UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Desarrollo de *pipeline* bioinformático para la detección de genes de resistencia a antibióticos en bacterias asociadas a la dieta de *Ceratitis capitata*

Trabajo de graduación presentado por Luis Pedro García Salazar para optar al grado académico de Licenciado en Ingeniería en Bioinformática

Guatemala,

2024

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

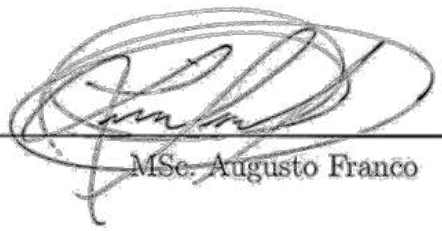
Desarrollo de *pipeline* bioinformático para la detección de genes de resistencia a antibióticos en bacterias asociadas a la dieta de *Ceratitis capitata*

Trabajo de graduación presentado por Luis Pedro García Salazar para optar al grado académico de Licenciado en Ingeniería en Bioinformática

Guatemala,

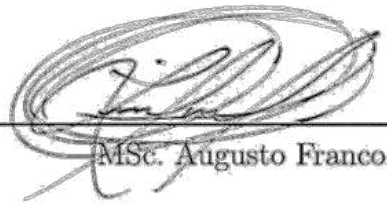
2024

Vo.Bo.:

(f) 

MSc. Augusto Franco

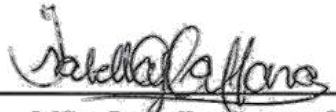
Tribunal Examinador:

(f) 

MSc. Augusto Franco

(f) 

MSc. Douglas Leonel Barrios Gonzalez

(f) 

MSc. Isabella García Caffaro

Fecha de aprobación: Guatemala, 23 de mayo de 2024.

Índice

Lista de figuras

Lista de cuadros

Resumen

En este proyecto, se desarrolló y validó un *pipeline* bioinformático destinado a la identificación de genes de resistencia a antibióticos en bacterias, particularmente, aquellas que forman parte de la dieta de *Ceratitis capitata*. La necesidad de este *pipeline* surge de la urgencia de comprender y mitigar los riesgos asociados con la liberación de insectos que pueden portar bacterias resistentes a antibióticos, lo que tiene implicaciones significativas para la salud pública.

El *pipeline* se diseñó para ser accesible y fácil de usar, permitiendo que usuarios sin conocimientos avanzados en bioinformática puedan operarlo. Utiliza herramientas estándar de bioinformática como *fastp* y *SPAdes* para procesar secuencias de ADN; y está equipado para proporcionar, no solo identificaciones de genes de resistencia, sino también información detallada sobre el genoma a lo largo de cada etapa de análisis.

Se validó utilizando un protocolo basado en el modelo de *Neisseria meningitidis*, empleando métricas como exactitud, precisión, sensibilidad y especificidad, las cuales demostraron la efectividad del *pipeline* al identificar los genes de resistencia. La aplicación del *pipeline* a muestras reales reveló la presencia de múltiples genes de resistencia, subrayando su relevancia y utilidad práctica.

Abstract

In this project, a bioinformatic *pipeline* was developed and validated for the identification of antibiotic resistance genes in bacteria, particularly in those that are part of the diet of *Ceratitis Capitata*. The need for this *pipeline* arises from the urgency to understand and mitigate the risks associated with the release of insects that may carry antibiotic-resistant bacteria, which have significant implications for public health.

The *pipeline* was designed to be accessible and easy to use, allowing users without advanced knowledge in bioinformatics to operate it effectively. It uses standard bioinformatics tools such as *fastp* and *SPAdes* to process DNA sequences and is equipped to provide not only resistance gene identifications but also detailed information about the genome throughout each stage of analysis .

It was validated using a protocol based on the *Neisseria meningitidis* model, using metrics such as accuracy, precision, sensitivity and specificity, which demonstrated the effectiveness of the *pipeline* in correctly identifying resistance genes. The application of the *pipeline* to real samples revealed the presence of multiple resistance genes, underlining its relevance and practical usefulness.

CAPÍTULO 1

Introducción

La resistencia a los antibióticos en bacterias representa un desafío que se ha colocado entre las primeras diez amenazas para la salud pública según la Organización Mundial de la Salud (OMS) (UN, 2015). Este fenómeno ha llegado a tal nivel de emergencia ya, que está asociado con la ineficacia de medicamentos, la dificultad para el tratamiento de enfermedades infecciosas, mayores tasas de hospitalización, costos de atención médica y mortalidad (Ciorba et al., 2015).

Ante el aumento de este fenómeno, se hace necesario desarrollar herramientas que permitan detectar de manera simple y precisa la presencia de genes de resistencia a antibióticos en bacterias. En este trabajo, se desarrolla una herramienta de este tipo en el contexto de la mosca *Ceratitis capitata*, también conocida como mosca de la fruta o del Mediterráneo. Esta mosca, considerada una plaga, tiene una importancia económica crucial debido a su papel en cultivos frutales. Por ello, es objeto de programas de control biológico mediante la técnica del insecto estéril, que implica la liberación de insectos estériles para competir con los fértiles y reducir la población total (Plá et al., 2021). Sin embargo, es esencial asegurarse de que estas moscas liberadas no porten bacterias con resistencia a antibióticos.

Por lo tanto, el objetivo principal de este trabajo es desarrollar un *pipeline*, o programa bioinformático, eficiente para la detección de genes de resistencia a antibióticos en bacterias asociadas a la dieta de *C. capitata*. De esta forma, se puede garantizar que las moscas liberadas no porten bacterias con resistencia a antibióticos en su microbioma.

La metodología para desarrollar el programa consistió en crear un sistema de procesamiento de los datos iniciales, es decir, la secuencia del genoma de las bacterias, para terminar con un análisis completo de los genes de resistencia que pueden presentarse. Entre los pro-

cesos del programa, se incluye un control de calidad de los datos, un ensamblaje y anotación genómica, y, al final, una comparación de los resultados con bases de datos de genes de resistencia conocidos.

El programa es usado inicialmente por usuarios que trabajan con moscas de la fruta, para detectar bacterias perjudiciales en su dieta. Las primeras pruebas del programa se realizaron con varios genomas bacterianos, logrando identificar la presencia o ausencia de bacterias con resistencia a antibióticos.

CAPÍTULO 2

Justificación

La creciente emergencia de resistencia a antibióticos en bacterias es un problema global que compromete la efectividad de los tratamientos médicos y aumenta los riesgos asociados a infecciones bacterianas. Según estimaciones recientes, en 2019, 1.27 millones de muertes fueron atribuidas directamente a infecciones resistentes a medicamentos a nivel mundial. Para 2050, se espera que hasta 10 millones de muertes ocurran anualmente debido a esta causa (UN, 2015). Algunos patógenos resistentes, como el *Staphylococcus aureus* resistente a la meticilina (MRSA), representan una amenaza particularmente grave. En Estados Unidos, el MRSA causa más muertes cada año que el VIH/SIDA, la enfermedad de Parkinson, el enfisema y los homicidios, combinados (Ventola, 2015).

Ante este panorama, resulta esencial hallar maneras de reducir la propagación de estas bacterias. Este trabajo aborda la detección de resistencia antibiótica en un vector biológico significativo: la mosca de la fruta o del Mediterráneo, *Ceratitis capitata*. Monitorear y controlar la resistencia a los antibióticos en este vector es crucial, dado su impacto en los ecosistemas agrícolas y su uso en programas de control biológico (Dionysopoulou et al., 2020). La detección de bacterias resistentes que forman parte de la dieta de *C. capitata* puede prevenir la introducción de cepas resistentes en ambientes naturales o agrícolas, donde podrían transferirse a otras especies bacterianas, amplificando el problema de la resistencia a los antibióticos a nivel global (Ventola, 2015).

Por lo tanto, el desarrollo de un programa bioinformático que simplifique este proceso de detección es fundamental. La mayoría de las metodologías existentes para identificar genes de resistencia a partir de secuencias genéticas implican numerosos pasos técnicos y requieren un conocimiento avanzado en tecnologías bioinformáticas. Sin embargo, este programa ofrece

una interfaz sencilla donde los usuarios pueden ingresar datos de secuenciación y recibir rápidamente un análisis sobre la presencia de genes de resistencia. Al hacerlo, el programa elimina obstáculos técnicos y facilita que un mayor número de personas pueda detectar genes de resistencia a antibióticos en bacterias en diversos contextos. Además, la relevancia de este programa se extiende más allá de su aplicación inicial con *C. capitata*. El programa también puede detectar la presencia de genes de resistencia a antibióticos en cualquier bacteria, lo que lo convierte una herramienta versátil en el campo de la microbiología y la salud pública.

CAPÍTULO 3

Objetivos

3.1. Objetivo general

Desarrollar un *pipeline* bioinformático eficiente para la detección de genes de resistencia a antibióticos en bacterias asociadas a la dieta de *Ceratitis capitata*

3.2. Objetivos específicos

- Desarrollar la capacidad de generar informes que presenten de manera clara y significativa los resultados del *pipeline*.
- Crear un programa interactivo y fácil de usar que genere datos complementarios útiles, enriqueciendo la utilidad del *pipeline*.

CAPÍTULO 4

Marco teórico

4.1. Antibióticos

Los antibióticos son compuestos químicos utilizados para prevenir y tratar infecciones causadas por bacterias. Estos medicamentos pueden ser de origen natural, semisintético o sintético y actúan específicamente contra las bacterias sin afectar significativamente a las células del huésped. La capacidad de los antibióticos para combatir infecciones bacterianas ha transformado la medicina moderna, permitiendo el tratamiento efectivo de enfermedades previamente mortales (Hutchings et al., 2019).

4.1.1. Mecanismos de acción

Los antibióticos pueden clasificarse según su mecanismo de acción en varias categorías principales:

- **inhibidores de la síntesis de la pared celular:** Estos antibióticos, como la penicilina y las cefalosporinas, impiden la formación de la pared celular bacteriana, lo que es esencial para la supervivencia de las bacterias. La ausencia de una pared celular robusta lleva a que las bacterias se vuelvan osmóticamente inestables y mueran (Hutchings et al., 2019).

- **Inhibidores de la síntesis de proteínas:** Antibióticos como los macrólidos, tetraciclinas y aminoglucósidos se unen a las subunidades ribosómicas bacterianas, impi-

diendo la síntesis de proteínas esenciales para el crecimiento y la replicación de las bacterias (Zinner, 2007).

- **Inhibidores de la síntesis de ácidos nucleicos:** Compuestos como las quinolonas y rifampicinas interfieren con las enzimas responsables de la replicación del ADN bacteriano o la transcripción del ARN, inhibiendo así la proliferación bacteriana (Zinner, 2007).

- **Alteradores de la membrana plasmática:** Los polimixinos interrumpen la estructura de la membrana plasmática bacteriana, causando la lisis celular (Hutchings et al., 2019).

- **Inhibidores de vías metabólicas esenciales:** Los sulfamídicos y trimetoprim actúan inhibiendo enzimas clave en la síntesis de folatos, una vía metabólica esencial para la síntesis de nucleótidos y, por ende, para la replicación del ADN bacteriano (Hutchings et al., 2019).

4.1.2. Clasificación

Los antibióticos también se pueden clasificar según su espectro de actividad en:

Espectro amplio: Son efectivos contra una amplia variedad de bacterias, tanto Gram-positivas como Gram-negativas. Ejemplos incluyen las tetraciclinas y las quinolonas (Kapoor et al., 2017).

Espectro estrecho: Actúan contra un grupo limitado de bacterias. La penicilina G, por ejemplo, es más efectiva contra bacterias Gram-positivas (Kapoor et al., 2017).

4.1.3. Desarrollo y descubrimiento de nuevos antibióticos

Dado el aumento de la resistencia a antibióticos en bacterias, el desarrollo de nuevos antibióticos y el descubrimiento de compuestos con mecanismos de acción novedosos son esenciales para mantenerse un paso adelante de las bacterias patógenas. Sin embargo, el desarrollo de nuevos antibióticos es un proceso complejo y costoso, lo que ha llevado a una disminución en el número de nuevos antibióticos introducidos en el mercado en las últimas décadas (Cuddy, 1997).

4.1.4. Antibióticos populares y sus mecanismos de acción

- **Penicilinas (β-lactámicos):** Las penicilinas, descubiertas por Alexander Fleming en 1928, son una clase de antibióticos β-lactámicos que actúan inhibiendo la síntesis de la pared celular bacteriana. Lo hacen al unirse a las proteínas de unión a penicilina (PBPs), inhibiendo la formación de enlaces cruzados en el peptidoglicano, lo que resulta en la lisis y muerte de la bacteria. Las penicilinas son especialmente efectivas contra bacterias Gram-positivas (Dever, 1991).

