Máster en Bioinformática

Desarrollo de un pipeline Bioinformático mediante nextflow:

Caracterización molecular de cepas de *Staphylococcus aureus* en pacientes con enfermedad invasora en España.

Autor: Víctor Alejandro Pizarro Riveros

Tutora(s): Sílvia García Cobos

María Pérez Vázquez

Curso 2023-24

Agradecimientos

Quiero expresar mi más profundo agradecimiento a María Pérez y Silvia García por brindarme la oportunidad de seguir la línea que me apasiona: el estudio microbiológico. Este campo, gracias a la Bioinformática, ha alcanzado un punto que no imaginaba posible. Su constante dedicación y apoyo frente a los nuevos desafíos, así como sus ideas para optimizar y mejorar el pipeline, me llevaron a explorar nuevas formas de abordar el trabajo y cumplir con las expectativas que depositaron en mí.

También quiero agradecer a mi madre y a mi círculo de amigos, quienes semana a semana escucharon con entusiasmo mis relatos tras cada reunión, celebrando conmigo cada logro y avance a lo largo de este desafío. Esta memoria no habría sido posible sin su constante apoyo y la confianza que siempre me inspiraron para dar lo mejor de mí.

Resumen

Objetivos: Desarrollar un pipeline de bioinformática utilizando los sistemas *Nextflow* y nf-core, que permiten crear flujos de trabajo automatizados, para analizar genomas completos (secuencias fastq) de *Staphylococcus aureus*.

Materiales y Métodos: Se emplearon treinta secuencias de genoma de *S. aureus,* obtenidas del Centro Nacional de Microbiología (ISCIII). El pipeline fue diseñado usando *Nextflow* y herramientas de *nf-core*, siguiendo las mejores prácticas para garantizar la reproducibilidad. Incluye módulos de control de calidad de lecturas (FastQC y FastP), ensamblaje (Unicycler y Shovill) y anotación (Prokka) de genomas bacterianos, y además proporciona, un análisis de genes de resistencia a antibióticos y genes de virulencia (ARIBA, Abricate y Mykrobe). También se utilizó IQ-TREE y MASH para la construcción de árboles filogenéticos.

Resultados: El pipeline ***StaPhyloRes*** mostró un desempeño eficiente en un sistema de computación de alto rendimiento (HPC, del inglés *High-performance computing*), generando resultados reproducibles y de alta calidad en cada módulo. La elección de Unicycler fue priorizada por una mayor calidad de los ensamblajes resultantes. Los resultados de la tipificación molecular (MLST, *spa*-tipo, *SCCmec* tipo y *agr* tipo) se agruparon en un resumen local para una visualización tipo resumen, al igual que los resultados de análisis de genes de resistencia a antibióticos y a virulencia que se resumieron según cada herramienta y base de datos utilizada. Finalmente, se logró la integración y consolidación de los datos de control de calidad de lecturas y ensamblaje en un único informe visual mediante MultiQC.

Conclusiones: ***StaPhyloRes*** es una herramienta útil y accesible para el estudio de genomas de *S. aureus*, permitiendo identificar genes de resistencia a antibióticos y virulencia, y proporcionando una base para estudios de vigilancia epidemiológica. Su diseño modular facilita la adaptación para otros patógenos y entornos de análisis bioinformático, aportando valor a la prevención de infecciones bacterianas hospitalarias.

Palabras Clave: *Staphylococcus aureus*, resistencia a antibióticos, *Nextflow*, tipificación molecular, análisis filogenético, bioinformática.

Abastract

Objectives: Develop a bioinformatics pipeline using *Nextflow* and *nf-core* systems, which enables the creation of automated workflows to analyse complete genomes (fastq sequences) of *Staphylococcus aureus.*

Materials and Methods: Thirty *S. aureus* genome sequences, obtained from the National Center for Microbiology (ISCIII), were used. The pipeline was designed using *Nextflow* and *nf-core* tools, adhering to best practices to ensure reproducibility. It includes reading quality control modules (FastQC and FastP), genome assembly (Unicycler and Shovill), and bacterial genome annotation (Prokka), as well as antibiotic resistance gene and virulence gene analysis (ARIBA, Abricate, and Mykrobe). IQ-TREE and MASH were also used for phylogenetic tree construction.

Results: *StaPhyloRes's* pipeline demonstrated an efficient performance in a high-performance computing system (HPC), generating reproducible and high-quality results across each module. Unicycler was prioritized for its higher assembly quality. Results from molecular typing (MLST, *spa* type, *SCCmec* type, and *agr* type) were compiled into a local summary for streamlined visualization, and antibiotic resistance and virulence gene analysis results were summarized by tool and database used. Finally, quality control and assembly data were integrated and consolidated into a single visual report via MultiQC.

Conclusions: *StaPhyloRes* is a valuable and accessible tool for studying *S. aureus genomes*, enabling the identification of antibiotic resistance and virulence genes, providing a foundation for epidemiological surveillance studies as well. Its modular design eases the adaptation for other pathogens and bioinformatics analysis environments, adding value to hospital bacterial infections prevention efforts.

Keywords: *Staphylococcus aureus*, antibiotic resistance, *Nextflow*, molecular typing, phylogenetic analysis, bioinformatics.

Índice

Lista de abreviaciones

1. AGR - Accessory Gene Regulator
2. AENOR - Asociación Española de Normalización y Certificación
3. bp - Base Pairs
4. CARD - Comprehensive Antibiotic Resistance Database
5. CNM - Centro Nacional de Microbiología
6. CPU - Central Processing Unit
7. CSV - Comma-Separated Values (valores separados por comas)
8. CSVTK - Comma Separated Values Tool Kit
9. ECDC - European Centre for Disease Prevention and Control
10. ET - Toxinas Exfoliativas
11. GC - Guanine-Cytosine (contenido de guanina y citosina en el ADN)
12. GB - Gigabyte
13. GFF - General Feature Format (formato de anotación genómica)
14. Gubbins - Genealogies Unbiased By recomBinations In Nucleotide Sequences
15. HPC - High-Performance Computing
16. ISCIII - Instituto de Salud Carlos III
17. ISO - International Organization for Standardization
18. LPV - Leucocidina de Panton-Valentine
19. MLST - Multilocus Sequence Typing
20. MRSA - Methicillin-resistant Staphylococcus aureus

21. OMS - Organización Mundial de la Salud
22. PFGE - Pulse Field Gel Electrophoresis
23. PRAN - Plan Nacional frente a la Resistencia a los Antibióticos
24. QUAST - Quality Assessment Tool for Genome Assemblies
25. RAM - Random Access Memory
26. RedLabRA - Red de Laboratorios para la Vigilancia de los Microorganismos Resistentes a los Antibióticos
27. SARM - Staphylococcus aureus resistente a meticilina
28. SCCmec - Staphylococcal Cassette Chromosome mec
29. SNP - Single Nucleotide Polymorphism
30. StaPhyloRes - Staphylococcus Phylogenetic Resistance
31. ST - Secuencia Tipo
32. VFDB - Virulence Factor Database
33. WGS - Whole-Genome Sequencing

1. Introducción

1.1.Staphylococcus aureus

Staphylococcus aureus es una bacteria que forma parte de las cocáceas Gram positivas, inmóvil, no esporulada y generalmente no encapsulada, aunque algunas cepas pueden producir una cápsula mucoide. Respecto a su metabolismo se caracteriza por ser anaerobio facultativo con producción de catalasa y coagulasa, pero negativo para oxidasa, con un crecimiento óptimo entre 18 y 40°C y amplios rangos de pH [1]. Este patógeno origina un gran número de enfermedades y es uno de los principales agentes causales de infecciones que van desde infecciones de piel y tejidos blandos, neumonías, infecciones osteoarticulares, infecciones endovasculares y meningitis. Las infecciones invasoras por *Staphylococcus aureus* se han asociado a una mayor mortalidad, mayor estancia hospitalaria y mayores costes sanitarios [1–3].

1.2.Staphylococcus aureus resistente a meticilina

S. aureus tiene una notable capacidad para adaptarse ante la presencia de los antibióticos usados para combatir las infecciones causados por este mismo. De modo que ya en la década los 40, con el uso de la penicilina para tratar infecciones por estafilococos, en un año de uso clínico ya se habían desarrollado e identificado cepas de *S. aureus* resistentes a este antibiótico. Resistencia producida por penicilinasas, las cuales corresponde a enzimas β-lactamasas codificadas por el gen *blaZ*, que pueden hidrolizar la penicilina. En la actualidad, más del 90% de las cepas de *S. aureus* poseen este mecanismo de resistencia [4].

La resistencia a la meticilina en *S. aureus* se debe a la producción de una proteína de unión a penicilina (PBP2a) de baja afinidad para antimicrobianos β-lactámicos, codificada por el gen *mecA*. Este gen se encuentra en una sección móvil del cromosoma de *S. aureus* llamada casete cromosómico estafilocócico *mec* (SCC*mec*), que es muy diverso y tiene un tamaño variable (tipos I a XI), cuya longitud está relacionada con el origen clonal de las cepas [5,6]. Las cepas de *S. aureus* que contienen el gen *mecA* son resistentes a todos los β-lactámicos, excepto a dos nuevas cefalosporinas: ceftobiprole

y ceftarolina. Esto representa un importante desafío terapéutico sumado a la capacidad de los estafilococos para adquirir resistencia a otros antibióticos [7].

1.3. Tipificación Molecular

La tipificación molecular de *Staphylococcus aureus* es realizada mediante diferentes técnicas con la finalidad de entender la epidemiología de *Staphylococcus aureus*, conocer clones predominantes y/o transmisibilidad para poder mantener una vigilancia y control frente a la diseminación del microorganismo.

1.3.1. Multilocus Sequence Typing (MLST)

El análisis consiste en examinar la secuencia de nucleótidos de fragmentos internos de ciertos genes conservados "housekeeping" que codifican enzimas metabólicas. Este análisis permite detectar variaciones neutras que definen líneas clonales relativamente estables. En el caso de *S. aureus*, se identifican a lo menos siete genes "housekeeping": *arcC* (carbamato quinasa), *aroE* (shikimato deshidrogenasa), *glpF* (glicerol quinasa), *gmk* (guanilato quinasa), *pta* (fosfato acetiltransferasa), *tpi* (triosafosfato isomerasa) y *yqiL* (acetil coenzima A acetiltransferasa). La secuencia de estos genes se compara con los alelos conocidos disponibles en la página web www.mlst.net, y el perfil alélico obtenido permite determinar la secuencia tipo (ST) de la cepa [8].

1.3.2. Spa-typing

La técnica de *spa*-typing utiliza la secuencia de un VNTR polimórfico en la región X codificante de la proteína A específica (*spa*) de *S. aureus*. La región X se caracteriza por un número variable (entre 3 y 15) de pequeñas repeticiones. Cada nueva composición de base del repetido polimórfico encontrada en una cepa se asigna un código único. La secuencia de repetición para una cepa determinada determina su *spa*-tipo. Aunque la tipificación *spa* es una técnica de tipificación de locus único, ofrece una resolución de subtipo comparable a técnicas más costosas y/o laboriosas como MLST y electroforesis de campo pulsado (*Pulse Field Gel Electrophoresis*, PFGE. La tipificación *spa* es suficientemente estable para la tipificación epidemiológica de cepas de *Staphylococcus*

aureus resistente a meticilina (SARM). Los resultados sugieren que los brotes hospitalarios pueden ser causados por dos o más cepas de SARM, haciendo de la tipificación *spa* una herramienta importante para entender la diseminación de SARM dentro y entre hospitales [9,10].

1.3.3. SCCmec

El análisis del polimorfismo del SCC*mec*, es una técnica basada en el análisis del polimorfismo asociado al SCC*mec* que contiene al gen *mecA*.

La importancia de conocer la constitución del SCC*mec* radica en que, a partir de los eventos de recombinación entre los genes *ccr* y *mecA*, se ha generado una variedad de SCC*mec* que permite clasificar al *Staphylococcus aureus* resistente a meticilina (SARM) según el tipo de SCC*mec* que posee. Inicialmente se describieron cinco tipos de SCC*mec* (I-V). Sin embargo, actualmente ya se cuenta con cuatro tipos extra como SCC*mec* VI-XI.

Además, los SCC*mec* se diferencian entre sí por sus determinantes de resistencia. Los SCC*mec* I, IV, V, VI y VII codifican exclusivamente para resistencia a los antibióticos betalactámicos, mientras que los SCC*mec* II, III y VIII poseen genes adicionales que confieren resistencia a múltiples clases de antibióticos, además de los antibióticos betalactámicos [11,12].

1.3.4. Locus agr

El locus del regulador de genes accesorios (*agr*) de *S. aureus* es un grupo de genes de *quorum sensing* compuesto por cinco genes (*hld, agrB, agrD, agrC y agrA*) que aumenta la producción de factores de virulencia secretados, incluyendo las hemolisinas alfa, beta y delta, y reduce la producción de factores de virulencia asociados a la célula [13,14].

Las cepas de *S. aureus* se pueden clasificar según el tipo de locus *agr* que posean, existiendo cuatro tipos diferentes en esta especie. Se ha observado que las cepas dentro

de un mismo grupo pueden activar la respuesta *agr* en otras cepas del mismo grupo e inhibir la respuesta en cepas de otros grupos [15].

1.4. Factores de Virulencia

1.4.1. Leucocidina de Panton-Valentine

Una de las toxinas más virulentas y conocidas es la Leucocidina de Panton-Valentine (LPV). Solo el 2-3% de las cepas de *S. aureus* producen esta leucotoxina. Este factor es una citotoxina que se une a los fosfolípidos de la membrana de los leucocitos y macrófagos, induciendo la formación de poros que alteran la permeabilidad celular, provocando la destrucción de leucocitos y necrosis tisular [16]. Se estipula que esta citotoxina es la responsable de causar desde infecciones leves en la piel y tejidos blandos hasta enfermedades invasivas como neumonía necrotizante, sepsis severa y fascitis necrotizante, que ponen en riesgo la vida del individuo [17].

1.4.2. Toxinas Exfoliativas

Las toxinas exfoliativas (ET) de *S. aureus* son proteínas extracelulares responsables en humanos (mediante toxinas A y B) de las manifestaciones cutáneas de la enfermedad denominada impétigo bulloso y del síndrome estafilocócico de la piel escaldada [18].

