNIVERSIDAD COMPLUTENSE DE MADRID.

[41] Zhang, Z., Pang, Y., Wang, Y., Cohen, C. and Zhao, Y. (2015). Differences in risk factors and drug susceptibility betweenMycobacterium avium and Mycobacterium intracellulare lung diseasesin China. *International Journal of Antimicrobial Agents*, pp.491–495.

[42] Sevilla Agirregomoskorta, I. (2007). *CARACTERIZACIÓN MOLECULAR, DETECCIÓN Y RESISTENCIA DE MYCOBACTERIUM AVIUM SUBESPECIE PARATUBERCULOSIS*. Doctorado. Universidad del País Vasco.

[43] Sandgren, A., Strong, M., Muthukrishnan, P., Weiner, B., Church, G. and Murray, M. (2009). Tuberculosis Drug Resistance Mutation Database. *PLoS Med*, 6(2), p.e1000002.

[44] Flandrois, J., Lina, G. and Dumitrescu, O. (2014). MUBII-TB-DB: a database of mutations associated with antibiotic resistance in Mycobacterium tuberculosis. *BMC Bioinformatics*, 15(1), p.107.

[45] Ncbi.nlm.nih.gov. (n.d.). *Using the SRA Toolkit to convert .sra files into other formats*. [online] Available at: https://www.ncbi.nlm.nih.gov/books/NBK158900/ [Accessed 22 Nov. 2016].

[46] Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

[47] Aguilar-Bultet, L. and Falquet, L. (2015). Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos. *Salud Anim*, 37(2), pp.125-132.

[48] Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11(1):31-46.

[49] Treangen TJ, Salzberg SL. Repetitive DNA and nextgeneration sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13(1):36-46.

[50] Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14(3):157-167.

[51] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, Glenn Tesler; QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013; 29 (8): 1072-1075. doi: 10.1093/bioinformatics/btt086.

[52] Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., … Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, *42*(Database issue), D206–D214. http://doi.org/10.1093/nar/gkt1226.

[53] Kleinheinz, K. A., Joensen, K. G., & Larsen, M. V. (2014). Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage*, *4*, e27943. http://doi.org/10.4161/bact.27943.

[54] Gupta, S., Padmanabhan, B., Diene, S., Lopez-Rojas, R., Kempf, M., Landraud, L. and Rolain, J. (2013). ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrobial Agents and Chemotherapy*, 58(1), pp.212-220.

[55] lida N, Yamao F, Nakamura Y, lida T: Mudi, a web tool for identifying mutations by bioinformatics analysis of whole-genome sequence. Genes to Cells 2014, 19:517-527.

[56] Zhang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B: Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis. BioMed Research International 2014, 2014: 1-16.

[57] Homolka, S. *et al. High genetic diversity among Mycobacterium tuberculosis complex strains from Sierra Leone*. *BMC Microbiology*8, 103. http://www.biomedcentral.com/1471-2180/8/103/abstract, doi: 10.1186/1471-2180-8-103 (2008).

[58] Feuerriegel, S. *et al. Sequence analysis for detection of first-line drug resistance in Mycobacterium tuberculosis strains from a high-incidence setting*. *BMC Microbiology*12, 90. http://www.biomedcentral.com/1471-2180/12/90/abstract, doi: 10.1186/1471-2180-12-90 (2012).

[59] Schleusener, V., Köser, C., Beckert, P., Niemann, S. and Feuerriegel, S. (2017). Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. *Scientific Reports*, [online] 7, p.46327. Available at: https://www.nature.com/articles/srep46327 [Accessed 6 Aug. 2017].

[60] Andrews S: FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2015.

[61] Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. http://doi.org/10.1093/bioinformatics/btw354.

[62] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ: ABySS: A parallel assembler for short read sequence data. Genome Res 2009, 19:1117–1123.

[63] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ: ABySS: A parallel assembler for short read sequence data. Genome Res 2009, 19:1117–1123.

[64] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V, Sirotkin A V, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA: SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 2012, 19:455–477.

[65] Zerbino DR, Birney E: Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008, 18:821–829.

[66] Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung D, Ming S, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Wah T, Wang J: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 2012, 1:1-6.

[67] Yandell M, Ence D: A beginner's guide to eukaryotic genome annotation. Nature Reviews 2012, 13:329-342.

[68] Steiner, A., Stucki, D., Coscolla, M., Borrell, S. and Gagneux, S. (2014). KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*, 15(1), p.881.

[69] Piersimoni C, Scarparo C. Extrapulmonary infections associated with nontuberculous mycobacteria in inmunocompetent persons. Emerg Infect dis, 2009; 15 (9):131-138

[70] Obata, S., Zwolska, Z., Toyota, E., Kudo, K., Nakamura, A., Sawai, T., Kuratsuji, T. and Kirikae, T. (2006). Association of rpoB mutations with rifampicin resistance in Mycobacterium avium. *International Journal of Antimicrobial Agents*, 27(1), pp.32-39.

[71] Organización Mundial de la Salud. (2017). *Tuberculosis*. [online] Available at: http://www.who.int/mediacentre/factsheets/fs104/es/ [Accessed 4 Apr. 2017].

[72] Uddin, R., Siddiqui, Q. N., Azam, S. S., Saima, B., & Wadood, A. (2018). Identification and characterization of potential druggable targets among hypothetical proteins of extensively drug resistant Mycobacterium tuberculosis (XDR KZN 605) through subtractive genomics approach. European Journal of Pharmaceutical Sciences, 114, 13–23. doi:10.1016/j.ejps.2017.11.014