- **Cefalosporinas (β-lactámicos):** Similar a las penicilinas, las cefalosporinas inhiben la síntesis de la pared celular bacteriana y tienen un espectro de actividad más amplio, incluyendo ciertas bacterias Gram-negativas. Son una opción común para tratar infecciones como la neumonía, infecciones de la piel y del tracto urinario (Dever, 1991).

- **Tetraciclinas:** Estos antibióticos se unen a la subunidad 30S del ribosoma bacteriano, impidiendo la adición de aminoácidos a la cadena peptídica emergente, lo que efectivamente detiene la síntesis de proteínas. Las tetraciclinas son efectivas contra una amplia variedad de bacterias y se utilizan para tratar infecciones como el acné, la clamidia y la enfermedad de Lyme (Dever, 1991).

- **Macrólidos:** Los macrólidos, como la eritromicina, se unen a la subunidad 50S del ribosoma bacteriano, inhibiendo la translocación peptídica, un paso en la síntesis de proteínas. Son útiles contra bacterias Gram-positivas y algunas Gram-negativas, y se prescriben comúnmente para infecciones respiratorias y de tejidos blandos (Thrum, 1977).

- **Quinolonas:** Las quinolonas, incluyendo la ciprofloxacina, inhiben las topoisomerasas bacterianas, enzimas necesarias para el desenrollamiento y enrollamiento del ADN, esenciales para la replicación y reparación del ADN. Son efectivas contra una amplia gama de bacterias Gram-positivas y Gram-negativas y se usan en el tratamiento de infecciones del tracto urinario y respiratorio, entre otras (Thrum, 1977).

- **Aminoglucósidos:** Estos antibióticos, incluyendo la gentamicina, se unen a la subunidad 30S del ribosoma bacteriano, causando una lectura errónea del ARNm y la incorporación incorrecta de aminoácidos, lo que lleva a la producción de proteínas defectuosas y la muerte bacteriana. Son efectivos principalmente contra bacterias Gram-negativas (Thrum, 1977).

- **Sulfonamidas:** Actúan como antagonistas competitivos de la para-aminobenzoico (PABA), un precursor vital para la síntesis de ácido fólico en bacterias. Al inhibir la

producción de ácido fólico, las sulfonamidas detienen la producción de nucleótidos y, por ende, la síntesis de ADN. Son útiles en el tratamiento de infecciones del tracto urinario y neumonía por *Pneumocystis jirovecii* (Thrum, 1977).

4.2. Resistencia a los antibióticos

La resistencia a los antibióticos es un fenómeno por el cual los microorganismos desarrollan la capacidad de sobrevivir a la exposición a antibióticos, medicamentos que originalmente eran efectivos para tratar infecciones causadas por estos organismos, ya no lo son. Este proceso representa una amenaza para la salud pública global, ya que puede conducir a infecciones que son más difíciles de tratar, aumentar la duración de las enfermedades, elevar los costos de atención médica y aumentar la mortalidad (Wright, 2010).

Los microorganismos adquieren resistencia a través de varios mecanismos, incluyendo mutaciones genéticas y la adquisición de genes de resistencia a través de la transferencia horizontal de genes. Estos mecanismos pueden alterar el sitio de acción del antibiótico, reducir la permeabilidad del microorganismo al antibiótico, modificar el antibiótico a través de enzimas o bombear el antibiótico fuera de la célula (Blair et al., 2014)

Por otro lado, la resistencia a los antibióticos en bacterias, específicamente, es un fenómeno que implica diversos mecanismos a través de los cuales las bacterias pueden evadir la acción de estos medicamentos. Cuando las bacterias que ya son difíciles de tratar adquieren una combinación adecuada de mecanismos de resistencia, todos los antibióticos o antifúngicos pueden volverse ineficaces, lo que resulta en infecciones intratables. Los mecanismos más comunes de resistencia incluyen:

- **Modificación del objetivo del antibiótico:** Cambios en las proteínas objetivo de los antibióticos pueden hacer que estos medicamentos sean menos efectivos o ineficaces. Este mecanismo es común en la resistencia a fluoroquinolonas y rifampicina (Blair et al., 2014)

- **Disminución de la permeabilidad de la membrana bacteriana:** Alteraciones en la estructura de la membrana celular pueden reducir la entrada del antibiótico a la célula bacteriana, disminuyendo su eficacia. Este mecanismo es notable en la resistencia a tetraciclinas y sulfonamidas (Thrum, 1977).

- **Bombeo activo del antibiótico fuera de la célula:** Las bacterias pueden utilizar bombas de expulsión para remover activamente los antibióticos de su interior, redu-

ciendo así la concentración del medicamento a niveles no letales. Este mecanismo es común en la resistencia a tetraciclinas y fluoroquinolonas (Munita & Arias, 2016).

- **Producción de enzimas que inactivan el antibiótico:** Algunas bacterias producen enzimas capaces de degradar antibióticos o modificarlos químicamente, haciéndolos ineficaces. La producción de β-lactamasas, que hidrolizan el anillo β-lactámico de penicilinas y cefalosporinas, es un ejemplo destacado de este mecanismo (Munita & Arias, 2016).

La resistencia a los antibióticos puede ser intrínseca, donde las bacterias naturalmente poseen características que las hacen resistentes a ciertos antibióticos, o adquirida, donde las bacterias obtienen nuevos genes de resistencia a través de la transferencia horizontal de genes. Esta transferencia puede ocurrir por conjugación, transformación o transducción, permitiendo la diseminación rápida de genes de resistencia entre diferentes especies bacterianas (Munita & Arias, 2016)

4.3. *Ceratitis capitata*

Ceratitis capitata, comúnmente conocida como la mosca de la fruta del Mediterráneo, es una de las plagas agrícolas más destructivas a nivel mundial. Este insecto, originario de África subsahariana, ha extendido su presencia a muchas áreas del mundo, incluidas Europa, América del Sur y partes de Norteamérica, donde las condiciones climáticas son favorables para su desarrollo y reproducción (Sciarretta et al., 2018).

4.3.1. Biología y ciclo de vida

La biología de *Ceratitis capitata* es particularmente adaptativa, lo que contribuye a su éxito como especie invasora. El ciclo de vida completo desde huevo hasta adulto puede variar de 30 a más de 100 días, dependiendo de las condiciones ambientales como la temperatura y la humedad. Los huevos son depositados bajo la piel de frutas maduras o en proceso de maduración, donde las larvas encuentran un ambiente nutritivo para desarrollarse. Tras completar tres estadios larvarios, las larvas se pupan en el suelo. Posteriormente, emergen adultos, completando el ciclo de vida (Tabilio et al., 2013).

4.3.2. Impacto en la agricultura

El impacto de *Ceratitis capitata* en la agricultura es significativo debido a su amplio rango de hospederos, que incluye más de 250 especies de frutas y vegetales. Las infestaciones causan daños directos a los cultivos al consumir la pulpa de las frutas, lo que no solo reduce la producción agrícola sino también afecta la calidad de los frutos, disminuyendo su valor de mercado y su idoneidad para la exportación. Además, su presencia en regiones no nativas ha llevado a la implementación de estrictas medidas de cuarentena y control, incrementando los costos de producción y manejo para los agricultores (Al-Behadili et al., 2020).

4.3.3. Estrategias de control

El manejo de *Ceratitis capitata* ha evolucionado para incluir una combinación de enfoques que integran métodos químicos, biológicos y culturales. Las prácticas comunes incluyen la eliminación de frutos infestados y la rotación de cultivos, estrategias diseñadas para disminuir la viabilidad del hábitat de la plaga. Estas medidas se complementan con el uso de trampas y atrayentes, así como insecticidas, que no solo ayudan en el monitoreo sino también en el control efectivo de la población (Navarro-Llopis et al., 2013). Además, métodos más sofisticados y respetuosos con el medio ambiente, como la técnica del insecto estéril (SIT), desempeñan un papel crucial. Esta técnica implica la liberación de machos estériles para interferir en la reproducción y, en consecuencia, reducir significativamente las poblaciones de la mosca (Pérez-Staples et al., 2021).

4.3.4. Técnica del insecto estéril (SIT)

Una de las estrategias más innovadoras y sostenibles para el control de *Ceratitis capitata* es la Técnica del Insecto Estéril (SIT, por sus siglas en inglés). Esta técnica implica la cría masiva de moscas de la fruta en instalaciones especializadas, seguida de la esterilización de machos mediante métodos como la radiación. Los machos estériles son luego liberados en áreas afectadas, donde compiten con los machos salvajes por aparearse con las hembras. Dado que los apareamientos con machos estériles no producen descendencia, la técnica reduce efectivamente la población de la plaga en las generaciones subsiguientes (Benedict, 2021).

La SIT es particularmente valiosa porque es una solución amigable con el medio ambiente, que no depende del uso de insecticidas químicos, y por lo tanto, evita los problemas asociados con la resistencia a pesticidas y el impacto negativo en organismos no objetivo (Bourtzis & Vreysen, 2021).

La implementación exitosa de la Técnica del Insecto Estéril (SIT) demanda una planifi-

cación meticulosa para establecer las condiciones óptimas de cría, esterilización y liberación de los insectos en cada área específica, asegurando así resultados efectivos. Además, es crucial la cooperación entre agricultores, investigadores y autoridades locales para coordinar adecuadamente las liberaciones y evaluar su eficacia.

A pesar de sus numerosas ventajas, esta técnica enfrenta varios desafíos significativos, incluyendo los costos asociados con la cría y esterilización de los insectos y la necesidad de realizar liberaciones continuas para mantener la plaga bajo control. Sin embargo, un desafío aún más importante es el de garantizar que las moscas liberadas no contribuyan a la propagación de patógenos ni porten organismos peligrosos, como bacterias con genes de resistencia a los antibióticos (Marec & Vreysen, 2019).

4.4. Conceptos y herramientas fundamentales

Para facilitar la comprensión de este trabajo, es esencial familiarizarse con ciertos términos y herramientas. Esta sección proporciona tanto definiciones como explicaciones de los conceptos clave y las herramientas esenciales empleadas en el estudio. Al comprenderlos, se pueden interpretar mejor los hallazgos y seguir la discusión subsecuente con mayor claridad.

4.4.1. *Pipelines* en bioinformática

Un *pipeline* en bioinformática se refiere a una serie de pasos secuenciales y automatizados, diseñados para procesar datos biológicos complejos. Puede incluir la adquisición de datos, el preprocesamiento, el análisis estadístico y la interpretación de los resultados. Los *pipelines* son esenciales para manejar eficientemente las grandes cantidades de datos generados por las tecnologías modernas de secuenciación de ADN y otras técnicas experimentales en biología molecular.

El desarrollo de un *pipeline* bioinformático implica la integración de diversos *software* y herramientas analíticas especializadas, como alineación de secuencias, anotación de genes y detección de variantes. Los sistemas automatizados ayudan a estandarizar los procesos de análisis, mejorando la reproducibilidad y precisión de los estudios científicos. Además, los *pipelines* permiten a los investigadores procesar y analizar rápidamente volúmenes de datos que serían inmanejables de forma manual (SoRelle et al., 2020).

4.4.2. Formato FASTQ

El formato FASTQ es ampliamente utilizado en bioinformática para almacenar secuencias de nucleótidos junto con sus correspondientes calidades de secuenciación. Cada registro en un archivo FASTQ contiene una secuencia única y su calidad asociada, lo que permite evaluar la confiabilidad de cada base secuenciada. Este formato es esencial para las aplicaciones de secuenciación de nueva generación (NGS), donde la precisión de cada lectura puede variar significativamente.

Un archivo FASTQ consta de cuatro líneas por secuencia: una línea de encabezado, que comienza con un símbolo '@', seguido por un identificador y una descripción opcional, una línea de secuencia que muestra las bases de nucleótidos, una línea de separador que comienza con un símbolo '+', y una línea de calidad que codifica la confianza en cada base de la secuencia. La calidad de la secuencia se codifica utilizando un sistema de puntuación ASCII que ayuda a identificar posibles errores en el proceso de secuenciación y a tomar decisiones informadas sobre el uso de los datos en análisis posteriores (Cock et al., 2010).

4.4.3. RefSeq (NCBI *reference sequence database*)

La base de datos de secuencias de referencia del National Center for Biotechnology Information (NCBI), más conocida como RefSeq, es una colección curada de secuencias de ADN, ARN y proteínas que sirve como un marco de referencia para la identificación genómica, genética y funcional. RefSeq proporciona una base de datos accesible y de alta calidad que incluye secuencias de genomas, genes y sus productos, así como todas las secuencias anotadas exhaustivamente.