1.4.3. Hemolisinas

Se han descrito cinco hemolisinas: alfa, beta, gamma, delta y una variante de gamma. Estas toxinas están generalmente presentes en la mayoría de las cepas de *S. aureus* [19].

1.5. Importancia de la vigilancia activa de infecciones por Staphylococcus aureus basada en secuenciación genómica

Desde al año 2021, la Organización Mundial de la Salud (OMS) ha definido una serie de microrganismos con tres niveles de prioridad en salud pública, que va desde el nivel Crítico, Elevado y Medio, encontrando el patógeno *Staphylococcus aureus* resistente a la meticilina, en el encuadre de prioridad elevada, motivo por el cual es un patógeno de interés en salud publica debido a su amplio espectro de asociaciones a diferentes infecciones mencionadas anteriormente [20].

El ECDC (del inglés *European Centre for Disease Prevention and Control*) propone la priorización en la implementación la secuenciación de genoma completo para su uso en las investigaciones de brotes y vigilancia de la salud pública, con tal de mejorar la prevención y control de las enfermedades [21].

Dentro de la lista de patógenos y enfermedades prioritarias para una integración a medio plazo (2019-2021) encontramos al *Staphylococcus* aureus resistente a la meticilina, priorizando su confirmación temprana para brotes multinacionales, identificación de los vectores genéticos que pueda presentar, entre otras medidas solicitadas [22].

Manteniendo los objetivos tanto de la OMS, la ECDC y del plan nacional de España frente a la resistencia a los antibióticos (PRAN), se vio la necesidad de la implementación de una red de laboratorios para la vigilancia de microorganismos resistentes, iniciando RedLabRA, la cual es una red de laboratorios de microbiología en España, coordinada e interconectada a nivel nacional, que trabaja en el diagnóstico y estudio molecular de microorganismos resistentes a antibióticos. Esta red responde a la necesidad de abordar el creciente impacto clínico y epidemiológico de los microorganismos multirresistentes, entre ellos SARM, y quiere estandarizar los procedimientos de detección y caracterización de los mecanismos de resistencia[23].

1.6.Secuenciación de Genoma Completo

La secuenciación del genoma completo (WGS, del inglés *whole-genome sequencing*) es una tecnología en rápida evolución, en donde gracias a la tecnología de secuenciación de segunda generación ha podido reducir costos y acercar esta tecnología en los laboratorios clínicos de rutina [24]. Es posible secuenciar muestras directamente en pocas horas, lo que trae grandes mejoras en el diagnóstico y trae consigo nuevos retos en el manejo de la información de valor generada por estos mismos análisis. Dentro de esa nueva información obtenida, trae el incremento de la rapidez en la detección de resistencia a los antibióticos para diversos patógenos de importancia sanitaria, además de poder ofrecer información valiosa para ser usada en vigilancia epidemiológica de los mismos [25].

1.7.NextFlow y nf-core

En análisis de grandes volúmenes de datos en bioinformática ha impulsado el desarrollo de herramientas capaces de manejar flujos de trabajo complejos de una forma totalmente reproducible y escalable. Para dar respuesta a este desafío surge *nextflow*, ofreciendo una plataforma flexible y escalable para manejar diversas tareas computacionales en bioinformática. *Nextflow*, basado en el modelo de flujo de datos, permite a los usuarios diseñar pipelines mediante la composición de procesos individuales, que se pueden ejecutar en plataformas locales o en entornos distribuidos como clústeres de computación de alto rendimiento (HPC), servicios en la nube y herramientas de software como Docker y Conda, hace posible la creación de pipelines portátiles que pueden ser ejecutados en una amplia variedad de entornos sin necesidad de reconfiguraciones complejas [26].

Gracias a *nextflow* nace la comunidad de nf-*core* [27], el cual recopila un conjunto curado de pipelines bioinformáticos construidos en *nextflow*, lo que les permite ejecutarse en la mayoría de las infraestructuras computacionales actuales. La comunidad nf-*core* ha desarrollado un conjunto de herramientas que automatizan la creación, prueba, implementación y sincronización de pipelines. La comunidad *nf*-core

se encarga de entregar el marco para las mejores prácticas de desarrollo, estandarización y documentación de código. La naturaleza *open source* de esta comunidad facilita el intercambio de ideas y colaboración sobre las mejores prácticas e implementación de pipelines específicos, con la participación de expertos de todo el mundo.[28,29]

1.8. Finalidad y Producto Esperado

La finalidad de este TFM tiene como objetivo realizar el análisis bioinformático y obtener información clínica relevante sobre los datos de genoma completo de cepas de *Staphylococcus aureus* provenientes de infecciones invasoras en hospitales de España. Esta información, recogida a través del Programa de Vigilancia de Infecciones Estafilocócicas, coordinado desde el Centro Nacional de Microbiología, del Instituto de Salud Carlos III, se analizará mediante la creación de un pipeline basado en *Nextflow* y módulos de *nf-core*, que automatizará los análisis requeridos para la vigilancia y diagnóstico microbiológico completo de infecciones y colonizaciones causadas por *Staphylococcus aureus* específicamente y aplicable también a otras bacterias resistentes a los antibióticos.

El pipeline ***StaPhyloRes*** está diseñado para el análisis genómico de secuencias cortas de *Staphylococcus aureus*, abarcando desde el control de calidad hasta el análisis de los genes de resistencia a antibióticos, identificación de factores de virulencia, predicción de resistencia a antibióticos, tipificación molecular y la construcción de árboles filogenéticos.

2. Objetivos

2.1. Objetivo General:

Desarrollar un pipeline de bioinformática utilizando los sistemas *Nextflow* y *nf-core*, que permiten crear flujos de trabajo automatizados, para analizar genomas completos (secuencias fastq) de *Staphylococcus aureus*. Este pipeline incluye los siguientes pasos: el análisis de calidad de las lecturas crudas, el trimado, el ensamblaje del genoma y calidad del ensamblado, la identificación de genes de resistencia a antibióticos y de virulencia, la tipificación molecular específica del género bacteriano, el análisis filogenético basado en polimorfismos de un solo nucleótido (SNP, de inglés *single-nucleotide polymorphisms*) y generación de informes detallados. Además, se estudiará la predicción fenotípica de la resistencia a antibióticos de interés clínico.

2.2. Objetivos Específicos:

2.2.1. Desarrollar e implementar procesos de control de calidad para evaluar y mejorar la calidad de las secuencias crudas y ensambladas. Permitiendo la generación de informes de calidad de las lecturas de manera eficiente y sistemática, asegurando así un control riguroso y continuo de la calidad de los datos en todas las etapas del análisis.

2.2.2. Integrar herramientas avanzadas de ensamblaje de genomas en el pipeline para ensamblar de manera eficiente las secuencias cortas de ADN de *Staphylococcus aureus.* Este enfoque permitirá la obtención de genomas completos y precisos, facilitando el análisis posterior de los genes de resistencia y virulencia, así como otros estudios genómicos relevantes.

2.2.3. Tipificación molecular para *S. aureus* incluyendo: secuencio tipo (ST, del inglés *Sequence Type*) basado en el esquema de MLST (del inglés, *Multilocus Sequence Typing*), *spa*-tipo, tipo de casete cromosómico estafilocócico (SCC*mec,* del inglés *Staphylococcal Chromosomal Cassette mec*) y *agr* locus.

2.2.4. Realizar una detección precisa de genes adquiridos, casetes de resistencia a antibióticos y mutaciones cromosómicas asociadas a la resistencia a

antibióticos, así como la identificación factores de virulencia y de genes plasmídicos (replicones) relevantes en las secuencias trimadas y ensambladas de *Staphylococcus aureus*.

2.2.5. Identificar y analizar distancia entre los ensamblados genómicos y variantes de SNPs en genomas de *Staphylococcus aureus*, previa identificación de un genoma de referencia, incluyendo la construcción de árboles filogenéticos basados tanto en distancias contrastadas ante genoma de referencia común y SNPs.

2.2.6. Implementar una herramienta bioinformática para predecir la resistencia a antibióticos de interés clínico a partir de los genomas de *Staphylococcus aureus*.

2.2.7. Automatizar y optimizar el pipeline bioinformático para garantizar que se ejecute de manera eficiente y reproducible en diferentes entornos de computación, local y HPC. Incluir opciones configurables que permitan a los usuarios elegir entre diferentes bases de datos y configuraciones de análisis mediante parámetros ajustables en el pipeline, facilitando así su adaptabilidad y personalización para diversas necesidades de investigación.

3. Materiales y Métodos

3.1. Origen de los datos

Los datos genómicos seudonimizados utilizados para la validación del pipeline provienen de cepas de *S. aureus* pertenecientes a la colección del Programa de Vigilancia Microbiológica de Infecciones Estafilocócicas del Centro Nacional de Microbiología (CNM) del Instituto de Salud Carlos III (ISCIII). Todo el proceso está certificado por la Asociación Española de Normalización y Certificación (AENOR) con la norma de calidad ISO 9001 desde 2012.

3.2. Software empleado en el TFM

El pipeline se ejecutó en una infraestructura de cómputo de alto rendimiento y sigue las mejores prácticas de la plataforma *nf-core* [27], lo que garantiza la reproducibilidad y escalabilidad del análisis. El pipeline fue escrito en con el lenguaje *Nextflow* mayoritariamente. La versión de *nexflow* utilizada fue 23.04.0 [26] y se escribió utilizando el software Visual Estudio Code 1.90.0 [30].

Como herramienta de apoyo en la programación y optimización del pipeline, se utilizó la herramienta ChatGPT[31], en su versión 4o, Este modelo de inteligencia artificial (IA) facilitó el proceso de depuración de código y la identificación de errores, así como la exploración de alternativas de programación para mejorar la eficiencia y funcionalidad del pipeline.

El desarrollo básico de este pipeline fue adaptado para ser usado tanto en computadores convencionales como clúster de computación, tomando los siguientes requerimientos como mínimos: 16 CPU y 12 GB de memoria. Se ha configurado cada proceso siguiendo las recomendaciones de *nf-core* definidas en su configuración base, siendo utilizado por cada proceso un máximo de 2 CPU y 12 GB de memoria excepto módulos de ensamblado, los cuales se ha definido un máximo de 6 CPU, 128 GB de memoria y un tiempo límite de 78 horas para aquel modulo. Para el uso en clúster de computación se ha realizado modificaciones a los recursos máximos, definiendo una

memoria máxima de 128 BG, 16 CPU y un tiempo de ejecución máxima para el pipeline de 120 horas.

El análisis bioinformático se ejecutó en un computador con las siguientes especificaciones: dispositivo de escritorio, procesador *AMD Ryzen 7 5800X 8-Core Processor* a 3.80 GHz, con *32 GB de RAM*. El sistema operativo utilizado fue de *64 bits* y el procesador basado en arquitectura *x64*. Este equipo proporcionó la capacidad computacional necesaria para ejecutar los distintos módulos del pipeline, permitiendo la paralelización eficiente de los procesos.

El uso de *Nextflow* permite escalar cualquier pipeline en un clúster de alta computación mediante la integración de *Slurm*, una herramienta altamente configurable que facilita la asignación personalizada de recursos. *Slurm*, permite definir parámetros como el número de nodos, la cantidad de tareas por nodo, y el número de CPUs por tarea, de esta forma optimizando la paralelización y reduciendo significativamente el tiempo en análisis complejos. Además, la memoria asignada puede configurarse por CPU, ajustando dinámicamente este valor según la cantidad de CPUs disponibles, lo que permite un uso eficiente de los recursos.

Otro aspecto configurable es el tiempo máximo de ejecución del pipeline en el entorno HPC, el cual puede extenderse independientemente del tiempo predeterminado en el pipeline, adaptándose a las necesidades específicas de cada etapa del análisis.

Esta capacidad de configuración garantiza la flexibilidad necesaria para adaptarse a distintas cargas de trabajo y permite realizar ajustes específicos en tareas individuales, como incrementar la cantidad de CPUs en procesos intensivos, asegurando así la escalabilidad y estabilidad del pipeline en diversos escenarios de análisis.

3.3. Validación del pipeline

Para la validación del pipeline en cada proceso se ha realizado una selección de 30 secuencias de genoma completo provenientes de cepas de *S. aureus* pertenecientes a

la colección del Programa de Vigilancia Microbiológica de Infecciones Estafilocócicas del Centro Nacional de Microbiología (CNM) del Instituto de Salud Carlos III (ISCIII), comprimidas en formato FASTQ (.fastq).

3.4. Preparación de datos

La preparación de los datos fue un paso clave en el proceso de análisis, y consistió en la creación de un archivo separado por comas (.csv) siguiendo las recomendaciones establecidas por la comunidad de *nf-core* [29] para garantizar la correcta ejecución de los pipelines bioinformáticos. Este archivo debe adoptar el formato estipulado por *nf-core*, en el que cada separación contiene los siguientes campos: sample, fastq_1 y fastq_2. La columna "*sample*" contiene la información de identificación única para cada genoma, mientras que las columnas fastq_1 y fastq_2 corresponden a las secuencias obtenidas de la secuenciación de ADN en formato FASTQ, representando las lecturas R1 y R2, respectivamente, que derivan de la secuenciación de extremos emparejados (*paired-end sequencing*).

3.5. Pre-procesamiento de lecturas crudas

El *subworkflow* encargado de realizar este paso se encuentra disponible en la comunidad *nf-core* [29] bajo el nombre de "fastq_trim_fastp_fastqc", y fue utilizado como inicio del pipeline al contener los pasos necesarios ya establecidos para el análisis requerido. Este *subworkflow* integra herramientas para el control de calidad y procesamiento de las lecturas crudas (fastq), optimizando el pre-procesamiento de datos de secuenciación.

Como *input* se ha utilizado el archivo CSV preparado con anterioridad, el cual contiene las rutas absolutas a las lecturas crudas (fastq) que se van a analizar. Este análisis inicial consistió en evaluar parámetros de calidad de secuenciación utilizando la herramienta FastQC v0.12.1 [32] con sus parámetros predeterminados. En este paso es generado un informe de calidad individualizado para cada lectura cruda (*forward* y *reverse*) para luego ser analizado.

Posteriormente, se realizó el trimado de las lecturas para eliminar adaptadores y bases de baja calidad mismas mediante la herramienta FastP v0.23.4 [33,34], mejorando de esta forma la calidad de las lecturas crudas al eliminar regiones no deseadas. Además,

se realizó un filtrado adicional de las secuencias trimadas para evitar fallos en los procesos posteriores.