Las secuencias de RefSeq son utilizadas ampliamente en diversas aplicaciones, incluyendo la anotación de nuevas secuencias genómicas, estudios comparativos de genomas, y análisis filogenéticos. Cada entrada en RefSeq está vinculada a información adicional sobre la función del gen, la estructura de la proteína, y la regulación, así como a enlaces a otras bases de datos relevantes de NCBI, lo que permite acceder a un panorama completo de la información genética (Pruitt et al., 2012).

4.4.4. Python

Python es un lenguaje de programación de alto nivel, interpretado y de propósito general, que se ha ganado la preferencia mundial debido a su sintaxis clara y legible. Diseñado por Guido van Rossum y lanzado por primera vez en 1991, Python facilita la programación eficiente y efectiva en diversos dominios de aplicación, desde el desarrollo web hasta la

ciencia de datos y la inteligencia artificial.

Uno de los principales atractivos de Python es su amplia biblioteca estándar, junto con un extenso ecosistema de paquetes de terceros. Estas bibliotecas ofrecen herramientas poderosas y especializadas para realizar una variedad de tareas, incluido el análisis de datos, la visualización, el procesamiento de datos en tiempo real y la automatización.

Además, Python es especialmente apreciado en la comunidad científica y académica por su simplicidad y flexibilidad. Permite a los investigadores desarrollar y prototipar rápidamente aplicaciones y herramientas de análisis, lo cual es crucial para proyectos que requieren análisis intensivo de datos y modelado estadístico (Lopez et al., 2013).

4.4.5. Paquetes de Python

En el ámbito de la programación, especialmente en Python, un paquete se refiere a un directorio que contiene archivos Python y un archivo `__init__.py`. Este archivo particular hace que Python trate los directorios como contenedores de módulos, lo que a su vez permite la organización del código de manera modular y reutilizable.

Los paquetes en Python facilitan la estructuración del código de manera que diferentes módulos que realizan funciones relacionadas pueden agruparse bajo un mismo paquete. Esto no solo mejora la legibilidad y mantenimiento del código, sino que también permite compartir y utilizar el código de manera más eficiente (X. Chen & Liu, 2022).

4.4.6. Anaconda

Anaconda es una distribución de los lenguajes de programación Python y R, diseñada especialmente para su uso en ciencia de datos, computación científica, análisis predictivo y procesamiento de datos a gran escala. Facilita la gestión de paquetes y entornos, permitiendo la instalación eficiente y la administración de versiones sin conflictos, lo cual es esencial para mantener la consistencia y reproducibilidad en proyectos complejos.

Además de simplificar la configuración de herramientas científicas y analíticas, Anaconda incluye el Anaconda Navigator, una interfaz gráfica de usuario que permite a los usuarios gestionar entornos virtuales y lanzar aplicaciones como Jupyter Notebook, Spyder y RStudio sin necesidad de comandos de consola. Esta capacidad hace de Anaconda una solución integral y accesible para investigadores y desarrolladores de todos los niveles técnicos (Adamowicz et al., 2022).

4.4.7. Conda

Conda es un gestor de paquetes y entornos diseñado para simplificar la instalación y gestión de software de los lenguajes de programación Python y R. Es una herramienta esencial dentro de la distribución Anaconda, pero también se puede utilizar de forma independiente para manejar bibliotecas y dependencias en diversos entornos científicos y de desarrollo.

El principal beneficio de Conda es su capacidad para crear entornos aislados que pueden contener diferentes versiones de paquetes y dependencias, lo que permite a los desarrolladores y científicos trabajar en proyectos múltiples sin riesgo de conflictos entre bibliotecas. Conda facilita la reproducibilidad científica al permitir que los entornos sean exactamente replicados en diferentes máquinas, lo que es crucial para colaboraciones científicas y validaciones de resultados. Además, su integración con Anaconda Navigator permite una gestión visual y más intuitiva de estos entornos y paquetes, acercando estas capacidades a un público más amplio que puede no estar familiarizado con la línea de comandos (Adamowicz et al., 2022).

4.4.8. Bioconda

Bioconda es un canal dentro del gestor de paquetes Conda que se especializa en paquetes de software para bioinformática. Esta canal ofrece un repositorio de fácil manejo para instalar más de 7,000 herramientas bioinformáticas, facilitando su implementación en entornos de investigación y desarrollo.

El objetivo principal de Bioconda es proporcionar una manera eficiente y sistemática de gestionar programas y bibliotecas bioinformáticas, garantizando la compatibilidad y actualización continua de las herramientas. Es particularmente valorado por su enfoque comunitario, que permite a los investigadores contribuir al desarrollo y mantenimiento de paquetes, asegurando que el software esté al día con los últimos avances científicos(Grüning et al., 2018).

En este trabajo, se emplearon varios paquetes de Bioconda que permitieron desarrollar el proceso de la detección de genes de resistencia a antibióticos en bacterias. Estos fueron:

- **Fastp:** es una herramienta versátil para el preprocesamiento de datos de secuenciación de alto rendimiento, especialmente diseñada para leer datos de FASTQ. Esta herramienta proporciona funcionalidades de filtrado de calidad, recorte de adaptadores, filtrado de lecturas contaminadas por polímeros y corrección de errores, entre otras. La eficacia de *fastp* radica en su capacidad para procesar datos de manera rápida y

con mínima pérdida de información útil, lo cual es crucial para garantizar la calidad y fiabilidad de los análisis posteriores.

Además, *fastp* es capaz de generar informes gráficos que permiten a los usuarios visualizar la calidad de los datos antes y después del preprocesamiento, facilitando así la evaluación y ajustes necesarios para optimizar el proceso de secuenciación(S. Chen et al., 2018).

- **SPAdes:** es una herramienta de ensamblaje de genomas diseñada para ensamblajes de nueva generación, incluyendo aquellos de tipo *de novo*. Es especialmente útil para ensamblar genomas de bacterias y pequeños eucariotas a partir de datos de secuenciación de próxima generación. *SPAdes* es conocido por su versatilidad y precisión, incorporando varios modos de ensamblaje optimizados para diferentes configuraciones y tipos de datos.

  Además de su funcionalidad principal de ensamblaje, *SPAdes* viene equipado con módulos útiles como error correction y ensamblaje de meta-genomas, que son esenciales para proyectos que manejan datos complejos y variados (Ishengoma & Rhode, 2022).

- **Kraken2:** es una herramienta de clasificación y asignación taxonómica rápida y altamente precisa para análisis de metagenómica. Utiliza un enfoque de k-mer para identificar qué organismos están presentes dentro de un conjunto de datos y cuál es su abundancia relativa. *Kraken2* se basa en una base de datos de secuencias de referencia para comparar k-mers encontrados en los datos de secuenciación, permitiendo una clasificación rápida y precisa a nivel de especie.

  La principal ventaja de *Kraken2* reside en su eficiencia y velocidad, lo que permite procesar gigabytes de datos en minutos, un aspecto crucial para el análisis de grandes volúmenes de datos metagenómicos. Además, *Kraken2* incluye opciones para trabajar con bases de datos personalizadas, lo que lo hace extremadamente flexible y adecuado para una variedad de proyectos metagenómicos, incluido el análisis de bacterias en la dieta de *Ceratitis capitata* (Wood et al., 2019).

- **Prokka:** es una herramienta de anotación de genomas bacterianos que automatiza el proceso de predicción de genes y la asignación de funciones genéticas en genomas bacterianos y de arqueas. Utiliza una serie de programas bioinformáticos bien establecidos, como *Prodigal* para la predicción de genes, *RNAmmer* para la identificación de ARN ribosomal, y *HMMER* para la búsqueda de homologías de proteínas, entre otros.

  Esta herramienta ofrece una solución integral para la anotación de genomas bacterianos, generando resultados detallados y fácilmente interpretables. *Prokka* es especialmente útil para proyectos de secuenciación de genomas, permitiendo a los investiga-

dores obtener información sobre la estructura y función de los genes de manera rápida y precisa (Seemann, 2014).

- **RGI:** es una herramienta desarrollada por la *Comprehensive Antibiotic Resistance Database* (CARD) para identificar y caracterizar genes de resistencia a antibióticos en datos genómicos. RGI utiliza modelos bioinformáticos para detectar la presencia de genes conocidos de resistencia a antibióticos y potenciales variantes no catalogadas mediante la comparación con la base de datos de CARD. Esta herramienta es esencial para estudios epidemiológicos y de salud pública donde la detección rápida y precisa de la resistencia es crítica.

  RGI es particularmente valioso por su capacidad para integrar y analizar secuencias genómicas complejas, proporcionando un perfil detallado de la resistencia antibiótica en muestras bacterianas. En este trabajo, RGI fue fundamental para determinar los perfiles de resistencia en las bacterias estudiadas, permitiendo una evaluación eficiente de las amenazas de resistencia y facilitando el desarrollo de estrategias de intervención adecuadas (Jia et al., 2017).

4.4.9. Bases de datos científicas

Las bases de datos científicas son plataformas o repositorios electrónicos que almacenan, organizan y proporcionan acceso no solo a literatura académica y científica, sino también a una amplia gama de datos biológicos, incluidos datos de secuenciación genómica, datos proteómicos y otros tipos de información biomolecular. Estas bases de datos varían en su enfoque, cobertura y características, adaptándose a las necesidades específicas de diferentes campos de estudio.

Entre las bases de datos científicas que se usaron para este trabajo se encuentran:

- **PubMed:** es un recurso gratuito que facilita la búsqueda y recuperación de literatura biomédica y de ciencias de la vida con el objetivo de mejorar la salud, tanto a nivel global como personal.

  La base de datos de PubMed contiene más de 36 millones de citas y resúmenes de literatura biomédica. No incluye artículos de revistas en su totalidad; sin embargo, a menudo se proporcionan enlaces al texto completo cuando están disponibles desde otras fuentes, como el sitio web del editor o PubMed Central (PMC)(Ossom Williamson & Minter, 2019).

- **Google Scholar:** es una herramienta de búsqueda gratuita que indexa literatura académica de diversas disciplinas y formatos. A diferencia de otras bases de datos,

Google Scholar rastrea y muestra resultados de diversas fuentes, incluyendo revistas, tesis, libros y conferencias (Vine, 2006).

- **BioProject de NCBI:** el BioProject del National Center for Biotechnology Information (NCBI) es una base de datos que proporciona un marco organizativo para agrupar todos los datos biológicos relacionados con un proyecto de investigación determinado. Un BioProject puede incluir desde estudios de secuenciación genómica hasta proyectos de investigación ambiental, y ofrece una vista integral de los datos experimentales, las publicaciones relacionadas y los enlaces a datos almacenados en otras bases de datos de NCBI como GenBank y Sequence Read Archive (SRA).

  Cada BioProject está asignado a un identificador único (ID) que permite a los usuarios y a los investigadores acceder de manera eficiente a la colección de datos vinculados al proyecto (Barrett et al., 2012).

- **SRA de NCBI:** el *Sequence Read Archive* (SRA) del NCBI es una de las mayores bases de datos de secuencias públicas del mundo. Esta base de datos almacena datos de secuencias en bruto y alineaciones, proporcionando acceso a series de datos derivados de estudios de secuenciación de alto rendimiento, incluidos aquellos que utilizan tecnologías como Illumina, 454, Ion Torrent, y PacBio.

  El SRA no solo sirve como un repositorio de datos sino también como una herramienta vital para la investigación biomédica y genómica, permitiendo a los investigadores acceder a un vasto conjunto de datos para análisis comparativos, estudios de asociación genética y exploraciones de diversidad genética. Los datos almacenados en SRA pueden ser accesados mediante herramientas de búsqueda y descarga del NCBI, facilitando la integración de estos datos en diversos proyectos de investigación genómica (Sayers et al., 2022).

4.4.10. Base de datos de genes de resistencia a antibióticos

CARD: Comprehensive Antibiotic Resistance Database

La *Comprehensive Antibiotic Resistance Database* (CARD) es una base de datos especializada que se centra en el acopio, organización y análisis de información genética y bioquímica relacionada con la resistencia a los antibióticos. CARD es una herramienta invaluable para investigadores que estudian los mecanismos de resistencia antibiótica y buscan desarrollar estrategias para mitigar esta amenaza creciente.

Esta base de datos no solo cataloga genes de resistencia a los antibióticos y sus mutaciones asociadas, sino que también proporciona herramientas analíticas como el sistema RGI

(Resistance Gene Identifier), que permite a los usuarios identificar genes de resistencia en secuencias genómicas de bacterias. El contenido de CARD se actualiza regularmente para reflejar los avances más recientes en la investigación de resistencia a los antibióticos, lo que la convierte en una fuente esencial para la investigación contemporánea en microbiología y farmacología (McArthur et al., 2013).

CAPÍTULO 5

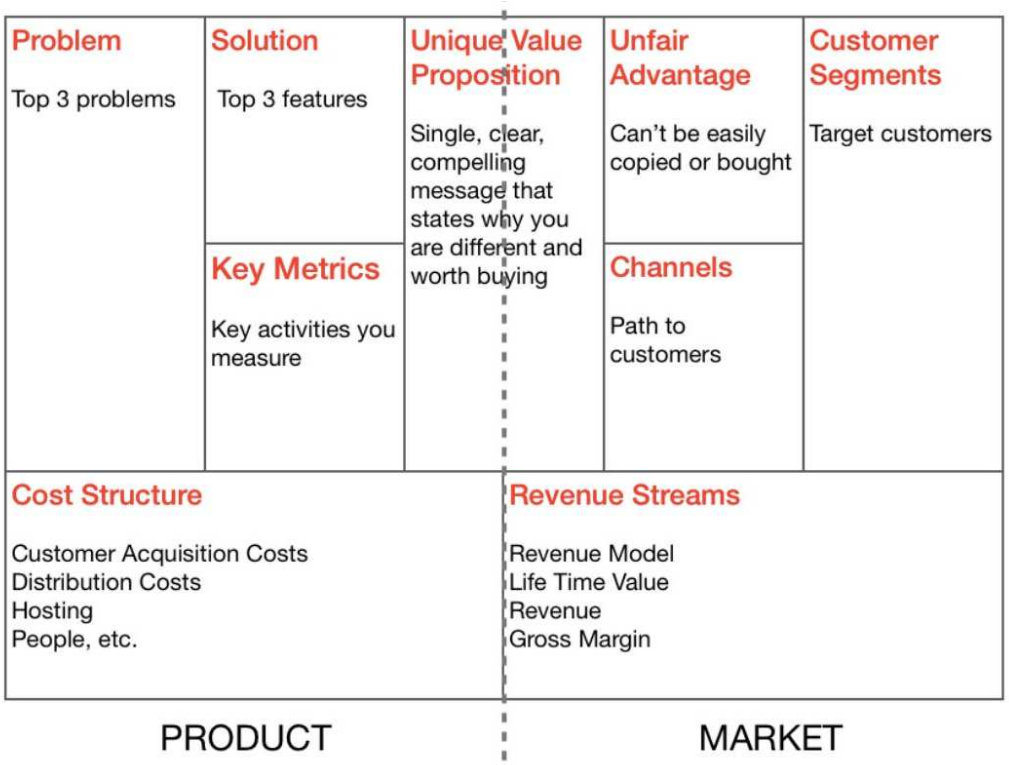
Metodología

5.1. Evaluación de viabilidad

Para determinar la viabilidad y el impacto potencial del *pipeline* desarrollado, se empleó una adaptación del modelo Lean Canvas. El Lean Canvas, creado originalmente por Ash Maurya, es un enfoque ágil que ayuda a definir y discutir modelos de negocio de manera estructurada y concisa. Este modelo incluye segmentos como el problema, soluciones propuestas, ventajas únicas y la propuesta de valor, los cuales son cruciales para la planeación estratégica (Chokshi & Mann, 2018).

Aunque tradicionalmente se utiliza en el ámbito empresarial para *startups*, su estructura es lo suficientemente flexible para adaptarse al contexto de un proyecto de investigación científica. Esta metodología permitió abordar de manera sistemática los principales desafíos, soluciones, y beneficios del proyecto.

Figura 1. Lean Canvas original



Adaptada de Mullen (2016).

A continuación, se muestran los segmentos más relevantes del Lean Canvas adaptados a este proyecto.

5.1.1. Problema

El principal problema que aborda este proyecto es la prevención de la propagación de bacterias portadoras de genes de resistencia a antibióticos, inicialmente aquellas que podrían encontrarse en el microbioma de *Ceratitis capitata* cuando se utilizan en la técnica del insecto estéril.

5.1.2. Solución

La solución propuesta a través de este proyecto es ofrecer una herramienta que permita analizar muestras de ADN y detectar la presencia de genes de resistencia a antibióticos de manera automatizada y eficiente.

5.1.3. Propuesta de valor único

La propuesta de valor único de este proyecto radica en su capacidad para integrar y automatizar el proceso completo de detección de genes de resistencia a antibióticos en un solo sistema. A diferencia de otras herramientas o métodos que requieren múltiples pasos manuales y experiencia técnica, este *pipeline* simplifica el análisis a través de una interfaz de usuario amigable y procesos automatizados. Por otra parte, hace que la tecnología de secuenciación y análisis genético sea más accesible a una variedad más amplia de investigadores, contribuyendo así al control eficaz de la propagación de resistencias en ambientes controlados y naturales.

5.1.4. Segmento de usuarios

El segmento de usuarios para este proyecto está dentro de los ámbitos académico y de investigación, especialmente aquellos que trabajan en el campo de la microbiología y genética. Incluye investigadores y científicos que estudian el impacto de la resistencia a antibióticos en diferentes ecosistemas, así como aquellos involucrados en programas de control biológico. Adicionalmente, debido a su diseño intuitivo, el programa también apunta a educadores y estudiantes que buscan herramientas prácticas para experiencias de aprendizaje en genómica y bioinformática.

5.1.5. Canales

Los canales de acceso y distribución deben permitir a los usuarios descargar y comenzar a utilizar el programa con facilidad. Primero, se planea ofrecer una versión compilada del programa como un archivo ejecutable (.exe), lo que permitiría a los usuarios de sistemas Windows instalarlo directamente sin necesidad de configuraciones adicionales. Además, para asegurar la accesibilidad y la colaboración continua, el código fuente del *pipeline* estará disponible en un repositorio público en GitHub.

Cuadro 1. Resumen de los segmentos del Lean Canvas aplicados al proyecto.

Segmentos	**Descripción**
Problema	Propagación de bacterias con genes de resistencia a antibióticos
Solución	Programa automatizado para detección eficiente de genes de resistencia
Propuesta de Valor Único	Conveniencia, accesibilidad, y automatización del proceso de detección
Segmento de Usuarios	Investigadores, científicos, educadores y estudiantes en microbiología y genética
Canales	Descarga directa (.exe) y acceso al código fuente a través de GitHub

Elaboración propia.

5.2. Recopilación de datos

Esta sección describe las fuentes y tipos de datos utilizados en la investigación para validar y probar el *pipeline* desarrollado. Los datos se dividieron en dos categorías principales: datos experimentales obtenidos de la dieta de *Ceratitis capitata* y datos de prueba obtenidos de bases de datos públicas para validar la eficacia del programa.

5.2.1. Datos de dieta de *Ceratitis capitata*

Los datos utilizados para la prueba aplicada del *pipeline* fueron proporcionados por la Licenciada en Bioquímica y Microbiología Isabella García Caffaro, colaboradora de la Universidad del Valle de Guatemala, quien ha trabajado en proyectos de investigación y desarrollo con Moscamed y otros entes relacionados con la producción de moscas de la fruta estériles mediante técnicas de biología molecular. Los datos consistieron en archivos FASTQ, derivados de muestras bacterianas recolectadas de la dieta de *Ceratitis capitata*. Estas muestras fueron esenciales para probar la capacidad del *pipeline* en la identificación de genes de resistencia a antibióticos en un contexto real y aplicado.

5.2.2. Datos de prueba de bases de datos públicas

Para la validación del *pipeline*, se emplearon datos de secuenciación obtenidos de la base de datos *Sequence Read Archive* (SRA) de NCBI. Se seleccionaron cuatro BioProjects que cumplían con criterios específicos para garantizar la relevancia y la calidad de los datos. Los criterios incluyeron (i) que las muestras provenían de bacterias asociadas, al menos en parte,

a la dieta de *Ceratitis capitata*; (ii) que las muestras habían sido secuenciadas utilizando la tecnología Illumina MiSeq; y (iii) que los proyectos presentaran un ensamblaje de genoma disponible como RefSeq.

Los BioProjects seleccionados (PRJNA307517, PRJNA1076266 y PRJNA392824) cumplieron con todos estos criterios y proporcionaron un conjunto de 53 SRA que contenían secuencias de diferentes bacterias con genes de resistencia conocidos. Por otro lado, se utilizó el BioProject PRJNA279657 como control negativo. Este proyecto incluía secuencias de *Klebsiella pneumoniae* que no reportaban genes de resistencia, lo cual fue crucial para verificar la capacidad del *pipeline* de detectar correctamente la ausencia de genes de resistencia en las muestras.

La presencia de un ensamblaje del genoma como RefSeq en los BioProjects fue esencial para permitir una comparación directa y precisa entre los genes de resistencia identificados por el *pipeline* y los genes documentados en la secuencia de referencia. Esto facilitó la confirmación de verdaderos y falsos positivos, así como de falsos negativos.

5.3. Desarrollo del *Pipeline*

El desarrollo del *pipeline* implicó una serie de etapas para asegurar tanto la funcionalidad como la usabilidad del programa. Incluyó la planeación y selección de tecnologías adecuadas, la integración de herramientas bioinformáticas y la creación de un diseño interactivo.

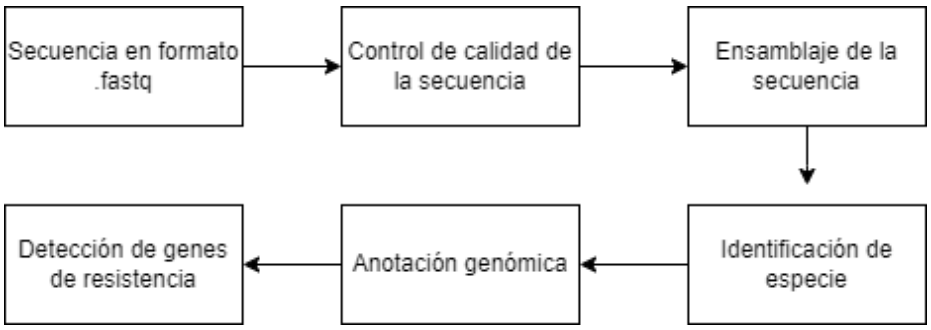
5.3.1. Planeación de los pasos del *pipeline*

Para el desarrollo de un *pipeline* intuitivo y eficiente, se requiere una estructura clara de los pasos a seguir para garantizar resultados precisos. Por esta razón, la primera etapa en el diseño del *pipeline* incluyó la definición de una secuencia lógica de pasos, estableciendo una ruta clara para el desarrollo del programa. Esta planificación se realizó para evitar redundancias y asegurar que cada etapa aportara un valor significativo al análisis final.

El primer paso planteado fue un control de calidad de las secuencias de entrada. Este paso es crucial porque las secuencias de *FASTQ* a menudo contienen segmentos de baja calidad que pueden comprometer la precisión de los análisis subsiguientes.

Una vez que las secuencias están limpias, el siguiente paso lógico sería un ensamblaje. Este proceso se encargaría de reconstruir la secuencia original, que debido a la naturaleza del proceso de secuenciación, resulta en fragmentos cortos de ADN. El ensamblaje proporciona una visión más completa y utilizable del material genético, que será la base para procesos

Figura 2. Diagrama de flujo del programa en la planeación.



Elaboración propia.

posteriores.

Tras el ensamblaje, se realizaría la identificación de la especie utilizando los contigs ensamblados. Este paso es crucial para enriquecer la comprensión del genoma ensamblado. Identificar correctamente la especie es esencial, ya que guía las fases subsiguientes de anotación y análisis funcional de los genes.

Con el genoma ya ensamblado y la especie identificada, el siguiente paso sería identificar las funciones de los genes y otras carácterísticas genómicas. Esto se haría mediante una anotación genómica. Esta información es indispensable para entender el contexto biológico de las secuencias y prepararlas para el paso final del análisis.

El último paso involucraría la detección de genes de resistencia a antibióticos. Utilizando las anotaciones, se compararían las secuencias con bases de datos de genes de resistencia conocidos, como la base de datos CARD. Este paso determina la presencia o ausencia de genes de resistencia en la muestra analizada, culminando el proceso de análisis.

Una vez establecida esta secuencia de operaciones, el siguiente paso fue seleccionar las herramientas específicas que mejor se adaptaran a cada una de estas tareas.

5.3.2. Planeación de las herramientas usadas

La selección de las herramientas a usar en cada etapa del *pipeline* fue un paso crucial para su desarrollo. Fue esencial elegir herramientas no solo capaces de realizar los análisis requeridos, sino que además ofrecieran una experiencia *user-friendly* y aseguraran una ejecución eficiente. Se evaluaron diversas opciones para cada herramienta, dando prioridad a aquellas que contaban con amplia documentación, lo cual facilitaría una mayor flexibilidad en su uso. También se consideraron criterios como la prevalencia de uso en la comunidad de bioinformática y la eficiencia en términos de velocidad de procesamiento.