Luego, cada lectura es evaluada nuevamente mediante herramienta FastQC para comprar métricas pre y post trimado de secuencias. Esta segunda evaluación permitió verificar la mejora de las lecturas en términos de calidad y la eliminación exitosa de regiones problemáticas.

Finalmente, las lecturas preprocesadas fueron almacenadas en un canal denominado *ch_trim_fastp*, el cual luego fue transformado y reasignado al canal *ch_unicycler* para ser utilizados en el siguiente nodo.

3.6. Ensamblaje y anotación de genomas bacterianos

Para el módulo de ensamblado de genomas se ha utilizado Unicycler v0.5.0 [35] y Shovill v1.1.0 [36], ejecutado con sus parámetros por defecto y sin utilizar un genoma de referencia, lo que permite un ensamblaje de *novo* adecuado para genomas bacterianos obtenidos a partir de secuencias cortas de Illumina.

Se realizaron pruebas locales comparativas entre Unicycler y la alternativa Shovill, para de esta forma definir la herramienta principal necesaria para la ejecución de este módulo. Shovill también utiliza SPAdes [37] en su núcleo, pero optimiza el preprocesamiento y la paralelización para obtener ensamblajes más rápidos, por lo cual era imperativo realizar una comparativa antes de definir el uso de alguna de estas herramientas.

Los ensamblajes generados fueron almacenados en un canal denominado *ch_assembly_read* el cual es utilizado dentro del pipeline para proveer el genoma ensamblado a los nodos siguientes que requieren este *input.*

Finalmente en paralelo se procede a realizar la anotación del genoma ensamblado utilizando Prokka v1.14.6 [38], generando archivos de anotación en formato *GFF* y *GenBank*, listos para ser utilizados en posteriores análisis comparativos o estudios de funcionalidad genómica.

Prokka se ejecutó con parámetros predeterminados, y se logró una anotación completa de las secuencias ensambladas, proporcionando una visión detallada de los genes presentes en las cepas de *Staphylococcus aureus* analizadas.

3.7. Predicción fenotípica de resistencia a antibióticos

Este módulo se encontró conformado por la herramienta Mykrobe v0.11.0 [39], la cual cuenta con una base de datos específica para *Staphylococcus aureus*. Mykrobe permite predecir la resistencia a antibióticos a partir de secuencias cortas de ADN, generando predicciones sobre la susceptibilidad o resistencia a los fármacos más comúnmente utilizados en la práctica clínica. Entre los antibióticos evaluados se encuentran: ciprofloxacina, clindamicina, eritromicina, ácido fusídico, gentamicina, meticilina, mupirocina, penicilina, rifampicina, tetraciclina, trimetoprima y vancomicina.

3.8. Identificación de genes de resistencia a antibióticos y virulencia

Para el desarrollo de este módulo se ha implementado el uso de tres herramientas que, en conjunto, permiten realizar una identificación exhaustiva de genes de resistencia a antibióticos y factores de virulencia.

En primera instancia, se ha utilizado la herramienta ARIBA v2.14.6 [40], que realiza el análisis de secuencias cortas preprocesadas y trimadas para la búsqueda de genes de resistencia a antibióticos y factores de virulencia en bases de datos especializadas, tales como ResFinder [41], PlasmidFinder [42], CARD [43] y VFDB [44]. Esta herramienta no solo identifica la presencia de genes asociados a resistencia y virulencia, sino que también localiza los *contigs* en los cuales se encuentran estos genes, lo cual es crucial para conocer el entorno genético de estos genes en estudios de epidemiología molecular y vigilancia de la resistencia a antibióticos.

La segunda herramienta utilizada es Abricate v1.0.1 [45], la cual se emplea para contrastar los hallazgos obtenidos a partir de los ensamblados generados en etapas previas. Abricate utiliza las mismas bases de datos, ResFinder y VFDB, para realizar una verificación adicional de los genes de resistencia a antibióticos y virulencia. Esta herramienta permite además crear bases de datos propias para el *screening* de genes

de interés. De forma complementaria, se ha utilizado una base de datos personalizada, denominada "*Staph*", que contiene información más específica sobre los factores de virulencia en *Staphylococcus aureus*, en comparación con la base de datos estándar VFDB.

Finalmente, como complemento al análisis basado en ensamblados, se ha utilizado STARAMR v0.10.0 [46]. Esta herramienta realiza un análisis adicional utilizando bases de datos faltantes como PointFinder [47] y PlasmidFinder, además de identificar variantes puntuales que podrían estar relacionadas con la resistencia. Adicionalmente, STARAMR ofrece un análisis de tipificación molecular mediante MLST (Multilocus Sequence Typing) [48], proporcionando información valiosa sobre los linajes bacterianos y la posible diseminación de clones resistentes.

3.9.Análisis filogenético

Para el módulo de análisis filogenético, se ha implementado un enfoque escalonado que comienza con la identificación de las especies y continúa con la selección automática de un genoma de referencia adecuado, seguido de un análisis de las distancias filogenéticas basado en SNP.

En primer lugar, se ha utilizado MASH v2.3 [49] para comparar nuestras secuencias ensambladas de *Staphylococcus aureus* con una base de referencia proporcionada por los desarrolladores de la herramienta. Este paso nos permite confirmar la identificación de la especie y detectar posibles contaminaciones con otras especies bacterianas presentes en las muestras.

En segundo lugar, se realiza una búsqueda y recuento de los genomas de referencia que concuerdan con cada secuencia ensamblada, permitiendo seleccionar automáticamente el genoma de referencia más representativo para el conjunto de secuencias de la corrida. Este genoma de referencia se descarga de manera automatizada mediante la herramienta NCBIGENOMEDOWNLOAD v0.3.3 [50], seleccionando el archivo en formato *GeneBank* para su posterior uso en el análisis de

variaciones de nucleótido único (SNP) con Snippy v4.6.0. En este análisis, se emplean las secuencias trimadas para identificar variaciones genéticas precisas entre las cepas.

El análisis de Snippy se enfoca en el "*core*" del genoma, lo que permite realizar la búsqueda de distancias filogenéticas basadas en *SNPs*. Para una visualización más clara y personalizable de los resultados, se utiliza la herramienta IQ-TREE v2.3.0 [51], que permite generar árboles filogenéticos al alineamiento de las secuencias con *SNPs* identificados para cada genoma, proporcionando una representación visual del parentesco filogenético entre las diferentes cepas analizadas.

Opcionalmente, se puede utilizar la herramienta Gubbins v.3.0.0 [52] para un análisis filogenético más detallado basado en *SNPs*, con el fin de corregir posibles recombinaciones genéticas, lo que puede mejorar la precisión del árbol filogenético en determinadas especies bacterianas y según la heterogeneidad de la población de estudio.

Finalmente, se puede realizar un análisis complementario de la distancia filogenética sin la inclusión de un genoma de referencia, utilizando la herramienta MASHTREE v1.2.0 [53], la cual construye árboles filogenéticos basados en distancias de MASH, permitiendo explorar relaciones evolutivas directamente entre las secuencias ensambladas.

3.10. Tipificación molecular

En el módulo de tipificación molecular se ha implementado la identificación de marcadores genéticos específicos en *Staphylococcus aureus*, utilizando varias herramientas especializadas que permiten el análisis de diferentes loci asociados.

En primer lugar, se ha empleado la herramienta MLST v2.23.0 [48] para llevar a cabo la tipificación por *Multilocus Sequence Typing* (MLST). Este método permite clasificar las cepas bacterianas en función de las variaciones en un conjunto de siete genes de referencia altamente conservados, proporcionando una asignación de los diferentes linajes o secuencias tipo (ST) de *Staphylococcus aureus*.

Seguidamente, se utilizó Spatyper v0.3.3 [54], una herramienta especializada en la identificación del *spa*-tipo. El *spa typing* es un método que se enfoca en la presencia de secuencias cortas repetitivas en la región *spa* del genoma de *Staphylococcus aureus.*

Asimismo, se incluyó el análisis de la región *SCCmec* mediante la herramienta Spathopia-sccmec v1.0.0 [55], la cual permite detectar y clasificar los distintos tipos de *SCCmec* presentes en las cepas de *Staphylococcus aureus* para identificar variantes del genoma relacionadas con la resistencia a meticilina y de esta forma realizar una caracterización de cepas MRSA.

Además, se ha utilizado Agrvate v1.0.2 [56] para analizar el locus *AGR*, el cual regula la virulencia en *Staphylococcus aureus*. Esta herramienta permite determinar el tipo de *agr* (I, II, III, IV) presente en cada cepa, proporcionando información relevante sobre la capacidad de las cepas para causar infecciones severas.

3.11. Informe de calidad

En el módulo final del pipeline, se ha implementado la herramienta MultiQC v1.16 [57], cuya función es consolidar en un único informe en formato HTML todos los informes de calidad generados a lo largo del análisis. MultiQC recopila y resume los resultados de las evaluaciones de calidad de las secuencias crudas, las secuencias trimadas y los ensamblados genómicos. Este enfoque permite visualizar de manera centralizada todos

los parámetros de calidad de estos tres procesos sin necesidad de cambiar de pantalla o analizar múltiples archivos individualmente.

4. Resultados y Discusión

Como resultado de este trabajo, se desarrolló un pipeline bioinformático modular y flexible denominado ***StaPhyloRes***, diseñado en *Nextflow* para facilitar la detección de factores de virulencia, resistencia a antibióticos, construcción de árboles filogenéticos y otros análisis moleculares, en cepas de *Staphylococcus aureus*. Este pipeline, disponible en un repositorio público de *GitHub*, garantiza su accesibilidad y reproducibilidad, permitiendo que otros investigadores puedan implementarlo y adaptarlo fácilmente para sus propios estudios.

En la actualidad, existe un desarrollo continuo de pipelines que integran diversas tareas y procesos bioinformáticos, tanto generales como específicos para diferentes tipos de organismos, tales como genomas bacterianos, detección de plásmidos, y análisis de secuencias virales. En este contexto, existen herramientas como ***BACTOPIA*** [58], un pipeline completo para análisis de genomas bacterianos que sirvió como fuente de inspiración para la construcción de ***StaPhyloRes***. Al igual que ***BACTOPIA***, ***StaPhyloRes*** ofrece una arquitectura modular y permite abordar diferentes aspectos genéticos y biológicos de *S. aureus*, proporcionando una base sólida para realizar estudios de epidemiología molecular, filogenia y vigilancia genética.

El código de ***StaPhyloRes*** está documentado y organizado siguiendo las mejores prácticas recomendadas por la comunidad *nf-core*, lo cual asegura una estructura de alta calidad, fácil mantenimiento y adaptabilidad. Esta organización permite su uso tanto en infraestructuras de cómputo convencionales como en plataformas de alto rendimiento (HPC), donde puede aprovechar al máximo los recursos computacionales para procesar grandes volúmenes de datos.

El repositorio, disponible en https://github.com/VictorPizarroR/StaPhyloRes, incluye documentación completa sobre la instalación, configuración y uso del pipeline. Esta documentación facilita su implementación y asegura que los usuarios dispongan de

toda la información necesaria para realizar un análisis genómico detallado de *Staphylococcus aureus.*

4.1.Requerimientos Básicos

Para ejecutar ***StaPhyloRes*** de manera óptima, se deben cumplir los siguientes requisitos de software y hardware:

- Sistema Operativo: Linux o macOS son recomendados, especialmente en entornos de alta computación (HPC) donde el rendimiento es crucial.
- Gestor de flujo de trabajo: Nextflow, versión 20.10.0 o superior, que coordina la ejecución y administración de los procesos del pipeline.
- Conda o Docker: Herramientas para la instalación y administración de entornos aislados. Conda permite la creación de entornos virtuales según las dependencias especificadas en el archivo YML, mientras que Docker facilita la ejecución en contenedores para mejorar la portabilidad.
- Slurm: Requerido si el pipeline se ejecuta en un clúster HPC que utiliza Slurm como gestor de recursos.
- Dependencias del Pipeline: Todas las herramientas necesarias están especificadas en el archivo YML para Conda y en los archivos Dockerfile incluidos en el repositorio, asegurando que el pipeline pueda instalarse y ejecutarse en diversos entornos sin conflictos de dependencias.
- CPU:

Mínimo: 4 núcleos (para pruebas o conjuntos de datos pequeños).

Recomendado: 16 núcleos o más en sistemas de HPC para optimizar el tiempo de procesamiento de grandes conjuntos de datos.

- Memoria RAM:

Mínimo: 16 GB.

Recomendado: 64 GB o más, especialmente para tareas de ensamblaje y análisis filogenético, que pueden consumir grandes cantidades de memoria.

- Espacio para procesamiento y salida:

Mínimo: 100 GB para ejecuciones pequeñas.

Recomendado: 500 GB o más para conjuntos de datos grandes, debido a la creación de archivos temporales, ensamblajes intermedios y resultados finales.

- Archivos temporales: Durante la ejecución, el pipeline generó archivos temporales de gran tamaño, por lo que se recomienda un espacio adicional de al menos 100 GB en el directorio de trabajo (*workdir*).

4.2.Flujo de Trabajo

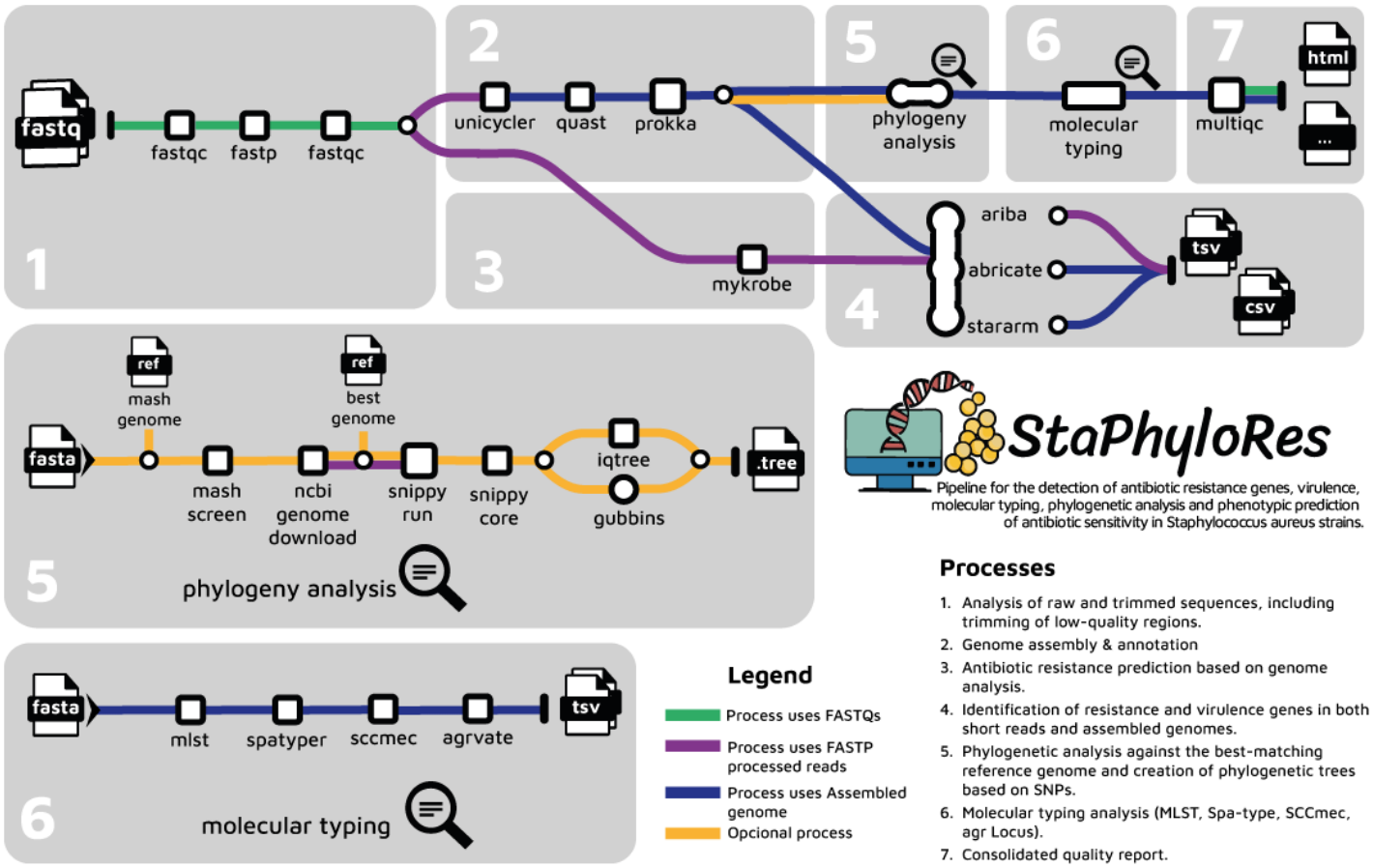

Figura 1. Diagrama del pipeline bioinformático ***StaPhyloRes*** para la detección de genes de resistencia a antibióticos, factores de virulencia, tipificación molecular, y análisis filogenético en cepas de *Staphylococcus aureus.* El flujo de trabajo se divide en siete módulos principales: (1) análisis de secuencias crudas y trimadas, (2) ensamblaje y anotación del genoma, (3) predicción de resistencia a antibióticos, (4) identificación de genes de resistencia a antibióticos y de virulencia, (5) análisis filogenético, (6) tipificación molecular, y (7) consolidación de informes de calidad. Las líneas de colores indican los diferentes tipos de procesos utilizados: secuencias FASTQ, secuencias trimadas, secuencias ensambladas y procesos opcionales.

4.3. Desarrollo y pruebas

Para facilitar la etapa de creación del archivo de entrada, se desarrolló el script en Bash "Crear_CSV.bash", el cual toma como argumento el directorio que contiene los archivos FASTQ comprimidos. El script genera automáticamente un archivo CSV en el directorio de ejecución, llamado "prepared_input.csv". El funcionamiento del script consta de la extracción del *sample ID* contenido en el nombre del archivo y posteriormente emparejando las lecturas R1 y R2 de cada muestra.

El script fue diseñado específicamente para secuencias que cumplen con el siguiente formato de nombre de archivo: "ID_SXX_R1_XXX.fastq" y "ID_SXX_R2_XXX.fastq". Esto garantiza que los archivos estén correctamente identificados y organizados en el archivo CSV, que servirá como entrada en el pipeline.

A continuación, se muestra el comando de ejecución utilizado (Comando de Ejecución 1), en el cual se especifica la ruta al directorio de secuencias mediante inputdir/. Esta automatización en la generación del archivo de entrada optimiza el flujo de trabajo y minimiza los errores derivados del manejo manual de grandes volúmenes de datos de secuenciación.

```
bash
./Crear_CSV.bash inputdir/
```

Comando de Ejecución 1: Ejemplo de ejecución del script "Crear_CSV.bash" con el directorio de entrada.

Una vez finalizado el preprocesamiento, el siguiente paso es el ensamblado de secuencias. Dada la variedad de herramientas disponibles para este propósito, se decidió realizar una comparación entre las opciones seleccionadas.

Los resultados del análisis mostraron que Shovill redujo considerablemente el tiempo de análisis en un 83.75% en comparación con Unicycler **(Tabla 1).** Esta reducción en el tiempo de ejecución convierte a Shovill en una opción atractiva para estudios en los que la rapidez es una variable crucial, especialmente cuando se manejan grandes volúmenes de datos o cuando se requieren resultados rápidos para aplicaciones clínicas.

Sin embargo, tras evaluar la calidad de los ensamblajes mediante QUAST v5.2.0 [59], se observó una notable diferencia en las métricas de calidad entre los ensamblajes generados por Shovill y Unicycler **(Tabla 1).** Los resultados indicaron que Shovill genera ensamblajes más contiguos con un mayor N50, lo que sugiere un ensamblaje más rápido y compacto, mientras que Unicycler presenta un mayor número de *contigs* grandes, lo que puede reflejar una mayor capacidad para ensamblar regiones genómicas complejas, las cuales son aquellas con secuencias repetidas y comúnmente relacionadas con la diseminación de genes, por ejemplo, plásmidos o casetes cromosómicos.

Considerando que la calidad del ensamblaje es importante para los objetivos de este trabajo, como es la detección de genes y estudio filogenético, se decidió mantener Unicycler en el flujo de trabajo, priorizando de esta forma la calidad sobre el tiempo de ejecución.

Métrica	Shovill v1.1.0	Unicycler v0.5.0
Tiempo de ejecución	3' 15''	20' 00''
# contigs (>= 1000 bp)	28	33
# contigs (>= 5000 bp)	23	25
# contigs (>= 50000 bp)	13	14
Contig más grande (bp)	710,547	710,324
Longitud total (bp)	2,829,276	2,820,945
N50 (bp)	243,770	164,468
auN (tamaño promedio de contig)	308,463.70	292,256.60
L50 (número de contigs en el 50% del genoma)	4	4
% GC	32.77%	32.75%
N's por 100 kbp	0	0

Tabla 1: Comparación de métricas de ensamblaje y tiempo requerido entre Shovill v1.1.0 y Unicycler v0.5.0.

Dado que la elección de un genoma de referencia adecuado es fundamental para los análisis filogenéticos, se presentó el desafío de determinar cuál secuencia utilizar para este propósito. Para resolverlo, se decidió realizar una búsqueda comparativa

basada en distancias mash entre cada aislado en estudio y la base de datos de MASH. Con esta finalidad, se adaptó el script "common_mash_reference.py" [60], el cual permite automatizar la búsqueda del genoma de referencia más representativo entre todas las cepas analizadas.

Este proceso se basa en un filtro que selecciona únicamente aquellos resultados que corresponden a genomas completos, excluyendo la presencia de secuencias relacionadas con fagos o secuencias obtenidas por secuenciación del tipo *shotgun*. Los resultados son ordenados en forma descendente según el porcentaje de concordancia, y se selecciona la primera línea, que corresponde al genoma con la mayor similitud respecto a las secuencias analizadas.

Simultáneamente, se genera un segundo archivo que incluye resultados sin aplicar el filtro de genoma completo, lo que permite obtener una lista más amplia de posibles candidatos a genoma de referencia. Al final de este paso, se generan dos archivos, cada uno con los posibles genomas candidatos a ser seleccionados como referencia para cada secuencia en estudio.

Posteriormente, todos los archivos generados se procesan para crear un ranking que cuantifica la frecuencia de coincidencia de cada genoma en las comparaciones, seleccionando finalmente el genoma más recurrente como el genoma de referencia. Este genoma es almacenado en el canal *ch_final_genome* y, a continuación, descargado en formato *GeneBank* (*gbk*) mediante la herramienta NCBIGENOMEDOWNLOAD, quedando listo para ser utilizado en las siguientes etapas del análisis.

4.4. Estructura de carpeta de resultados

La estructura de la carpeta de resultados generada durante la ejecución del pipeline se presenta en un formato organizado, donde cada herramienta y proceso cuenta con un directorio dedicado. Esta organización facilita el acceso a los archivos de salida de cada etapa del análisis y mejora la trazabilidad y gestión de datos en proyectos de gran escala. A continuación, se detalla la estructura de carpetas:

```
bash
.
├── abricate
├── agrvate
├── ariba
├── fastp
├── fastqc
├── genome
├── iqtree
├── mash
├── mashtree
├── mlst
├── multiqc
├── mykrobe
├── ncbigenomedownload
├── pipeline_info
├── prokka
├── quast
├── snippy
├── spatyper
├── staphopiasccmec
├── staramr
├── summary
└── unicycler
```

Estructura de carpetas obtenido mediante herramienta *tree.*

4.5. Revisión de resultados obtenidos

Durante la revisión de los archivos generados, se corroboró que estos siguen una estructura organizada en el formato: "Nombre de Herramienta > SampleID > Outputs". Este orden es útil para la mayoría de los análisis y para la visualización de resultados. Sin embargo, en los análisis de tipificación molecular, y en las comparativas de resistencia y virulencia, mantener los resultados de forma aislada presenta un desafío a la hora de tabular los datos, especialmente cuando los identificadores tienen pequeñas variaciones.

Para abordar esta problemática, se realizaron modificaciones en los scripts de los módulos de ARIBA, ABRICATE y STARARM, además de la implementación de la herramienta CSVTK_CONKAT v0.30.0 [61]. Estas adaptaciones permiten que cada herramienta mantenga una estructura de salida homogénea, facilitando así las

comparaciones entre los *outputs* de cada herramienta y asegurando una organización uniforme de los resultados. Esto garantiza una mayor facilidad en el proceso de integración y tabulación de los datos.

Para reducir la cantidad de código y mantener las configuraciones de cada herramienta bien organizadas, se han definido *subworkflows* locales, en los cuales la estructura de cada uno se enfoca en tres etapas: ejecución de la herramienta, generación de un resumen de resultados y consolidación de estos en un archivo único. A continuación, se presenta el flujo de trabajo diseñado específicamente para ARIBA y ABRICATE, que incluye los pasos principales:

```
nextflow
workflow ARIBA {
    take:
    input_reads // channel: trimreads from fastp
    db_name     // string:  db version to use

    main:
    ch_versions = Channel.empty()

    //OBTENER BASE DE DATOS
    ARIBA_GETREF(db_name)

    ch_ariba_db                = ARIBA_GETREF.out.db.dump(tag:
'Ariba_db')

    //RUN ARIBA
    ARIBA_RUN(input_reads, ch_ariba_db)

    ARIBA_RUN.out.report.collect{meta, report -> report}
        .map{ report -> [[id:"ariba-${db_name}-report"], report]}
        .set{ ch_merge_report }

    ARIBA_RUN.out.summary.collect{meta, summary -> summary}
        .map{ summary -> [[id:"ariba-${db_name}-summary"], summary]}
        .set{ ch_merge_summary }

    SUMMARY_REPORT(ch_merge_report, 'tsv', 'tsv', '-C "$" --lazy-
quotes')

    SUMMARY_ARIBA(ch_merge_summary, 'csv', 'csv', '--lazy-quotes')
}
```

Workflow Local 1. Flujo de trabajo del subworkflow ARIBA: se realiza la obtención de la base de datos especificada y la ejecución de ARIBA con las lecturas de entrada procesadas. Los resultados se agrupan en un reporte y un resumen, que luego son consolidados en archivos únicos en formatos TSV y CSV, respectivamente, mediante la herramienta CSVTK, para facilitar la visualización y comparación de los resultados.

```
nextflow
workflow ABRICATE {
    take:
    input_reads // channel: trimreads from fastp
    db_name     // string:  db version to use

    main:
    ch_versions = Channel.empty()

    //RUN ABRICATE
    ABRICATE_RUN (input_reads, db_name)

    ABRICATE_RUN.out.summary.collect{meta, summary -> summary}
        .map{ summary -> [[id:"abricate-${db_name}-summary"],
summary]}
        .set{ ch_merge_summary }

    //RUN SUMMARY CSVTK
    SUMMARY_REPORT(ch_merge_summary, 'tsv', 'tsv', '--lazy-quotes')

}
```

Workflow Local 2. Flujo de trabajo del *subworkflow* ABRICATE: el proceso consiste en ejecutar ABRICATE con las lecturas de entrada y la base de datos especificada. Los resultados se recopilan y agrupan en un canal de resumen, que se consolida en un archivo TSV mediante la herramienta CSVTK, optimizando la organización y visualización de los datos obtenidos.

Como parte del análisis de tipificación molecular en *Staphylococcus aureus*, se desarrolló un *subworkflow* de consolidación, denominado ***spatypes***, que integra todos los resultados de tipificación molecular obtenidos para cada genoma analizado. Este *subworkflow* fue diseñado para facilitar la clasificación e identificación de cepas en función de sus características moleculares, optimizando el flujo de trabajo para la identificación de patrones de resistencia, virulencia y variantes clonales en las muestras analizadas.

4.6.Ejecución del pipeline y opciones configurables

Para la ejecución del pipeline, se siguieron una serie de pasos que comenzaron con la preparación del entorno de trabajo, pasando por la ejecución del análisis y la configuración de opciones avanzadas y complementarias, adaptando el pipeline a los objetivos específicos del estudio.