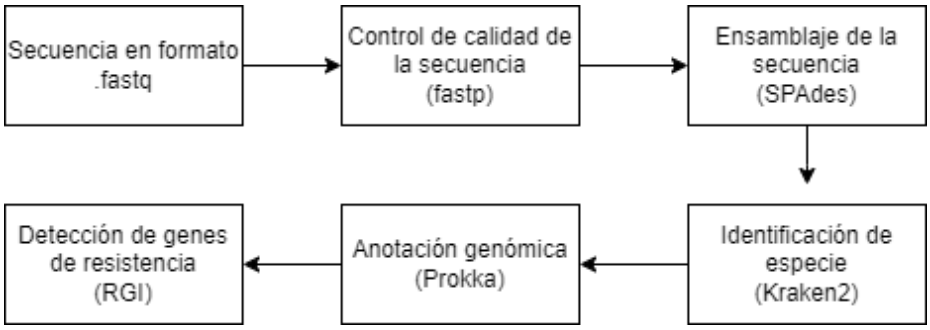
Primero, se seleccionó un lenguaje de programación adecuado. Python fue la elección natural, dada su amplia aceptación en la comunidad científica y bioinformática, lo que garantiza una vasta documentación y una extensa biblioteca de paquetes útiles para el análisis de datos, especialmente biológicos(Fourment & Gillings, 2008).

Dada la variedad de paquetes y librerías de Python necesarios para el proyecto, se optó por utilizar Conda como gestor de paquetes. Conda es un gestor de paquetes y entornos que facilitó la instalación y el manejo del software requerido para el análisis bioinformático. Este sistema permitió la creación de entornos aislados que albergaban todas las herramientas necesarias, garantizando así la compatibilidad y previniendo conflictos entre paquetes.

Para los pasos específicos del *pipeline*, como el control de calidad, el ensamblaje de secuencias, la identifiación de especie, la anotación y la detección de genes de resistencia, se seleccionaron herramientas específicas de Bioconda que son reconocidas por su rendimiento.

Se escogió `fastp` para el control de calidad por su rapidez y eficiencia en el filtrado y limpieza de datos de secuenciación. `SPAdes` fue la herramienta elegida para el ensamblaje debido a su capacidad para manejar diversos tipos de secuenciación y sus algoritmos optimizados para ensamblajes complejos. Para la identificación de especie después del ensamblaje, se utilizó `Kraken2` por su alta precisión y rapidez en la clasificación taxonómica. `Prokka` se utilizó para la anotación rápida de características genómicas, dada su capacidad para generar anotaciones de alta calidad en tiempo reducido. Finalmente, `RGI` fue seleccionado para la identificación de genes de resistencia mediante su comparación con la base de datos CARD, debido a su precisión en la detección de patrones de resistencia a antibióticos en secuencias genómicas.

Figura 3. Diagrama de flujo del programa en la planeación con las herramientas a usar.



Elaboración propia.

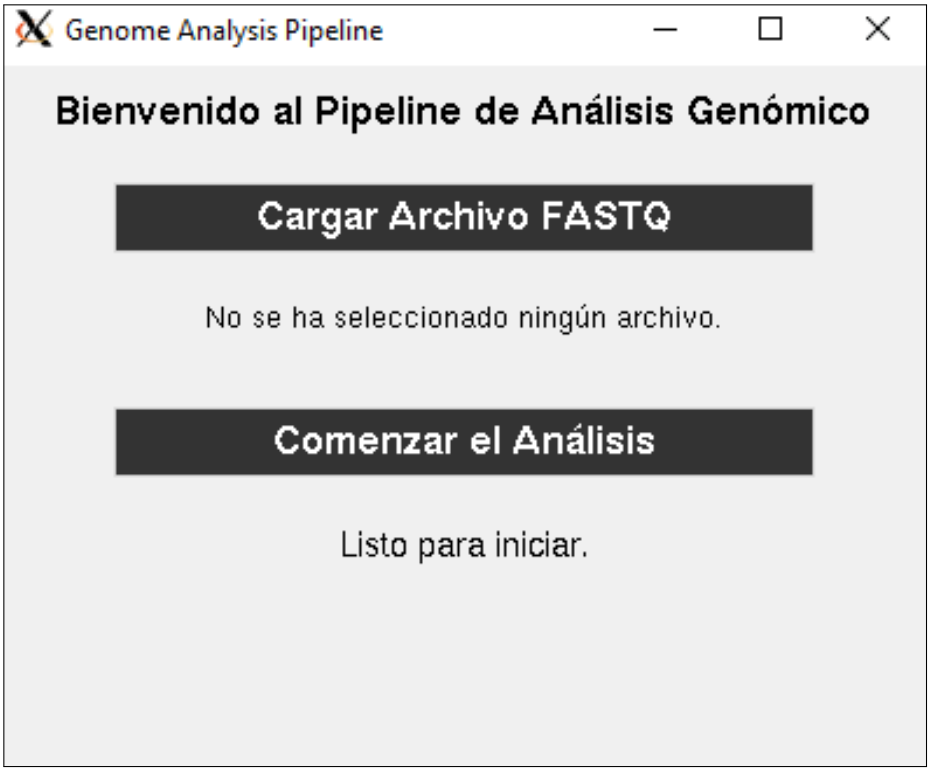
5.3.3. Diseño y usabilidad

El diseño de la interfaz gráfica del usuario fue un componente crucial en el desarrollo del *pipeline* bioinformático. El objetivo era crear una herramienta intuitiva y fácil de usar, eliminando la necesidad de conocimientos técnicos avanzados.

Primer prototipo

El primer diseño de la GUI fue intencionalmente minimalista para evitar abrumar al usuario y asegurar que se pudiera comprender fácilmente cómo operar el programa. El enfoque se basó en realizar iteraciones sobre un diseño inicial que incluía únicamente los elementos esenciales: botones para cargar los archivos *fastq* y comenzar el análisis, y un área de texto para indicar el estado del proceso. Esta área de texto ofrecía actualizaciones en tiempo real sobre el estado del análisis, manteniendo al usuario informado desde el control de calidad hasta la anotación final, culminando en la detección de genes de resistencia.

Figura 4. Primer prototipo del menú principal.



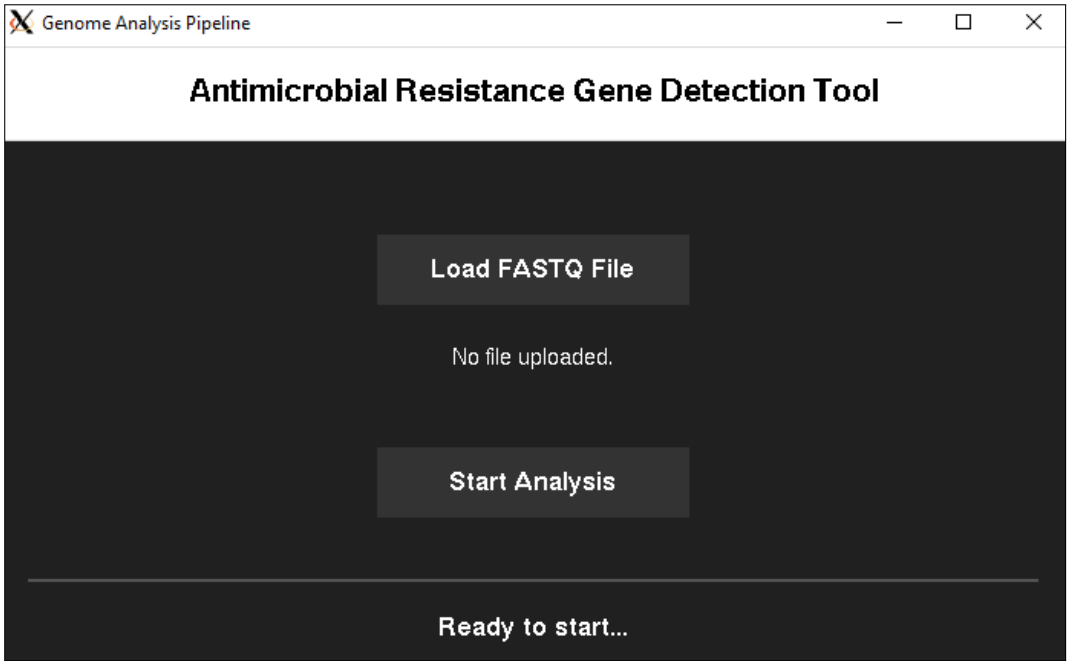
Elaboración propia.

Mejoras estéticas

Inicialmente, aunque el diseño de la GUI cumplía con su función, su apariencia era bastante básica y podría percibirse como poco atractiva. Por ello, se hicieron mejoras estéticas significativas para mejorar la interfaz sin sacrificar su simplicidad operativa. El diseño se mantuvo centrado exclusivamente en los dos botones esenciales —cargar los archivos *fastq* y comenzar el análisis—, pero se refinaron sus aspectos visuales para ofrecer una experiencia más agradable al usuario.

Esta nueva versión de la interfaz fue presentada al usuario final, quien expresó su satisfacción con las mejoras, lo que llevó a la decisión de mantener este diseño.

Figura 5. Segundo prototipo del menú principal.



Elaboración propia.

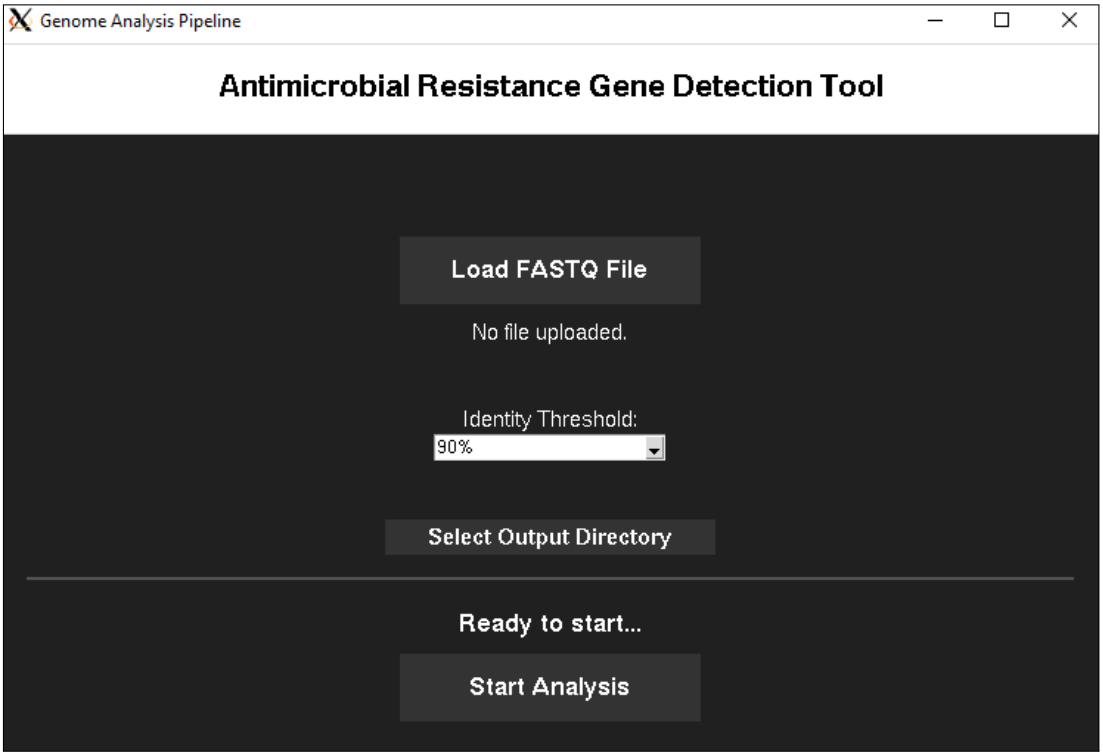
Mejoras funcionales

Además de las mejoras estéticas, se realizaron ajustes funcionales para mejorar la experiencia del usuario sin alterar el diseño fundamental de la interfaz. Uno de los ajustes fue añadir un botón adicional para permitir a los usuarios seleccionar el directorio de salida, facilitando así la organización de los archivos resultantes.

Posteriormente, se introdujo un control adicional en forma de botón desplegable, permitiendo a los usuarios ajustar el porcentaje de identidad para los resultados. El porcentaje de

identidad se refiere a la similitud genética entre las secuencias observadas y las secuencias de referencia conocidas, indicando la certeza de que una determinada secuencia corresponde a un gen de resistencia específico. Inicialmente fijado en 90 %, ahora se podía variar entre 80 % y 95 %, ofreciendo a los usuarios mayor flexibilidad y control sobre los análisis. Estas modificaciones no solo mejoraron la utilidad del programa sino que también reforzaron la autonomía del usuario en el manejo del mismo.

Figura 6. Tercer prototipo del menú principal.



Elaboración propia.