4.6.1. Preparación del Entorno

Se realizo una copia del repositorio de Git disponible en https://github.com/VictorPizarroR/StaPhyloRes, este directorio fue copiado en la ruta local o HPC, según corresponda:

```
Bash
git clone https://github.com/VictorPizarroR/StaPhyloRes.git
```

Una vez clonado el repositorio, se creó un entorno de Conda a partir del archivo YML proporcionado al interior de la carpeta “resorces”, que contiene todas las dependencias necesarias para ejecutar el pipeline:

```
bash
conda env create -f StaPhyloRes/resources/resvirpredictor.yml --name
env_name
```

Una vez creado el entorno, este se activó con el siguiente comando:

```
bash
conda activate env_name
```

4.6.2. Ejecución Básica del Pipeline

La ejecución estándar del pipeline se realiza con el siguiente comando:

```
bash
nextflow run StaPhyloRes/ --input prepared_input.csv --outdir
outdirpath/
```

Este comando ejecuta los módulos principales del pipeline, que incluyen control de calidad, ensamblaje, análisis de resistencia y virulencia, tipificación molecular y generación de informes.

Los pipelines creados con *Nextflow* ofrecen una gran versatilidad al permitir la continuación de trabajos previamente iniciados que se hayan detenido por cualquier motivo. Esto se logra mediante el comando `-resume`, que indexa los datos en la carpeta definida como *workdir* y aprovecha el almacenamiento en caché para evitar la repetición de análisis ya realizados, optimizando así el tiempo y los recursos computacionales.

```
bash
nextflow run StaPhyloRes/ --input prepared_input.csv --outdir
outdirpath/ -resume
```

4.6.3. Entornos de ejecución disponibles

El pipeline puede ejecutarse en distintos entornos mediante *profiles* que permiten la creación de entornos **Conda** o **Docker** aislados para cada herramienta utilizada, proporcionando todas las dependencias necesarias automáticamente y sin necesidad de utilizar el archivo YML. Los *profiles* se activan con los siguientes comandos:

1. Conda

```
bash
nextflow run StaPhyloRes/ --input prepared_input.csv --outdir
outdirpath/ -profile conda
```

2. Docker

```
bash
nextflow run StaPhyloRes/ --input prepared_input.csv --outdir
outdirpath/ -profile docker
```

Ambos *profiles* incluyen tanto los módulos de Ejecución Básica (control de calidad, ensamblaje, análisis de resistencia y virulencia, entre otros) como el módulo de Ejecución de Filogenia, permitiendo realizar un análisis integral que cubre todas las fases de procesamiento.

4.6.4. Uso de una Base de Datos Personalizada

El pipeline permite la integración de una base de datos personalizada para el análisis con Abricate, según los pasos descritos en la página del autor. Esta base de datos fue denominada "*staph*", al momento de su configuración y se encuentra en el directorio *resources* de GitHub del TFM.

La ejecución con una base de datos personalizada se realiza mediante el siguiente comando:

```
bash
nextflow run StaPhyloRes/ --input prepared_input.csv --outdir
outdirpath/ --abricate_db true
```

4.6.5. Análisis de filogenia

Para realizar el análisis filogenético, primero se descarga la base de datos de referencia de MASH desde el siguiente enlace: refseq.genomes.k21s1000.msh. Luego, se especifica su ruta en la ejecución del pipeline:

```
bash
nextflow run StaPhyloRes/ --input prepared_input.csv --outdir
outdirpath/ --phylogeny true --mash_reference
/path/to/mash_reference.msh
```

4.6.6. Parámetros configurables

La tabla a continuación resume las opciones configurables para adaptar el pipeline a diferentes necesidades de análisis:

Categoría	Opción	Descripción
Opciones de Entrada y Salida	--input [string]	Ruta al archivo CSV de muestras.
	--outdir [string]	Directorio de salida para los resultados.
	--abricate_db [boolean]	Utiliza una base de datos personalizada.
	--email [string]	Correo para notificaciones de finalización.
Omisión de Análisis Específicos	--skip_unicycler	Omite el ensamblaje con Unicycler.

	--skip_ariba	Omite el análisis de resistencia con ARIBA.
	--skip_assemblyanalisis	Omite el análisis con Abricate y Staramr.
	--skip_mykrobe	Omite el análisis con Mykrobe.
	--skip_mlst	Omite el análisis de MLST.
Parámetros de Filogenia	--phylogeny	Activa el análisis filogenético.
	--mash_reference [string]	Ruta a la referencia MASH.
Solicitud de Recursos	--max_cpus [integer]	Número máximo de CPUs (por defecto 16).
	--max_memory [string]	Memoria máxima (por defecto 12 GB).
	--max_time [string]	Tiempo máximo (por defecto 120 horas).

Tabla 3. Opciones configurables del pipeline ***StaPhyloRes***. La tabla muestra las distintas categorías de parámetros ajustables, que permiten adaptar el pipeline a necesidades específicas de análisis. Incluye opciones para especificar rutas de entrada y salida, omitir etapas específicas del análisis, activar el análisis filogenético y gestionar la asignación de recursos computacionales. Estas configuraciones brindan flexibilidad para personalizar el pipeline según los requerimientos del estudio.

4.7. Resultados de validación

Para la validación final del pipeline, se generó un archivo CSV que contiene la totalidad de secuencias a utilizar, el cual se ejecutó en un entorno Conda creado a partir del archivo YML disponible en el repositorio de GitHub, en la carpeta *resources*. El comando de ejecución fue el siguiente:

```
bash
nextflow run StaPhyloRes/--input prepared_input.csv --outdir
/home/results --abricate_db true --phylogeny true --mash_reference
/home/refseq.genomes.k21s1000.msh --max_memory 12.GB --max_cpus 16
```

Bajo los recursos especificados, el pipeline se completó con los siguientes resultados:

```
bash
-[StaPhyloRes] Pipeline completed successfully, but with errored
process(es) -
Completed at: XX-XX-2024 XX:XX:XX
Duration    : 19h 48m 30s
CPU hours   : 89.9 (0.1% failed)
Succeeded   : 658
Ignored     : 1
Failed      : 1
```

Durante esta ejecución, se identificó una incapacidad para contrastar las secuencias de un aislado con la base de datos de VFDB de Ariba. Sin embargo, el pipeline ha sido configurado para que estos errores menores no detengan la ejecución general, permitiendo una mayor tolerancia a fallos y una finalización completa de los análisis en caso de errores aislados. Esta configuración mejora la robustez del pipeline frente a posibles interrupciones y asegura la continuidad en el procesamiento de grandes volúmenes de datos.

4.7.1. Informe de resultados de predicción a los antibióticos

El *output summary* generado por Mykrobe proporciona un análisis detallado de resistencia a antibióticos, basado en datos de secuenciación genómica. Entre los datos obtenidos, se encuentran:

```
R
[1] "sample"
 [2] "drug"
 [3] "susceptibility"
 [4]
"variants..dna_variant.AA_variant.ref_kmer_count.alt_kmer_count.conf..
.use...format.json.for.more.info."
 [5]
"genes..prot_mut.ref_mut.percent_covg.depth...use...format.json.for.mo
re.info."
 [6] "mykrobe_version"
 [7] "files"
 [8] "probe_sets"
 [9] "genotype_model"
[10] "kmer_size"
[11] "phylo_group"
[12] "species"
[13] "lineage"
[14] "phylo_group_per_covg"
[15] "species_per_covg"
[16] "lineage_per_covg"
[17] "phylo_group_depth"
[18] "species_depth"
[19] "lineage_depth"
```

Este *summary* proporciona información detallada sobre cada muestra, incluyendo la predicción de la sensibilidad o resistencia a diferentes antibióticos, el linaje bacteriano y la identificación de genes de resistencia, entre otros. Esta versatilidad permite un análisis exhaustivo de las características de las cepas de *Staphylococcus aureus* estudiadas. Para este TFM, se ha seleccionado un subconjunto de las columnas generadas, en particular *sample, drug y susceptibility*, con el objetivo de realizar un

análisis exploratorio de la susceptibilidad a antibióticos en las muestras obtenidas durante la validación del pipeline.

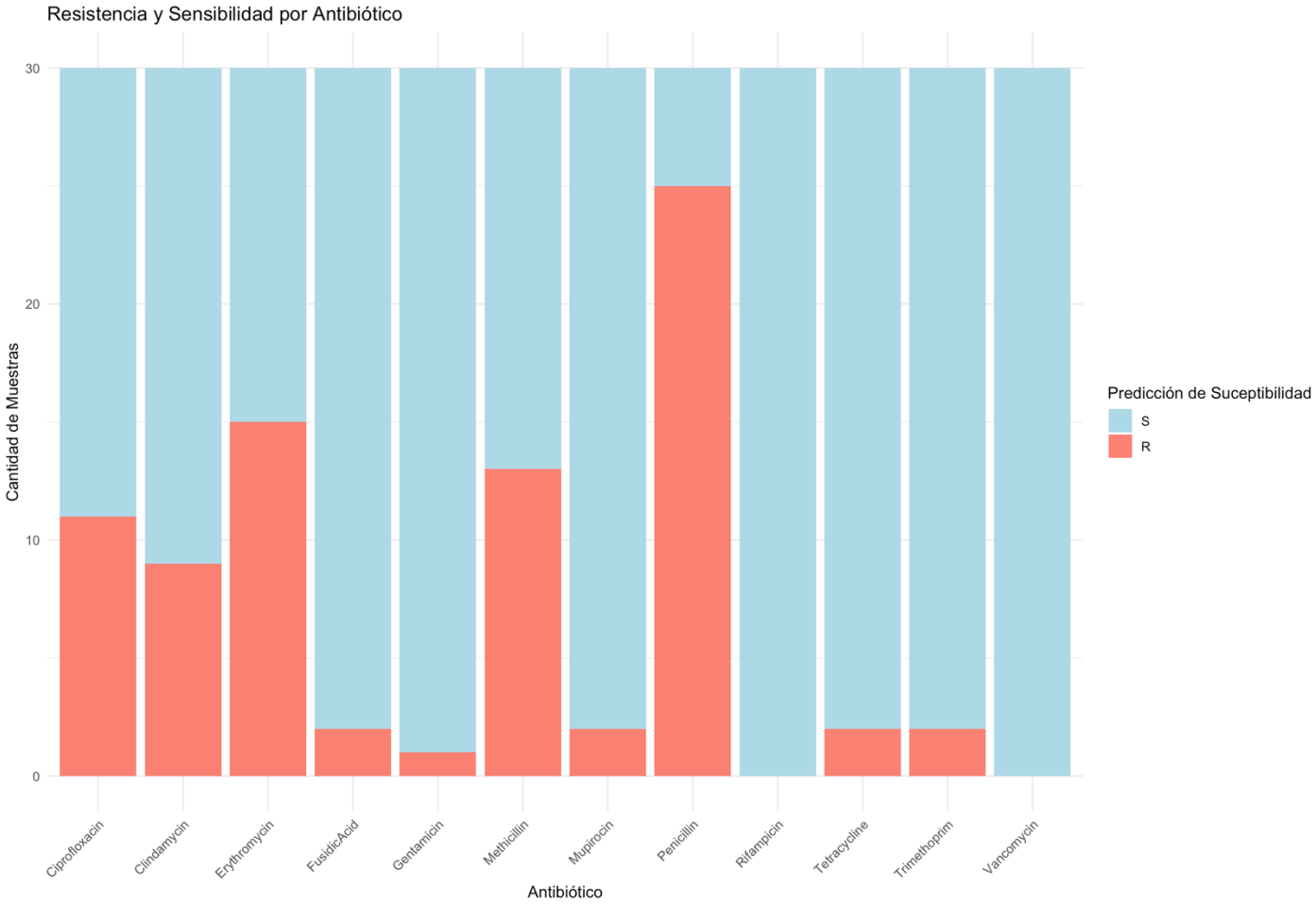

Figura 2: Gráfico de barras apiladas que muestra la cantidad de muestras de *Staphylococcus aureus* clasificadas como resistentes (R) o sensibles (S) a distintos antibióticos, de acuerdo con los datos obtenidos mediante *Mykrobe*.

4.7.2. Visualización del árbol filogenético de distancias

El análisis filogenético de los aislados de *Staphylococcus aureus* mediante la construcción de un árbol basado en *core-SNPs* **(Figura 3)** permitió identificar las relaciones genéticas y posibles agrupaciones evolutivas entre los diferentes aislados y una secuencia de referencia. Este enfoque aporta información valiosa sobre la diversidad genética y la cercanía evolutiva de los aislados estudiados. A su v es estos resultados pueden ser utilizados por otras herramientas de análisis para poder obtener la matriz de distancias genómicas, la cual se basa en la comparación de los SNPs presentes en los genomas de los aislados, proporcionando así una medida cuantitativa de la similitud o divergencia genética.

La secuencia de referencia utilizada en este análisis corresponde al genoma de *Staphylococcus aureus subsp. aureus USA300_TCH1516* (GCF_000253135.1). El análisis de distancias de *SNPs* entre los distintos nodos revela subgrupos con alta similitud genética, como el formado por los aislados 24Sau021, 24Sau006 y 24Sau022, que presentan distancias de *SNPs* mínimas entre ellos. Esta proximidad genética sugiere una evolución reciente compartida o la posibilidad de un brote común. Otros grupos, como el de los aislados 24Sau010, 24Sau018 y 24Sau007, también presentan distancias cortas entre sí, lo que indica relaciones evolutivas cercanas.

Por otro lado, las mayores distancias observadas en ramas largas, especialmente en aislados como 24Sau039 y 24Sau002, sugieren una mayor variabilidad genética, lo que podría representar linajes divergentes o adaptaciones a condiciones distintas. Estos aislados más distantes podrían corresponder a variantes genéticas más antiguas o a linajes más diferenciados dentro de *Staphylococcus aureus*.

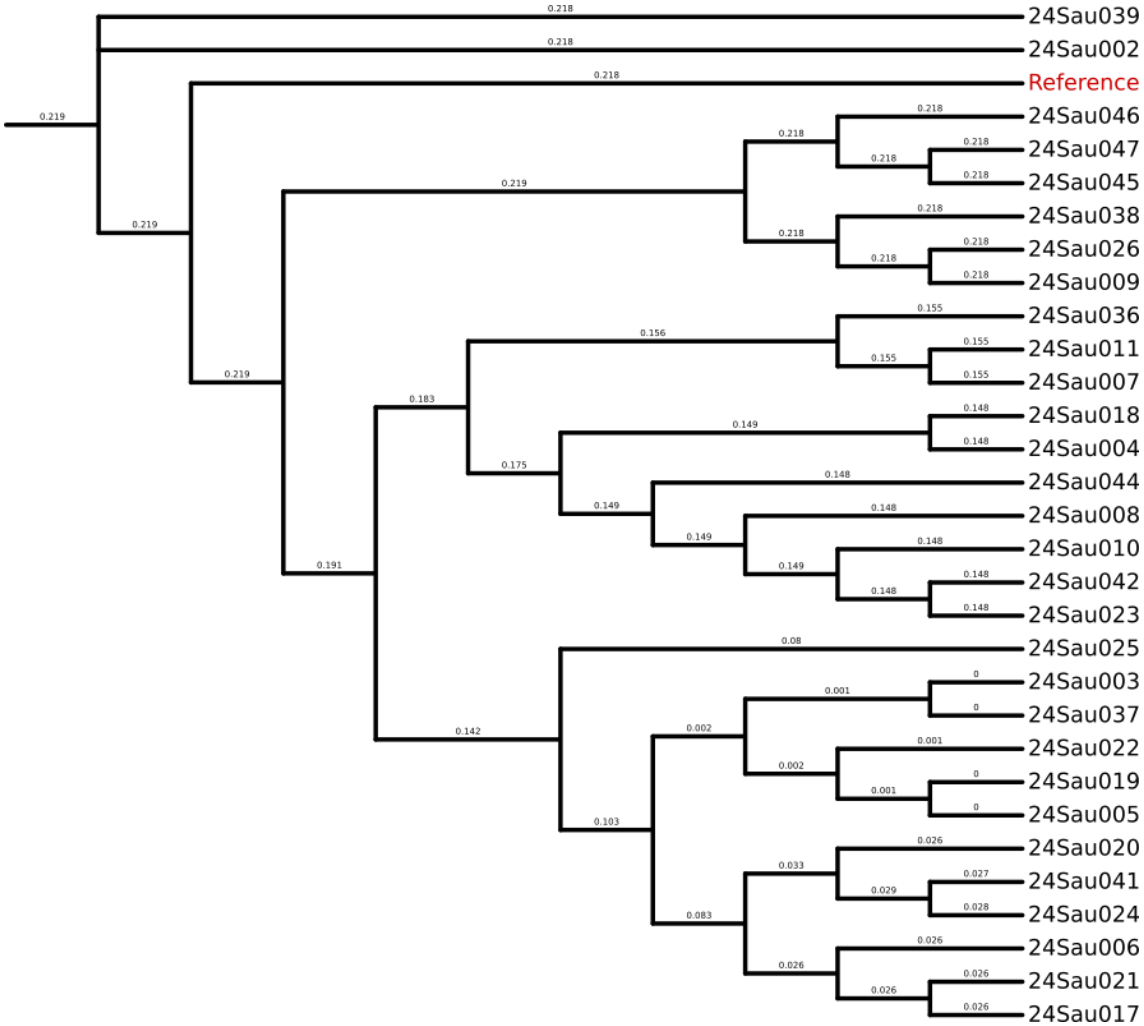

Figura 3. Árbol filogenético basado en SNPs de aislados de *Staphylococcus aureus*. El árbol muestra las relaciones evolutivas entre los diferentes aislados, con base en las variaciones de SNPs.

4.7.3. Informes consolidados de virulencia y resistencia

Para este apartado, se ha elegido la base de datos de CARD analizada por Ariba. El análisis consolidado de los genes de virulencia y resistencia antimicrobiana reveló la presencia de varios genes de alta relevancia clínica, los cuales fueron visualizados mediante un *heatmap* **(figura 4),** que resume su distribución y prevalencia entre las cepas analizadas.

Entre los genes identificados, destacan *blaZ* y *fosB*, los cuales están asociados a resistencia a antibióticos comúnmente utilizados. El gen *blaZ* es responsable de conferir resistencia a antibióticos betalactámicos, una clase fundamental en el tratamiento de infecciones bacterianas, mientras que *fosB* se asocia con resistencia a fosfomicina, un antibiótico que suele usarse en casos de infecciones resistentes. Estos genes son particularmente relevantes en infecciones invasivas causadas por *S. aureus*, donde las opciones de tratamiento son limitadas.

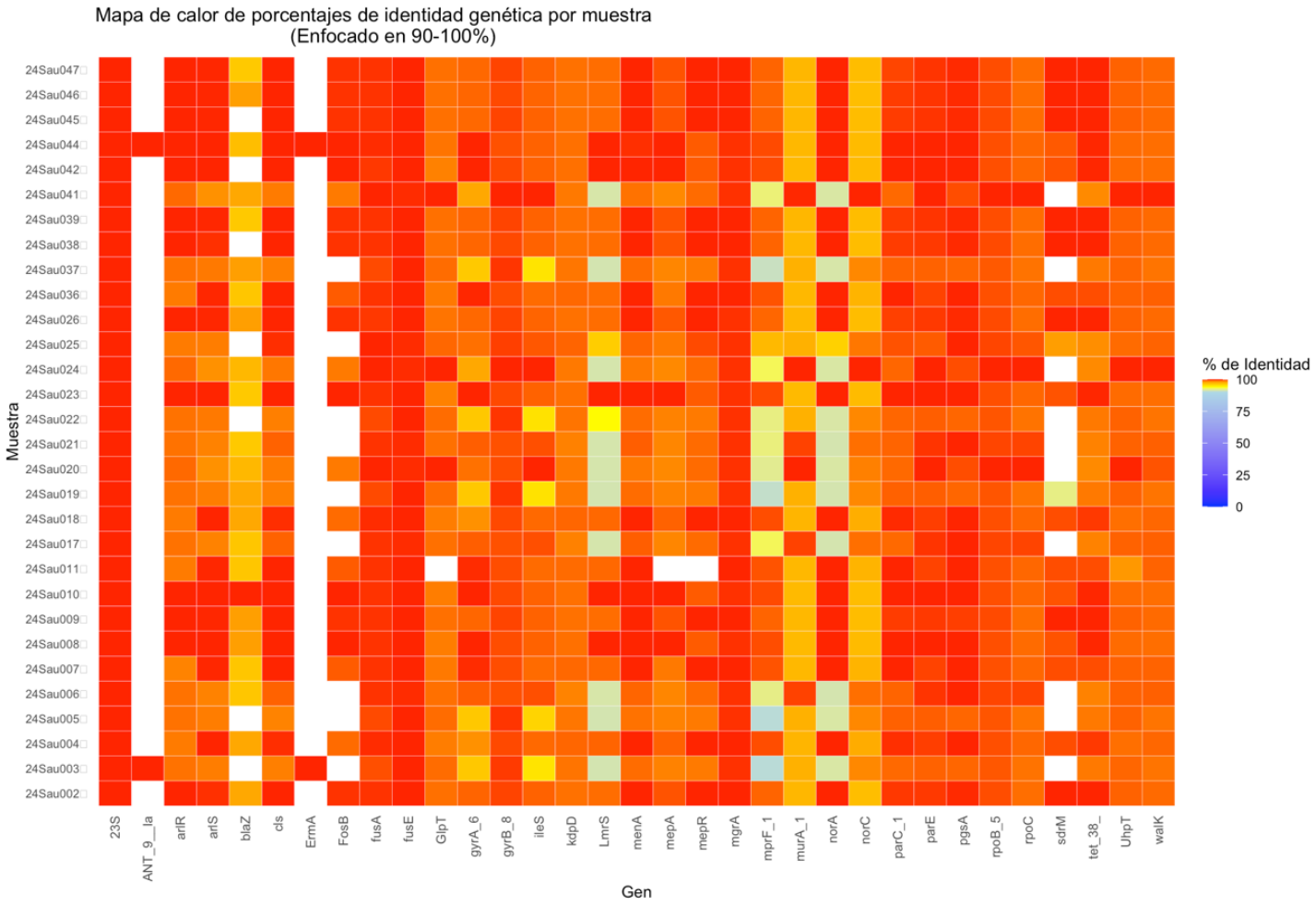

Figura 4. *Heatmap* que representa los porcentajes de identidad genética para diversos genes en cepas de *Staphylococcus aureus*. El mapa se centra en un rango de identidad del 90-100%, mostrando variaciones en la presencia de genes de resistencia y virulencia. Los colores van de azul (baja identidad) a rojo (alta identidad), indicando la variabilidad en la concordancia genética entre las muestras y los genes analizados.

4.7.4. Informe de resultados tipificación molecular

El *subworkflow* ***spatypes*** ejecuta de manera secuencial cada herramienta de tipificación molecular seleccionada, generando archivos de salida específicos y resúmenes detallados para cada muestra. Posteriormente, estos resultados se consolidan en un archivo final, que organiza los datos de manera uniforme, tal como se presenta en la **Tabla 3**. Este archivo integrado permite una comparación directa de los

resultados de tipificación molecular, estandarizando la disposición de la información y facilitando el análisis comparativo entre diferentes cepas.

Esta estructura de consolidación automatizada permite que el proceso de tipificación molecular sea más rápido y eficiente, minimizando los posibles errores de integración y reduciendo el tiempo dedicado a la tabulación manual de resultados. Además, al centralizar los resultados en un solo archivo, se mejora la visualización y análisis de los datos, lo que resulta particularmente útil para estudios epidemiológicos y la caracterización de cepas en función de sus perfiles genéticos y fenotípicos.

Sample	mlst	spatyper	agr_group	scc_mec
24Sau017	398	t1451	gp1	
24Sau021	398	t1451	gp1	
24Sau004	15	t084	gp2	
24Sau008	8	t008	gp1	meca-IVc
24Sau011	72	07-23-12-21-12-17-21-12-17-20-17-12-12-17	gp1	meca-IVc
24Sau007	72	t6685	gp1	meca-IVc
24Sau006	398	t17617	gp1	
24Sau009	125	t837	gp2	IV-meca-IVc
24Sau003	45	t706	gp1	
24Sau002	5	t002	gp2	
24Sau020	34	04-44-24-16-34-16-12-25-22-34	gp3	
24Sau018	15	t084	gp2	
24Sau022	45	t505	gp1	
24Sau005	45	t230	gp1	
24Sau019	45	t230	gp1	
24Sau010	8	t008	gp1	IV-meca-IVc
24Sau023	8	t024	gp1	
24Sau038	125	t1084	gp2	IV-meca-IVc
24Sau036	72	t148	gp1	IV-meca-IVa
24Sau042	8	t024	gp1	
24Sau024	30	t012	gp3	
24Sau045	5	t002	gp2	IV-meca-IVc
24Sau039	5	t002	gp2	IV-meca-IVc
24Sau025	59	t216	gp1	
24Sau041	30	t342	gp3	

24Sau047	-	t002	gp2	IV-meca-IVc
24Sau044	8	t008	gp1	
24Sau046	5	t002	gp2	IV-meca-IVc
24Sau037	45	09-02-13-17-34-34-242-34	gp1	meca
24Sau026	125	t067	gp2	IV-meca-IVc

Tabla 3. Resultados de tipificación molecular de cepas de *Staphylococcus aureus* obtenidos mediante el módulo ***spatypes***. La tabla muestra los resultados de identificación de secuencias multilocus (*mlst*), tipificación de *spa* (*spatyper*), grupo agr (*agr_group*) y el tipo de SCCmec (*scc_mec*) para cada muestra de validación.

El análisis de la tabla revela que algunas cepas presentan el gen de resistencia *mecA*, asociado a diferentes variantes de *SCCmec* (como *meca-IVc* y *meca*), lo cual es indicativo de cepas resistentes a la meticilina (MRSA). Por otro lado, la tipificación *spa* muestra diversidad, con variantes como *t084*, *t002* y *t1451*, lo que sugiere una variedad de linajes en las muestras. El grupo *agr*, que regula factores de virulencia en *S. aureus*, también varía entre cepas, destacándose los grupos *gp1* y *gp2*, lo que puede influir en la patogenicidad de las cepas.

4.7.5. Informe de calidad de secuencias crudas, trimadas y ensamblados

En este análisis, se utilizaron los datos de secuenciación crudos de múltiples aislados de *Staphylococcus aureus*, los cuales fueron evaluados con FastQC. La gráfica generada mediante MultiQC muestra el conteo de lecturas únicas y duplicadas para cada aislado, permitiendo observar la proporción de duplicación en las secuencias y evaluar la calidad del conjunto de datos para el análisis posterior.

En la **figura 5** se presenta el número de lecturas únicas (en azul) y duplicadas (en gris) para cada aislado de *Staphylococcus aureus*. Se puede observar una variabilidad en el conteo de lecturas únicas y duplicadas entre los diferentes aislados. Aislados como 24Sau017 y 24Sau046 muestran una alta proporción de lecturas duplicadas en comparación con otras muestras, lo que podría sugerir sobre-secuenciación o redundancia en los datos. Esta alta proporción de lecturas duplicadas puede afectar la

calidad del ensamblaje genómico y los análisis de diversidad genética, ya que la duplicación puede influir en la cobertura y precisión del ensamblaje. Por otro lado, muestras con un balance entre lecturas únicas y duplicadas, como 24Sau021 y 24Sau023, representan un conjunto de datos más equilibrado para el análisis.