Visualización de resultados

A pesar de que el diseño inicial de la GUI cumplió con las expectativas funcionales, la presentación de los resultados necesitaba mejoras significativas. Inicialmente, los resultados se mostraban en una tabla creada con *Tkinter*. Sin embargo, esta solución presentaba varios problemas: el diseño era visualmente poco atractivo y el texto de una columna se solapaba con el de otra, debido a la limitada capacidad de ajuste del tamaño de las celdas.

Figura 7. Primera implementación de la tabla de resultados con *Tkinter*.

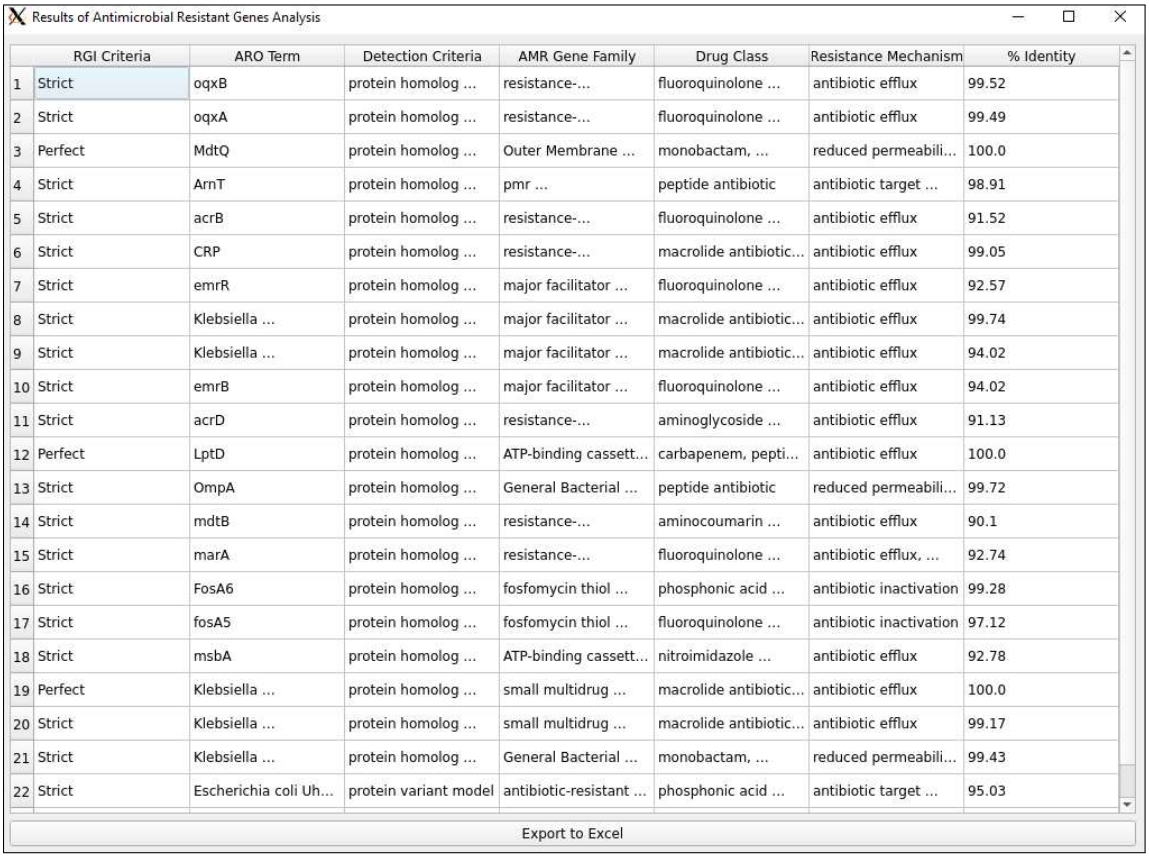
RGI Criteria	ARO Term	Detection Criteria	AMR Gene Family	Drug Class	Resistance Mechani:	% Identity
Strict	VIM-17	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	99.17
Strict	VIM-2	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	99.17
Strict	VIM-48	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	99.17
Strict	VIM-62	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	99.59
Strict	VIM-63	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-15	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-53	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	99.17
Strict	VIM-30	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-10	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-51	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-41	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-46	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-36	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-8	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-67	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-45	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-9	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-11	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-44	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-72	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-16	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-23	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-58	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-20	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-24	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-73	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-56	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.76
Strict	VIM-65	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.35
Strict	VIM-3	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.35
Strict	VIM-6	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.35
Strict	VIM-50	protein homolog mode	VIM beta-lactamase	carbapenem, cephal	antibiotic inactivatior	98.35

Elaboración propia.

Ante la imposibilidad de resolver estos problemas con ajustes menores, se decidió explorar otras herramientas más versátiles para la visualización de datos. La elección recayó en *PyQt5*, que ofrecía mejores opciones para la manipulación y presentación de tablas. La nueva implementación no solo resolvió los problemas de solapamiento de texto, sino que también mejoró considerablemente la estética de la tabla de resultados.

Esta nueva tabla fue bien recibida por el usuario final, quien apreció la claridad y la facilidad de lectura. Además, a solicitud del usuario, se añadió un botón para exportar los resultados directamente a *Excel*, aumentando así la funcionalidad del sistema y facilitando la manipulación posterior de los datos.

Figura 8. Tabla de resultados final implementada con *PyQt5*.



	RGI Criteria	ARO Term	Detection Criteria	AMR Gene Family	Drug Class	Resistance Mechanism	% Identity
1	Strict	oqxB	protein homolog ...	resistance-...	fluoroquinolone ...	antibiotic efflux	99.52
2	Strict	oqxA	protein homolog ...	resistance-...	fluoroquinolone ...	antibiotic efflux	99.49
3	Perfect	MdtQ	protein homolog ...	Outer Membrane ...	monobactam, ...	reduced permeabili...	100.0
4	Strict	ArnT	protein homolog ...	pmr ...	peptide antibiotic	antibiotic target ...	98.91
5	Strict	acrB	protein homolog ...	resistance-...	fluoroquinolone ...	antibiotic efflux	91.52
6	Strict	CRP	protein homolog ...	resistance-...	macrolide antibiotic...	antibiotic efflux	99.05
7	Strict	emrR	protein homolog ...	major facilitator ...	fluoroquinolone ...	antibiotic efflux	92.57
8	Strict	Klebsiella ...	protein homolog ...	major facilitator ...	macrolide antibiotic...	antibiotic efflux	99.74
9	Strict	Klebsiella ...	protein homolog ...	major facilitator ...	macrolide antibiotic...	antibiotic efflux	94.02
10	Strict	emrB	protein homolog ...	major facilitator ...	fluoroquinolone ...	antibiotic efflux	94.02
11	Strict	acrD	protein homolog ...	resistance-...	aminoglycoside ...	antibiotic efflux	91.13
12	Perfect	LptD	protein homolog ...	ATP-binding cassett...	carbapenem, pepti...	antibiotic efflux	100.0
13	Strict	OmpA	protein homolog ...	General Bacterial ...	peptide antibiotic	reduced permeabili...	99.72
14	Strict	mdtB	protein homolog ...	resistance-...	aminocoumarin ...	antibiotic efflux	90.1
15	Strict	marA	protein homolog ...	resistance-...	fluoroquinolone ...	antibiotic efflux, ...	92.74
16	Strict	FosA6	protein homolog ...	fosfomycin thiol ...	phosphonic acid ...	antibiotic inactivation	99.28
17	Strict	fosA5	protein homolog ...	fosfomycin thiol ...	fluoroquinolone ...	antibiotic inactivation	97.12
18	Strict	msbA	protein homolog ...	ATP-binding cassett...	nitroimidazole ...	antibiotic efflux	92.78
19	Perfect	Klebsiella ...	protein homolog ...	small multidrug ...	macrolide antibiotic...	antibiotic efflux	100.0
20	Strict	Klebsiella ...	protein homolog ...	small multidrug ...	macrolide antibiotic...	antibiotic efflux	99.17
21	Strict	Klebsiella ...	protein homolog ...	General Bacterial ...	monobactam, ...	reduced permeabili...	99.43
22	Strict	Escherichia coli Uh...	protein variant model	antibiotic-resistant ...	phosphonic acid ...	antibiotic target ...	95.03

Elaboración propia.

5.4. Validación del programa

La validación del *pipeline* desarrollado fue un paso crucial para asegurar su fiabilidad y exactitud. Este proceso permitió verificar la funcionalidad de la herramienta, facilitar la corrección e iteración del diseño, y evaluar su capacidad para identificar correctamente la presencia o ausencia de genes de resistencia, reduciendo así los riesgos de falsos positivos y negativos.

5.4.1. Pruebas con genomas conocidos

Para validar el *pipeline*, la base de datos SRA de NCBI fue de gran utilidad, ya que contiene genomas de una amplia variedad de organismos, incluyendo bacterias comunes en las dietas de *Ceratitis capitata*, como *Klebsiella pneumoniae* y *Escherichia coli*. Utilizando la barra de búsqueda avanzada y operadores lógicos como `AND`, junto con términos como ‘‘`antimicrobial resistance`’’, se buscaron genomas donde la presencia de genes de re-

sistencia a antibióticos era conocida, así como aquellos en los que no se esperaban estos genes. Este enfoque permitió probar el programa bajo condiciones controladas y evaluar su precisión.

Se analizaron 53 genomas de *Escherichia coli* y *Klebsiella pneumoniae* en total para validar el *pipeline*. Los datos provinieron de tres BioProjects específicos: PRJNA307517, PRJNA1076266 y PRJNA392824, que contenían secuencias de lectura (SRA) disponibles en NCBI y que documentaban resistencia a antibióticos. Además, se incluyó un BioProject de control, PRJNA279657, con una secuencia de *Klebsiella pneumoniae* sin genes de resistencia reportados, para verificar la capacidad del *pipeline* para identificar correctamente la ausencia de genes de resistencia.

Cuadro 2. Distribución de las secuencias SRA utilizadas para la validación del *pipeline*.

BioProject	**Número de SRA**
PRJNA307517 - (*Klebsiella pneumoniae*)	24
PRJNA392824 - (*Klebsiella pneumoniae*)	12
PRJNA1076266 - (*Escherichia coli*)	16
PRJNA279657 (*Klebsiella pneumoniae*) (Control)	1
Total	**53**

Elaboración propia.

Para la detección de genes de resistencia, se empleó la herramienta *Resistance Gene Identifier* (RGI) versión 6.0.3 con un umbral de identidad del 90 %. La validación del *pipeline* se basó en el método propuesto por Bogaerts et al., utilizando las fórmulas de precisión, exactitud, sensibilidad y especificidad(Bogaerts et al., 2019). Estas métricas se evaluaron clasificando los resultados en verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN). Un TP indica un gen detectado por el *pipeline* y que está presente en el genoma de referencia; un FP, un gen detectado por el *pipeline* pero ausente en el genoma de referencia; un TN, un gen no detectado por el *pipeline* y ausente en el genoma de referencia; y un FN, un gen no detectado por el *pipeline* pero presente en el genoma de referencia.

Cuadro 3. Parámetros evaluados en el análisis de rendimiento y validación del *pipeline*.

Término	**Descripción**	**Fórmula**
Exactitud (*Accuracy*)	Probabilidad de que los resultados del ensayo sean correctos	$\frac{TP+TN}{TP+TN+FP+FN} \times 100\,\%$
Precisión (*Precision*)	Probabilidad de que los resultados detectados sean verdaderamente positivos	$\frac{TP}{TP+FP} \times 100\,\%$
Sensibilidad (*Sensitivity*)	Probabilidad de que el resultado sea detectado correctamente en el ensayo cuando está presente	$\frac{TP}{TP+FN} \times 100\,\%$
Especificidad (*Specificity*)	Probabilidad de que un resultado no sea detectado falsamente en un ensayo cuando está ausente	$\frac{TN}{TN+FP} \times 100\,\%$

TP = verdadero positivo; TN = verdadero negativo; FP = falso positivo; FN = falso negativo.

Elaboración propia.

CAPÍTULO 6

Resultados y discusión

Para validar el *pipeline*, se procesaron 53 genomas de *Klebsiella pneumoniae* y *Escherichia coli*. Estos genomas pasaron por las etapas de control de calidad, ensamblaje, identificación de especie, anotación y, finalmente, detección de genes de resistencia a antibióticos.

6.1. Resultados con genomas de validación

Los genes de resistencia a antibióticos (AMR's) más prevalentes en las 53 muestras analizadas, tanto de *Klebsiella pneumoniae* como de *Escherichia coli*, fueron *sul2*, con un 64 % de presencia; *tet(A)*, con un 57 %; *APH(6)-Id*, con un 46 %; *APH(3")-Ib*, con un 43 %; y *KPC-2*, con un 41 %. La lista de los genes más presentes en los genomas se puede observar en el Cuadro 4.

Cuadro 4. Proporción de genes de resistencia en las 53 muestras con su correspondiente resistencia a antibióticos.