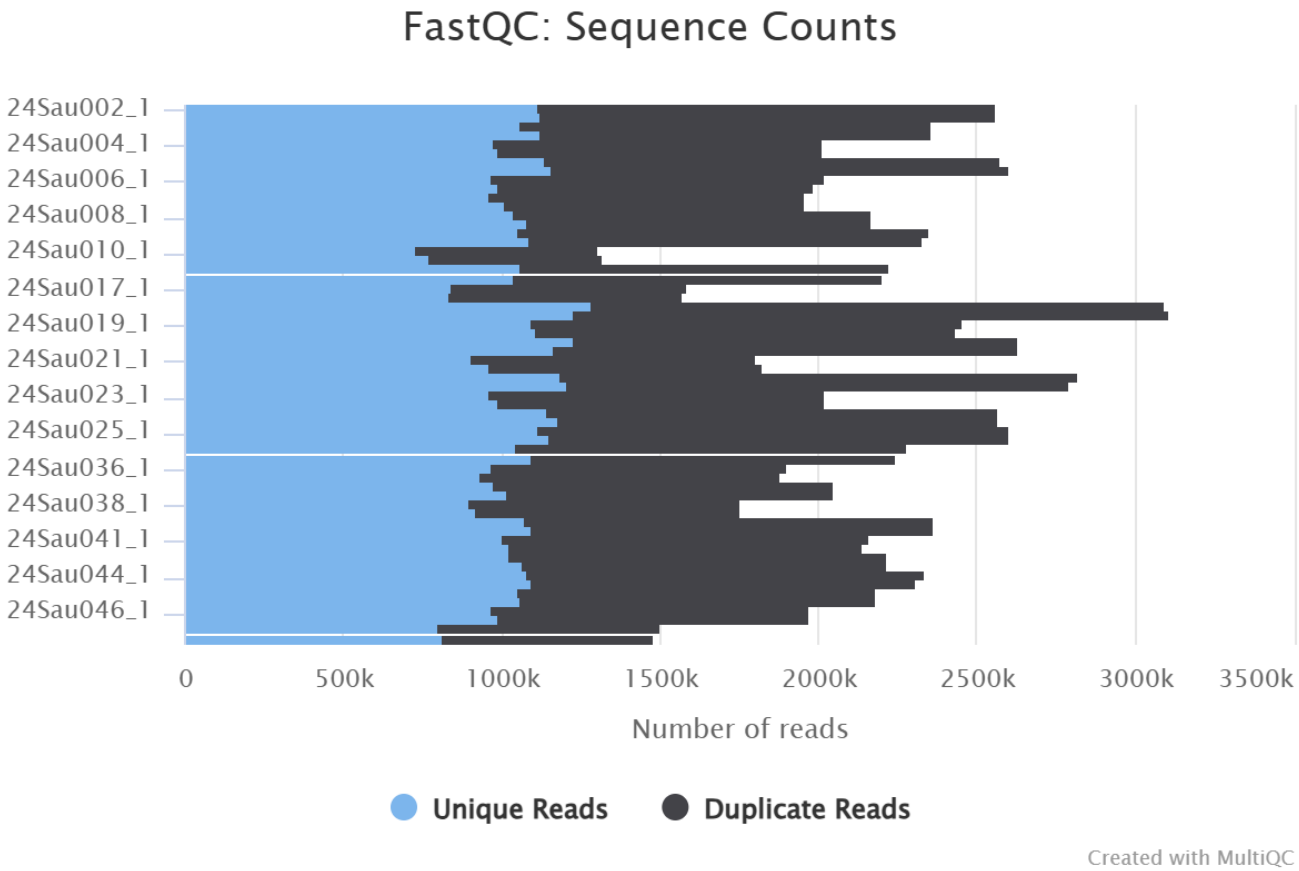

Figura 5. Conteo de secuencias únicas y duplicadas por aislado de *Staphylococcus aureus*. La gráfica muestra el número de lecturas únicas (en azul) y duplicadas (en gris) para cada muestra, evaluadas con FastQC y visualizadas mediante MultiQC. La proporción de lecturas duplicadas puede impactar la calidad del análisis genómico, influyendo en la cobertura y precisión del ensamblaje.

Posteriormente el análisis de secuencias mediante FastP Para este propósito, se utilizó Fastp en el análisis de múltiples aislados de *Staphylococcus aureus*, con el objetivo de identificar y filtrar lecturas en función de criterios de calidad, longitud y composición. La gráfica generada con MultiQC muestra las lecturas que superaron el filtro de calidad y aquellas que fueron descartadas por distintos motivos.

En la **figura 6**, se presenta el número de lecturas filtradas para cada aislado de *Staphylococcus aureus*. La mayoría de los aislados muestran una alta proporción de lecturas que pasaron el filtro, lo cual indica un buen estado general de los datos.

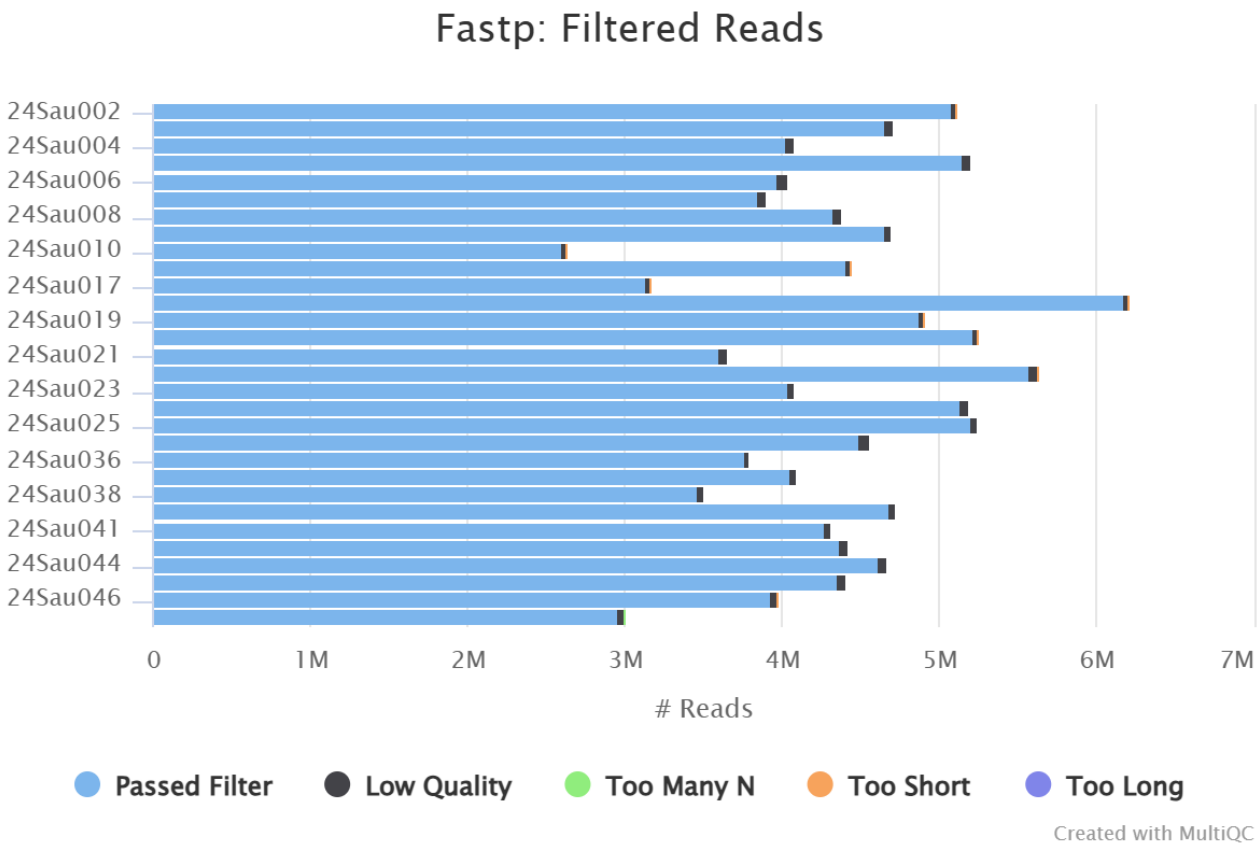

Figura 6. Lecturas filtradas por Fastp para cada aislado de *Staphylococcus aureus*. La gráfica muestra el número de lecturas que superaron el filtro de calidad (en azul) y las lecturas eliminadas debido a baja calidad (en negro), exceso de bases N (en verde), longitud inadecuada (en naranja y morado).

Finalmente, la evaluación de la calidad de ensamblaje permite determinar la integridad y continuidad de los *contigs* obtenidos. La gráfica generada por MultiQC muestra la distribución de *contigs* en distintas categorías de tamaño para cada aislado, lo cual permite identificar la calidad del ensamblaje y la fragmentación de las secuencias obtenidas.

En la **figura 7**. Podemos identificar aislados como 24Sau021 y 24Sau023 muestran una proporción equilibrada de *contigs* en varias categorías, lo cual podría sugerir un

ensamblaje moderadamente fragmentado, pero relativamente completo. En contraste, aislados como 24Sau017 y 24Sau010 tienen un mayor número de *contigs* cortos (rojo y morado), lo que sugiere una mayor fragmentación y posiblemente una calidad de ensamblaje más baja.

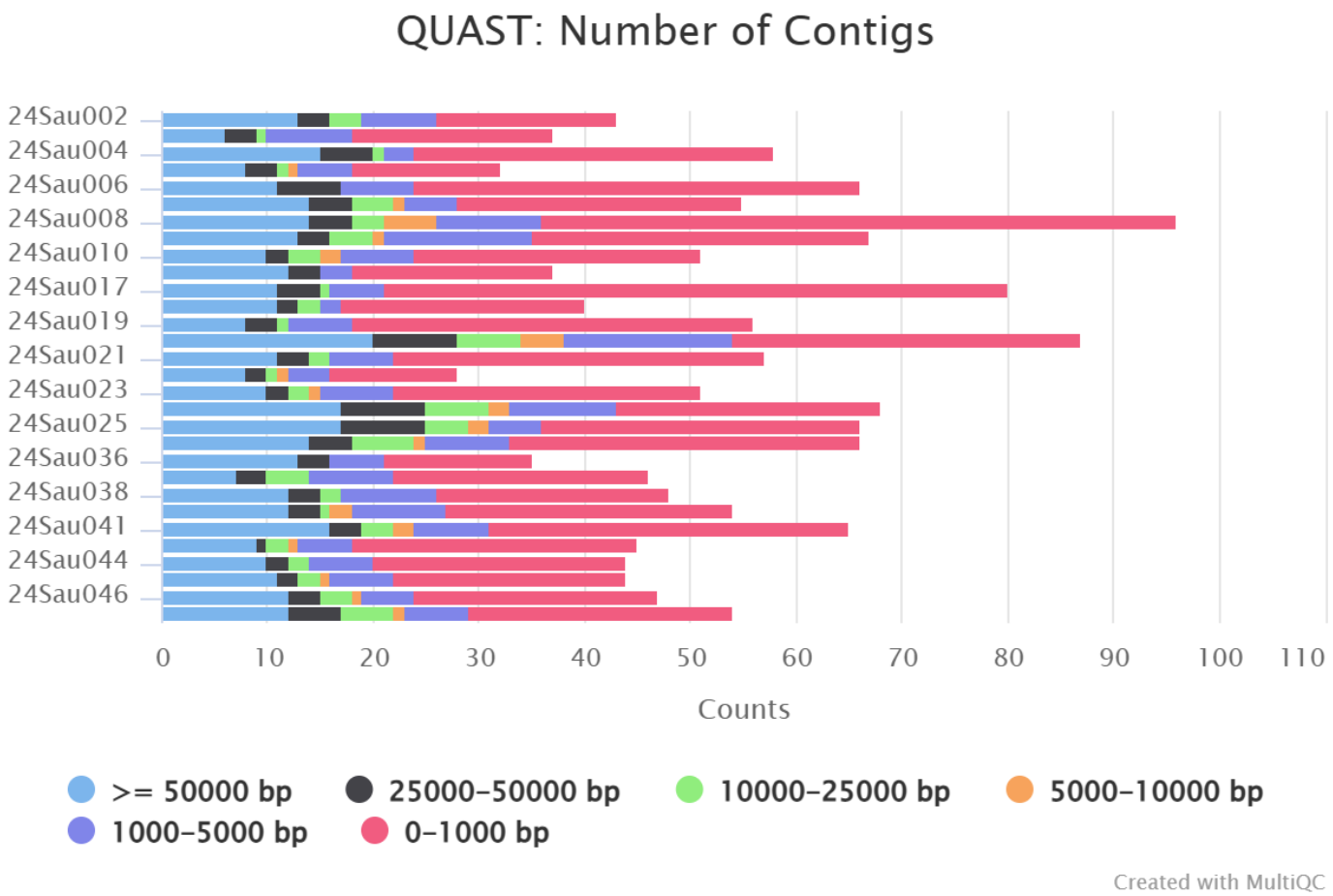

Figura 7. Número de *contigs* en diferentes categorías de tamaño para cada aislado de *Staphylococcus aureus*, evaluado con QUAST y visualizado mediante MultiQC. La gráfica muestra la distribución de *contigs* según su longitud: ≥50,000 bp (en azul), 25,000–50,000 bp (en negro), 10,000–25,000 bp (en verde), 5,000–10,000 bp (en naranja), 1,000–5,000 bp (en morado), y 0–1,000 bp (en rojo). La cantidad de *contigs* largos indica un ensamblaje de mayor calidad, mientras que un alto número de *contigs* cortos sugiere mayor fragmentación del ensamblaje.

5. Principales conclusiones

Este proyecto ha culminado en el diseño y desarrollo exitoso de un pipeline bioinformático modular y flexible, denominado ***StaPhyloRes***, específicamente orientado a la caracterización molecular de cepas de *Staphylococcus aureus*. ***StaPhyloRes*** está basado en *Nextflow*, lo cual permite su escalabilidad y adaptabilidad en infraestructuras de alto rendimiento (HPC) y sistemas convencionales, lo que facilita su uso en entornos de investigación y clínicos donde la eficiencia y reproducibilidad son esenciales.

Este pipeline permite un entendimiento detallado de los mecanismos genéticos subyacentes en la resistencia a antibióticos como la meticilina, macrólidos y fluoroquinolonas. La presencia de genes como *blaZ*, *FosB*, *ermA*, y *ermC*, entre otros, resalta la capacidad de adaptación de estas cepas y subraya la importancia de un monitoreo constante en contextos clínicos para optimizar estrategias de tratamiento.

La inclusión de herramientas de tipificación molecular y análisis filogenético proporciona una visión integral de la diversidad genética y las relaciones evolutivas entre las cepas de *S. aureus*. La identificación de variantes en los grupos *agr*, tipos de *spa*, y *SCCmec*, junto con el análisis filogenético basado en *SNPs*, permite la identificación de linajes clonales y la evaluación de la dispersión genética en diferentes entornos.

StaPhyloRes ha demostrado ser eficaz en la identificación de genes clave de resistencia antimicrobiana y factores de virulencia en cepas de *S. aureus*. A comparación de **BACTOPIA, *StaPhyloRes*** se centra en un análisis automatizado siguiendo un flujo definido, el cual es capaz de optimizar tiempo de ejecución al enfocarse en tareas prioritarias en la práctica clínica y la vigilancia epidemiológica, por ejemplo, en el caso de enfermedades invasoras producidas por *Staphylococcus aureus.* Esta ventaja permite generar informes rápidos y consolidados para clara visualización., aunque también define una limitación frente a estudios específicos, evidenciando la necesidad de continuar desarrollando nuevos módulos y *subworkflows* que puedan adaptarse a entornos clínicos y de investigación más complejos.