Gen	**Resistencia a medicamento**	**% de detección**
sul2	Sulfonamida	64 %
tet(A)	Tetraciclina	57 %
APH(6)-Id	Aminoglicosido	46 %
APH(3")-Ib	Aminoglicosido	43 %
KPC-2	Carbapenem, penam, cefalosporina	41 %
TEM-1	Penam, monobactam, penem, cefalosporina	29.5 %
sul1	Sulfonamida	27 %
floR	Fenicol	27 %
AAC(3)-IId	Aminoglicosido	28 %
dfrA14	Diaminopirimidina	21 %
CTX-M-65	Cefalosporina	19 %

Elaboración propia.

Las métricas para la validación del *pipeline* en cada BioProject se presentan en el Cuadro 5. Los cálculos de exactitud, precisión, sensibilidad y especificidad se realizaron comparando el número y tipo de los genes obtenidos por el *pipeline* para cada SRA con su secuencia de referencia (RefSeq) correspondiente. Esta comparación fue esencial, razón por la cual fue indispensable que los BioProjects seleccionados contaran con un RefSeq.

Cuadro 5. Métricas de rendimiento del *pipeline* por cada BioProject.

Métrica	**BioProject 1**	**BioProject 2**	**BioProject 3**	**BioProject 4**
Exactitud (*Accuracy*)	97.13 %	98.13 %	86.41 %	100.00 %
Precisión (*Precision*)	95.43 %	97.32 %	90.32 %	100.00 %
Sensibilidad (*Sensitivity*)	93.24 %	89.34 %	83.34 %	100.00 %
Especificidad (*Specificity*)	100.00 %	100.00 %	99.20 %	100.00 %

Elaboración propia.

En el BioProject 4, que correspondía al genoma de control de *Klebsiella pneumoniae*, el cual no presentaba genes de resistencia a antibióticos, no se detectó ningún gen de resistencia. Como resultado, sus métricas alcanzaron un 100 %.

6.2. Aplicación del *pipeline* en muestras reales de la dieta de *Ceratitis capitata*

Después de validar el *pipeline* con genomas de referencia, se analizaron las muestras reales obtenidas de la dieta de *Ceratitis capitata*. Se procesaron cuatro muestras, de las cuales todas presentaron genes de resistencia a antibióticos, detectados con un umbral mínimo del 90 % de identidad.

Las muestras analizadas fueron identificadas como de las especies *Klebsiella pneumoniae* (dos muestras), *Kluyvera cryocrescens* y *Kluyvera ascorbata*. En los siguientes cuadros, se detallan los genes de resistencia identificados.

Cuadro 6. Genes de resistencia identificados en las primer muestra de *Klebsiella pneumoniae*.

Gen	**Mecanismo de resistencia**	**% de identidad**
eptB	Alteración del objetivo antibiótico	99.28
ArnT	Alteración del objetivo antibiótico	98.91
acrD	Eflujo antibiótico	91.13
oqxA	Eflujo antibiótico	100
oqxB	Eflujo antibiótico	99.71
OmpA	Reducción de la permeabilidad al antibiótico	99.72
FosA6	Inactivación del antibiótico	99.28
fosA5	Inactivación del antibiótico	97.12
LptD	Eflujo antibiótico	100
KpnF	Eflujo antibiótico	98.17
KpnE	Eflujo antibiótico	99.17
OmpK37	Reducción de la permeabilidad al antibiótico	99.2
CRP	Eflujo antibiótico	99.05
acrB	Eflujo antibiótico	91.5
emrR	Eflujo antibiótico	92.57
KpnG	Eflujo antibiótico	99.74
KpnH	Eflujo antibiótico	94.02
emrB	Eflujo antibiótico	94.02
SHV-37	Inactivación del antibiótico	100
msbA	Eflujo antibiótico	92.78
mdtC	Eflujo antibiótico	91.61
MdtQ	Reducción de la permeabilidad al antibiótico	99.79
marA	Eflujo antibiótico	92.74
UhpT	Alteración del objetivo antibiótico	95.03
baeR	Eflujo antibiótico	93.75

Elaboración propia.

Cuadro 7. Genes de resistencia identificados en la muestra de *Kluyvera cryocrescens*

ARO Term	**Mecanismo de resistencia**	**% de identidad**
mdtB	Eflujo antibiótico	91.06
mdtC	Eflujo antibiótico	91.22
CTX-M-95	Inactivación del antibiótico	90.38
CTX-M-77	Inactivación del antibiótico	90.03
CTX-M-165	Inactivación del antibiótico	90.03
CTX-M-76	Inactivación del antibiótico	90.03
acrB	Eflujo antibiótico	91.71
msbA	Eflujo antibiótico	92.78
marA	Eflujo antibiótico	91.94
CRP	Eflujo antibiótico	99.05
emrR	Eflujo antibiótico	91.43
emrB	Eflujo antibiótico	92.03
KpnH	Eflujo antibiótico	91.83
acrD	Eflujo antibiótico	90.07
UhpT	Alteración del objetivo antibiótico	95.03
baeR	Eflujo antibiótico	92.92

Elaboración propia.

Cuadro 8. Genes de resistencia identificados en la muestra de *Kluyvera ascorbata* (Parte 1).

Gen	**Mecanismo de resistencia**	**% de identidad**
mdtB	Eflujo de antibióticos	90.87
mdtC	Eflujo de antibióticos	91.12
APH(6)-Id	Inactivación de antibióticos	99.64
APH(3")-Ib	Inactivación de antibióticos	99.63
msbA	Eflujo de antibióticos	92.61
tet(C)	Eflujo de antibióticos	100
CRP	Eflujo de antibióticos	99.05
acrD	Eflujo de antibióticos	90.07
marA	Eflujo de antibióticos, reducción de permeabilidad	91.94
KpnH	Eflujo de antibióticos	91.07
CTX-M-95	Inactivación de antibióticos	92.48
CTX-M-77	Inactivación de antibióticos	92.11
CTX-M-2	Inactivación de antibióticos	91.73
CTX-M-124	Inactivación de antibióticos	91.73
CTX-M-97	Inactivación de antibióticos	91.73
CTX-M-59	Inactivación de antibióticos	91.73

Elaboración propia.

Cuadro 8. Genes de resistencia identificados en la muestra de *Kluyvera ascorbata* (Continuación de Cuadro 8).

Gen	**Mecanismo de resistencia**	**% de identidad**
CTX-M-20	Inactivación de antibióticos	91.73
CTX-M-76	Inactivación de antibióticos	92.11
CTX-M-31	Inactivación de antibióticos	91.35
CTX-M-56	Inactivación de antibióticos	91.35
CTX-M-115	Inactivación de antibióticos	91.35
CTX-M-44	Inactivación de antibióticos	91.35
CTX-M-92	Inactivación de antibióticos	91.35
CTX-M-141	Inactivación de antibióticos	91.35
CTX-M-165	Inactivación de antibióticos	91.35
CTX-M-200	Inactivación de antibióticos	91.35
CTX-M-171	Inactivación de antibióticos	91.35
CTX-M-35	Inactivación de antibióticos	91.35
CTX-M-131	Inactivación de antibióticos	91.35
CTX-M-5	Inactivación de antibióticos	91.73
CTX-M-43	Inactivación de antibióticos	90.98
EF-Tu	Inactivación de antibióticos	98.22
baeR	Eflujo de antibióticos	92.5

Elaboración propia.

Cuadro 9. Genes de resistencia identificados en la segunda muestra de *Klebsiella pneumoniae*.

Gen	**Mecanismo de resistencia**	**% de identidad**
oqxB	Eflujo de antibióticos	99.52
oqxA	Eflujo de antibióticos	99.49
MdtQ	Reducción de la permeabilidad a antibióticos	100
ArnT	Alteración del objetivo de antibióticos	98.91
acrB	Eflujo de antibióticos	91.52
CRP	Eflujo de antibióticos	99.05
emrR	Eflujo de antibióticos	92.57
KpnG	Eflujo de antibióticos	99.74
KpnH	Eflujo de antibióticos	94.02
emrB	Eflujo de antibióticos	94.02
acrD	Eflujo de antibióticos	91.13
LptD	Eflujo de antibióticos	100
OmpA	Reducción de la permeabilidad a antibióticos	99.72
mdtB	Eflujo de antibióticos	90.1
marA	Reducción de la permeabilidad a antibióticos	92.74
FosA6	Inactivación de antibióticos	99.28
fosA5	Inactivación de antibióticos	97.12
msbA	Eflujo de antibióticos	92.78
KpnF	Eflujo de antibióticos	100
KpnE	Eflujo de antibióticos	99.17
OmpK37	Reducción de la permeabilidad a antibióticos	99.43
UhtP	Alteración del objetivo de antibióticos	95.03
baeR	Eflujo de antibióticos	93.75

Elaboración propia.

Se identificaron un total de 97 genes de resistencia a lo largo de las cuatro muestras, incluyendo repeticiones. Los genes que mostraron mayor prevalencia, estando presentes en todas las muestras, incluyen acrD, CRP, msbA, marA y baeR. A continuación, se presenta un cuadro que destaca los genes que aparecieron con mayor frecuencia en todas las muestras analizadas.

Cuadro 10. Genes con mayor presencia en las cuatro muestras de bacterias presentes en la dieta de *Ceratitis capitata*.

Gen	**% de presencia**
acrD	100
CRP	100
acrB	75
emrB	75
msbA	100
mdtC	75
marA	100
baeR	100

Elaboración propia.

6.3. Discusión

En este proyecto se validó un *pipeline* bioinformático diseñado para la identificación de genes de resistencia a antibióticos en bacterias. Posteriormente, fue aplicado a bacterias presentes en la dieta de *Ceratitis Capitata*.

A diferencia de otras herramientas existentes, este *pipeline* se diseñó para ser especialmente accesible y fácil de usar, sin requerir conocimientos técnicos o bioinformáticos avanzados. Se enfocó en ser una solución *"plug-and-play"*, lo cual reduce significativamente la barrera de entrada y permite que un público más amplio pueda realizar análisis de secuencias de manera efectiva.

Este programa incorpora herramientas ampliamente utilizadas como *fastp* y *SPAdes*, por lo que, en teoría, no se deberían de generar resultados inesperadamente erróneos. Sin embargo, esto no elimina la necesidad de una validación adecuada. La combinación de diversas herramientas, la forma en que los datos se procesan a través de cada etapa y hasta las particularidades de la programación pueden afectar los resultados finales. Por esta razón, una validación meticulosa es crucial antes de aplicar el *pipeline* a datos experimentales. De no

llevarse a cabo, no se podría estar seguro de la veracidad de los resultados obtenidos.

Para validar este *pipeline*, se adoptó el protocolo propuesto por Bogaerts et al. (Bogaerts et al., 2019), diseñado específicamente para la validación de flujos de trabajo bioinformáticos que incluyen la detección de genes de resistencia a antibióticos. El protocolo emplea métricas de rendimiento tales como exactitud, precisión, sensibilidad y especificidad, para evaluar la caracterización de genes de resistencia, utilizando 131 secuencias de *Neisseria meningitidis* como modelo de prueba. Estas métricas permiten una evaluación de la capacidad del *pipeline* para identificar correctamente los genes de resistencia en los organismos analizados.

En este proyecto, se analizaron 53 muestras de *Klebsiella pneumoniae* y *Escherichia coli* utilizando las métricas de exactitud, precisión, sensibilidad y especificidad para validar el *pipeline*, siguiendo el proceso y las fórmulas propuestas por Bogaerts et al. Estas muestras formaban parte de cuatro BioProjects, y las métricas se calcularon de forma individual para cada uno.

Todos los BioProjects, incluido el BioProject 4 que actuó como control negativo, mostraron métricas relativamente altas. La sensibilidad más baja se observó en el BioProject 3 con un 83.34 %, que incluía muestras de *Escherichia coli*. Este BioProject presentó las métricas más bajas, destacando una de las limitaciones principales del estudio: dado que las muestras se obtuvieron de bases de datos públicas, las condiciones exactas de secuenciación y procesamiento no son conocidas, introduciendo un elemento de incertidumbre sobre por qué las métricas resultaron de la manera en que lo hicieron. Sin embargo, los BioProjects 1 y 2, que incluían genomas de *K. pneumoniae*, no solo mostraron valores altos sino también consistentes, sugiriendo que el *pipeline* mantiene una coherencia en los resultados entre muestras del mismo organismo.

El hecho de que la métrica más baja fuera del 83.34 % y que la gran mayoría de las demás superara el 95 %, demuestra que la validación del *pipeline* fue exitosa y que se puede confiar significativamente en su uso. Una vez validado, se procedió con el análisis de las muestras experimentales obtenidas de la dieta de *Ceratitis capitata*, incluyendo dos muestras de *Klebsiella pneumoniae*, una de *Kluyvera cryocrescens* y otra de *Kluyvera ascorbata*.

En el análisis de estas muestras, empleando nuevamente un umbral mínimo de identidad del 90 %, se detectaron genes de resistencia a antibióticos en todas ellas, identificando un total de 97 genes diferentes. Este hallazgo comienza a subrayar la utilidad de haber desarrollado un programa como este. Esto, ya que las bacterias pueden permanecer en el microbioma de las moscas y, al ser liberadas, pueden diseminarse, representando un riesgo significativo para la salud pública.

Los genes de resistencia presentes en todas las muestras analizadas fueron acrD, CRP, msbA, marA y baeR. Todos estos genes, a excepción de baeR, pertenecen a familias de

bombas de eflujo y están involucrados en mecanismos que reducen la efectividad de los antibióticos. Las bombas de eflujo son proteínas transportadoras que expulsan una variedad de sustancias, incluidos los antibióticos, fuera de las células bacterianas. Este mecanismo de resistencia permite a las bacterias sobrevivir en ambientes con altas concentraciones de antibióticos, lo que puede dar lugar a infecciones crónicas y difíciles de tratar (Soto, 2013). El hecho de que la mayoría de genes de resistencia presentes en estas muestras tengan este mecanismo, puede significar que las poblaciones bacterianas en el entorno de *Ceratitis Capitata* están adaptadas para resistir tratamientos antimicrobianos convencionales.

Además de los genes de resistencia que este *pipeline* identifica, otro factor de valor agregado es que el programa genera datos de cada etapa a medida que la información avanza a través de cada paso del *pipeline*. Es decir, se almacena información residual detallada del control de calidad, el ensamblaje, la anotación, entre otros. Esto significa que el *pipeline* no solo provee información sobre los genes de resistencia identificados, sino también sobre el genoma en general. Esto, de hecho, fue uno de los aspectos que más apreció el usuario inicial.

En resumen, este estudio no solo validó el *pipeline* para la detección de genes de resistencia a antibióticos, sino que también demostró su aplicabilidad en muestras reales, ofreciendo una herramienta accesible y potente para múltiples usuarios. Su diseño intuitivo y la vasta información que proporciona permiten que pueda ser una herramienta valiosa en la lucha contra la propagación de resistencias bacterianas. El uso de este *pipeline* podría contribuir significativamente a los esfuerzos para monitorizar y prevenir la diseminación de bacterias resistentes, apoyando tanto la investigación científica como la implementación de medidas de salud pública eficaces.

CAPÍTULO 7

Conclusiones

- El *pipeline* bioinformático desarrollado fue validado exitosamente, demostrando alta exactitud, precisión, sensibilidad y especificidad en la identificación de genes de resistencia a antibióticos.
- La implementación del *pipeline* facilitó el análisis de secuencias genéticas sin necesidad de conocimientos técnicos avanzados, lo cual democratiza el acceso a herramientas bioinformáticas y puede ampliar su uso en estudios microbiológicos y epidemiológicos.
- Los resultados de las muestras experimentales de la dieta de *Ceratitis capitata* revelan una prevalencia significativa de genes de resistencia a antibióticos, lo cual no solo subraya el riesgo potencial para la salud pública si estas bacterias se propagan, sino también demuestra la utilidad del *pipeline* en la detección precisa de estos elementos.
- La capacidad del *pipeline* para almacenar datos detallados en cada etapa del análisis no solo proporciona resultados finales de resistencia a antibióticos, sino también información valiosa sobre el genoma que puede ser utilizada para investigaciones adicionales.
- Este estudio contribuye significativamente al campo de la bioinformática aplicada al control de enfermedades transmitidas por vectores, mediante la propuesta de una herramienta que no solo evalúa la resistencia a los antibióticos, sino que también facilita la gestión de riesgos asociados con la liberación de insectos modificados.

CAPÍTULO 8

Recomendaciones

- Se recomienda realizar estudios adicionales para explorar la prevalencia y distribución de genes de resistencia en otras poblaciones de bacterias asociadas a la dieta de *Ceratitis capitata*, así como en otras especies que puedan estar involucradas en programas similares de control de plagas, con el fin de entender mejor el riesgo para la salud pública.
- Dada la importancia de los datos de secuenciación precisos para la identificación de genes de resistencia, se sugiere una estandarización en los procesos de secuenciación y manejo de muestras para asegurar la consistencia y fiabilidad de los resultados con el *pipeline*.

CAPÍTULO 9

Referencias

Adamowicz, L., Christakis, Y., Czech, M. D., & Adamusiak, T. (2022). SciKit Digital Health: Python package for streamlined wearable inertial sensor data processing. *JMIR MHealth UHealth*, *10*(4), e36762.

Al-Behadili, F. J. M., Agarwal, M., Xu, W., & Ren, Y. (2020). Mediterranean fruit fly Ceratitis capitata (Diptera: Tephritidae) eggs and larvae responses to a low-oxygen/high-nitrogen atmosphere. *Insects*, *11*(11), 802.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K. D., Resenchuk, S., Tatusova, T., Yaschenko, E., & Ostell, J. (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, *40*(Database issue), D57-63.

Benedict, M. Q. (2021). Sterile insect technique: Lessons from the past. *J. Med. Entomol.*, *58*(5), 1974-1979.

Blair, J. M. A., Webber, M., Baylay, A., Ogbolu, D., & Piddock, L. (2014). Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology*, *13*, 42-51. https://doi.org/10.1038/nrmicro3380

Bogaerts, B., Winand, R., Fu, Q., Van Braekel, J., Ceyssens, P.-J., Mattheus, W., Bertrand, S., De Keersmaecker, S. C. J., Roosens, N. H. C., & Vanneste, K. (2019). Validation of a bioinformatics workflow for routine analysis of whole-genome sequencing data and related challenges for pathogen typing in a European national reference center: Neisseria meningitidis as a proof-of-concept. *Front. Microbiol.*, *10*, 362.

Bourtzis, K., & Vreysen, M. J. B. (2021). Sterile insect technique (SIT) and its applications. *Insects*, *12*(7), 638.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890.

Chen, X., & Liu, W. (2022). The value of Python programming in general education and comprehensive quality improvement of medical students based on a retrospective cohort study. *J. Healthc. Eng.*, *2022*, 4043992.

Chokshi, S. K., & Mann, D. M. (2018). Innovating from within: A process model for user-centered digital development in academic medical centers. *JMIR Hum. Factors*, *5*(4), e11048.

Ciorba, V., Odone, A., Veronesi, L., Pasquarella, C., & Signorelli, C. (2015). Antibiotic resistance as a major public health concern: epidemiology and economic impact. *Ann. Ig.*, *27*(3), 562-579.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, *38*(6), 1767-1771.

Cuddy, P. G. (1997). Antibiotic classification. *Crit. Care Nurs. Q.*, *20*(3), 89.

Dever, L. A. (1991). Mechanisms of bacterial resistance to antibiotics. *Arch. Intern. Med.*, *151*(5), 886.

Dionysopoulou, N. K., Papanastasiou, S. A., Kyritsis, G. A., & Papadopoulos, N. T. (2020). Effect of host fruit, temperature and Wolbachia infection on survival and development of Ceratitis capitata immature stages. *PLoS One*, *15*(3), e0229727.

Fourment, M., & Gillings, M. R. (2008). A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics*, *9*(1), 82.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J., & Bioconda Team. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, *15*(7), 475-476.

Hutchings, M. I., Truman, A. W., & Wilkinson, B. (2019). Antibiotics: past, present and future. *Curr. Opin. Microbiol.*, *51*, 72-80.

Ishengoma, E., & Rhode, C. (2022). Using SPAdes, AUGUSTUS, and BLAST in an automated pipeline for clustering homologous exome sequences. *Curr. Protoc.*, *2*(5), e449.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., Doshi, S., Courtot, M., Lo, R., Williams, L. E., Frye, J. G., Elsayegh, T., Sardar, D., Westman, E. L., Pawlowski, A. C., ... McArthur, A. G. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, *45*(D1), D566-D573.

Kapoor, G., Saigal, S., & Elongavan, A. (2017). Action and resistance mechanisms of antibiotics: A guide for clinicians. *J. Anaesthesiol. Clin. Pharmacol.*, *33*(3), 300.

Lopez, C. F., Muhlich, J. L., Bachman, J. A., & Sorger, P. K. (2013). Programming biological models in Python using PySB. *Mol. Syst. Biol.*, *9*(1), 646.

Marec, F., & Vreysen, M. J. B. (2019). Advances and challenges of using the sterile insect technique for the management of pest Lepidoptera. *Insects*, *10*(11), 371.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J. V., Spanogiannopoulos, P., ... Wright, G. D. (2013). The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.*, *57*(7), 3348-3357.

Munita, J. M., & Arias, C. A. (2016, enero). Mechanisms of antibiotic resistance. En *Virulence Mechanisms of Bacterial Pathogens, Fifth Edition* (pp. 481-511). American Society of Microbiology.

Navarro-Llopis, V., Primo, J., & Vacas, S. (2013). Efficacy of attract-and-kill devices for the control of Ceratitis capitata. *Pest Manag. Sci.*, *69*(4), 478-482.

Ossom Williamson, P., & Minter, C. I. J. (2019). Exploring PubMed as a reliable resource for scholarly communications services. *J. Med. Libr. Assoc.*, *107*(1), 16-29.

Pérez-Staples, D., Díaz-Fleischer, F., & Montoya, P. (2021). The sterile insect technique: Success and perspectives in the Neotropics. *Neotrop. Entomol.*, *50*(2), 172-185.

Plá, I., García de Oteyza, J., Tur, C., Martínez, M. Á., Laurín, M. C., Alonso, E., Martínez, M., Martín, Á., Sanchis, R., Navarro, M. C., Navarro, M. T., Argilés, R., Briasco, M., Dembilio, Ó., & Dalmau, V. (2021). Sterile insect technique programme against Mediterranean fruit fly in the Valencian Community (Spain). *Insects*, *12*(5).

Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, *40*(Database issue), D130-5.

Sayers, E. W., O'Sullivan, C., & Karsch-Mizrachi, I. (2022). Using GenBank and SRA. *Methods Mol. Biol.*, *2443*, 1-25.

Sciarretta, A., Tabilio, M. R., Lampazzi, E., Ceccaroli, C., Colacci, M., & Trematerra, P. (2018). Analysis of the Mediterranean fruit fly [Ceratitis capitata (Wiedemann)] spatio-temporal distribution in relation to sex and female mating status for precision IPM. *PLoS One*, *13*(4), e0195097.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069.

SoRelle, J. A., Wachsmann, M., & Cantarel, B. L. (2020). Assembling and validating bioinformatic pipelines for next-generation sequencing clinical assays. *Arch. Pathol. Lab. Med.*, *144*(9), 1118-1130.

Soto, S. M. (2013). Role of efflux pumps in the antibiotic resistance of bacteria embedded in a biofilm. *Virulence*, *4*(3), 223-229.

Tabilio, M. R., Fiorini, D., Marcantoni, E., Materazzi, S., Delfini, M., De Salvador, F. R., & Musmeci, S. (2013). Impact of the Mediterranean fruit fly (medfly) Ceratitis capitata

on different peach cultivars: the possible role of peach volatile compounds. *Food Chem.*, *140*(1-2), 375-381.

Thrum, H. (1977). Antibiotikaklassen und ihre Wirkungsmechanismen. *Zeitschrift fur die gesamte innere Medizin und ihre Grenzgebiete*, *32*, 209-214.

UN. (2015). Antimicrobial resistance: a global threat. *Ann. Ig.*, *27*(3), 562-579.

Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *P T*, *40*(4), 277-283.

Vine, R. (2006). Google scholar. *Journal of the Medical Library Association*, *94*(1), 97.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.*, *20*(1), 257.

Wright, G. D. (2010). The antibiotic resistome. *Expert Opinion on Drug Discovery*, *5*, 779-788. https://doi.org/10.1517/17460441.2010.497535

Zinner, S. H. (2007). Antibiotic use: present and future. *Newmicrobiologica*, *9*(4), 249-252.

CAPÍTULO 10

Anexos

10.1. Enlaces y documentos

10.1.1. Código de Github

En el siguiente enlace, se encuentra el código del *pipeline*: https://github.com/luispedro10/resistance_pipeline

10.1.2. Guías y tutoriales

En el siguiente enlace, se encuentra la guía de cómo instalar y correr el *pipeline*: https://docs.google.com/document/d/1Wm3qnEtW2-0Z_YjDGm0EiiDdN2SrTsJFlqXUvwVDxjc/edit?usp=sharing

En el siguiente enlace, se proporciona la demostración del programa: https://youtu.be/GZPqERyf4ls

En el siguiente enlace, se encuentra un video que explica cómo instalar y correr el programa, utilizando la guía como referencia: https://youtu.be/Lm3WByBnFac