Finalmente, como perspectiva a futuro, la adaptabilidad de ***StaPhyloRes*** permite su expansión a otros patógenos de interés clínico, ampliando su utilidad y contribución en el campo de la bioinformática aplicada a la salud pública. Además, al estar disponible en un repositorio público en GitHub, ***StaPhyloRes*** ofrece a la comunidad científica una herramienta accesible y reproducible, fomentando la colaboración y el avance en el estudio de la resistencia antimicrobiana y la vigilancia epidemiológica.

6. Declaración obligatoria del uso de herramientas de inteligencia artificial

Se declara que se ha utilizado y se mantendrá el uso de una herramienta de inteligencia artificial como asistente de redacción de este Trabajo de Fin de Máster (TFM). Adicionalmente, se ha utilizado una herramienta de inteligencia artificial explicativa para las secciones de código y verificación de errores de programación, apoyando en la resolución de errores, optimización de código y en la implementación de buenas prácticas de programación y documentación. Este enfoque permitió un avance eficiente en el desarrollo de ***StaPhyloRes***, incrementando la calidad y reproducibilidad del pipeline.

Estas herramientas se emplean con el objetivo de mejorar la estructura del texto, la identificación de entidades nombradas (Ej, nombre de genes y bacterias), con la finalidad de asegurar el correcto formato del documento.

Todo contenido generado y/o explicado mediante las herramientas de IA, han sido y será revisado a detalle para garantizar su precisión, coherencia y cohesión, manteniendo la alineación con los objetivos de este TFM.

7. Bibliografía

1. Arunachalam, K., Pandurangan, P., Shi, C., & Lagoa, R. (2023). Regulation of Staphylococcus aureus Virulence and Application of Nanotherapeutics to Eradicate S. aureus Infection. *Pharmaceutics*, *15*(2), 310. https://doi.org/10.3390/pharmaceutics15020310

2. *Staphylococcus aureus resistente a meticilina y personal sanitario | Enfermedades Infecciosas y Microbiología Clínica*. (s. f.). Recuperado 29 de julio de 2024, de https://www.elsevier.es/es-revista-enfermedades-infecciosas-microbiologia-clinica-28-articulo-staphylococcus-aureus-resistente-meticilina-personal-S0213005X13001195

3. Larsen, A. R., Stegger, M., Böcher, S., Sørum, M., Monnet, D. L., & Skov, R. L. (2009). Emergence and Characterization of Community-Associated Methicillin-Resistant Staphyloccocus aureus Infections in Denmark, 1999 to 2006. *Journal of Clinical Microbiology*, *47*(1), 73-78. https://doi.org/10.1128/JCM.01557-08

4. Fernández, C. L. (2012). *Epidemiología molecular de Staphylococcus aureus resistente a meticilina del linaje CC398 de distintos orígenes: Resistencia, virulencia y contenido plasmídico*. https://www.semanticscholar.org/paper/Epidemiolog%C3%ADa-molecular-de-Staphylococcus-aureus-a-Fern%C3%A1ndez/0e43631112eb8f5a8f37a414c109250ee805ee35

5. Kong, X., Gong, S., Su, L., Howard, N., & Kong, Y. (2017). Automatic Detection of Acromegaly From Facial Photographs Using Machine Learning Methods. *EBioMedicine*, *27*, 94-102. https://doi.org/10.1016/j.ebiom.2017.12.015

6. Martínez-Medina, R. M., Montalvo-Sandoval, F. D., Magaña-Aquino, M., Terán-Figueroa, Y., Pérez-Urizar, J. T., Martínez-Medina, R. M., Montalvo-Sandoval, F. D., Magaña-Aquino, M., Terán-Figueroa, Y., & Pérez-Urizar, J. T. (2020). Prevalencia y caracterización genotípica de cepas de Staphylococcus aureus resistente a meticilina aisladas en un

hospital regional mexicano. *Revista chilena de infectología*, *37*(1), 37-44. https://doi.org/10.4067/S0716-10182020000100037

7. Chung, M., Antignac, A., Kim, C., & Tomasz, A. (2008). Comparative study of the susceptibilities of major epidemic clones of methicillin-resistant Staphylococcus aureus to oxacillin and to the new broad-spectrum cephalosporin ceftobiprole. *Antimicrobial Agents and Chemotherapy*, *52*(8), 2709-2717. https://doi.org/10.1128/AAC.00266-08

8. Enright, M. C., Day, N. P. J., Davies, C. E., Peacock, S. J., & Spratt, B. G. (2000). Multilocus Sequence Typing for Characterization of Methicillin-Resistant and Methicillin-Susceptible Clones of Staphylococcus aureus. *Journal of Clinical Microbiology*, *38*(3), 1008-1015. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC86325/

9. Hallin, M., Friedrich, A. W., & Struelens, M. J. (2009). Spa typing for epidemiological surveillance of Staphylococcus aureus. *Methods in Molecular Biology (Clifton, N.J.)*, *551*, 189-202. https://doi.org/10.1007/978-1-60327-999-4_15

10. van Belkum, A., Riewerts Eriksen, N., Sijmons, M., van Leeuwen, W., VandenBergh, M., Kluytmans, J., Espersen, F., & Verbrugh, H. (1996). Are variable repeats in the spa gene suitable targets for epidemiological studies of methicillin-resistant Staphylococcus aureus strains? *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology*, *15*(9), 768-770. https://doi.org/10.1007/BF01691972

11. Zhang, K., McClure, J.-A., Elsayed, S., & Conly, J. M. (2009). Novel Staphylococcal Cassette Chromosome mec Type, Tentatively Designated Type VIII, Harboring Class A mec and Type 4 ccr Gene Complexes in a Canadian Epidemic Strain of Methicillin-Resistant Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *53*(2), 531-540. https://doi.org/10.1128/AAC.01118-08

12. Ito, T., Katayama, Y., Asada, K., Mori, N., Tsutsumimoto, K., Tiensasitorn, C., & Hiramatsu, K. (2001). Structural Comparison of Three Types of Staphylococcal Cassette Chromosome mec Integrated in the Chromosome in Methicillin-Resistant Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *45*(5), 1323-1336. https://doi.org/10.1128/AAC.45.5.1323-1336.2001

13. Morfeldt, E., Tegmark, K., & Arvidson, S. (1996). Transcriptional control of the agr-dependent virulence gene regulator, RNAIII, in Staphylococcus aureus. *Molecular Microbiology*, *21*(6), 1227-1237. https://doi.org/10.1046/j.1365-2958.1996.751447.x

14. Sakoulas, G., Eliopoulos, G. M., Moellering, R. C., Wennersten, C., Venkataraman, L., Novick, R. P., & Gold, H. S. (2002). Accessory Gene Regulator (agr) Locus in Geographically Diverse Staphylococcus aureus Isolates with Reduced Susceptibility to Vancomycin. *Antimicrobial Agents and Chemotherapy*, *46*(5), 1492-1502. https://doi.org/10.1128/AAC.46.5.1492-1502.2002

15. Shopsin, B., Mathema, B., Alcabes, P., Said-Salim, B., Lina, G., Matsuka, A., Martinez, J., & Kreiswirth, B. N. (2003). Prevalence of agr Specificity Groups among Staphylococcus aureus Strains Colonizing Children and Their Guardians. *Journal of Clinical Microbiology*, *41*(1), 456-459. https://doi.org/10.1128/JCM.41.1.456-459.2003

16. Romero A, S., Castellano G, M., Ginestre P, M., & Perozo M, A. (2016). Leucocidina de Panton Valentine en cepas SAMR aisladas de pacientes del Hospital Universitario de Maracaibo. *Kasmera*, *44*(2), 111-120. http://ve.scielo.org/scielo.php?script=sci_abstract&pid=S0075-52222016000200005&lng=es&nrm=iso&tlng=es

17. Lozano, D., Díaz, L., Echeverry, M., Pineda, S., & Máttar, S. (2010). Staphylococcus aureus resistentes a meticilina (SARM) positivos para PVL aislados en individuos sanos de

Montería-Córdoba. *Universitas Scientiarum*, *15*(2), 159-165. http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0122-74832010000200007&lng=en&nrm=iso&tlng=es

18. Kato, F., Kadomoto, N., Iwamoto, Y., Bunai, K., Komatsuzawa, H., & Sugai, M. (2011). Regulatory Mechanism for Exfoliative Toxin Production in Staphylococcus aureus. *Infection and Immunity*, *79*(4), 1660-1670. https://doi.org/10.1128/IAI.00872-10

19. Dinges, M. M., Orwin, P. M., & Schlievert, P. M. (2000). Exotoxins of Staphylococcus aureus. *Clinical Microbiology Reviews*, *13*(1), 16-34. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC88931/

20. *Patógenos multirresistentes que son prioritarios para la OMS - OPS/OMS | Organización Panamericana de la Salud*. (2021, marzo 4). https://www.paho.org/es/noticias/4-3-2021-patogenos-multirresistentes-que-son-prioritarios-para-oms

21. *Whole genome sequencing: New ECDC framework suggests priority diseases and implementation options*. (2019, abril 4). https://www.ecdc.europa.eu/en/news-events/whole-genome-sequencing-new-ecdc-framework-suggests-priority-diseases-and

22. *ECDC strategic framework for the integration of molecular and genomic typing into European surveillance and multi-country outbreak investigations*. (2019, abril 4). https://www.ecdc.europa.eu/en/publications-data/ecdc-strategic-framework-integration-molecular-and-genomic-typing-european

23. *Red de Laboratorios para la Vigilancia de los Microorganismos Resistentes (RedLabRA)*. (s. f.). Recuperado 28 de julio de 2024, de https://www.isciii.es/QueHacemos/Servicios/DiagnosticoMicrobiol%C3%B3gicoyProgramasVigilancia/Paginas/RedLabRA.aspx

24. Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews. Genetics*, *11*(1), 31-46. https://doi.org/10.1038/nrg2626

25. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A., & Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews. Genetics*, *13*(9), 601-612. https://doi.org/10.1038/nrg3226

26. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316-319. https://doi.org/10.1038/nbt.3820

27. *Nf-core*. (s. f.). Recuperado 17 de octubre de 2024, de https://nf-co.re/

28. Langer, B. E., Amaral, A., Baudement, M.-O., Bonath, F., Charles, M., Chitneedi, P. K., Clark, E. L., Tommaso, P. D., Djebali, S., Ewels, P. A., Eynard, S., Yates, J. A. F., Fischer, D., Floden, E. W., Foissac, S., Gabernet, G., Garcia, M. U., Gillard, G., Gundappa, M. K., … Community, T. N.-C. (2024). *Empowering bioinformatics communities with Nextflow and nf-core* (p. 2024.05.10.592912). bioRxiv. https://doi.org/10.1101/2024.05.10.592912

29. Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276-278. https://doi.org/10.1038/s41587-020-0439-x

30. *Visual Studio Code—Code Editing. Redefined*. (s. f.). Recuperado 17 de octubre de 2024, de https://code.visualstudio.com/

31. *ChatGPT*. (s. f.). Recuperado 5 de noviembre de 2024, de https://chatgpt.com

32. *Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data*. (s. f.). Recuperado 18 de octubre de 2024, de https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

33. Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta*, *2*(2), e107. https://doi.org/10.1002/imt2.107

34. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

35. Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, *13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595

36. Seemann, T. (2024). *Tseemann/shovill* [Perl]. https://github.com/tseemann/shovill (Obra original publicada en 2016)

37. *Using SPAdes De Novo Assembler—Prjibelski—2020—Current Protocols in Bioinformatics—Wiley Online Library*. (s. f.). Recuperado 19 de octubre de 2024, de https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.102

38. Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069. https://doi.org/10.1093/bioinformatics/btu153

39. Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L. J., Anson, L., de Cesare, M., Piazza, P., Votintseva, A. A., Golubchik, T., Wilson, D. J., Wyllie, D. H., Diel, R., Niemann, S., Feuerriegel, S., Kohl, T. A., … Iqbal, Z. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nature Communications*, *6*(1), 10063. https://doi.org/10.1038/ncomms10063

40. Hunt, M., Mather, A. E., Sánchez-Busó, L., Page, A. J., Parkhill, J., Keane, J. A., & Harris, S. R. (2017). ARIBA: Rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial Genomics*, *3*(10), e000131. https://doi.org/10.1099/mgen.0.000131

41. Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *The Journal of Antimicrobial Chemotherapy*, *67*(11), 2640-2644. https://doi.org/10.1093/jac/dks261

42. Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F., & Hasman, H. (2014). In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, *58*(7), 3895-3903. https://doi.org/10.1128/AAC.02412-14

43. 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 Research*, *45*(D1), D566-D573. https://doi.org/10.1093/nar/gkw1004

44. Chen, L., Zheng, D., Liu, B., Yang, J., & Jin, Q. (2016). VFDB 2016: Hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Research*, *44*(D1), D694-697. https://doi.org/10.1093/nar/gkv1239

45. Seemann, T. (2024). *Tseemann/abricate* [Perl]. https://github.com/tseemann/abricate (Obra original publicada en 2014)

46. Bharat, A., Petkau, A., Avery, B. P., Chen, J. C., Folster, J. P., Carson, C. A., Kearney, A., Nadon, C., Mabon, P., Thiessen, J., Alexander, D. C., Allen, V., El Bailey, S., Bekal, S.,

German, G. J., Haldane, D., Hoang, L., Chui, L., Minion, J., … Mulvey, M. R. (2022). Correlation between Phenotypic and In Silico Detection of Antimicrobial Resistance in Salmonella enterica in Canada Using Staramr. *Microorganisms*, *10*(2), Article 2. https://doi.org/10.3390/microorganisms10020292

47. Zankari, E., Allesøe, R., Joensen, K. G., Cavaco, L. M., Lund, O., & Aarestrup, F. M. (2017). PointFinder: A novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *Journal of Antimicrobial Chemotherapy*, *72*(10), 2764-2768. https://doi.org/10.1093/jac/dkx217
48. Seemann, T. (2024). *Tseemann/mlst* [Shell]. https://github.com/tseemann/mlst (Obra original publicada en 2014)
49. Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., & Phillippy, A. M. (2019). Mash Screen: High-throughput sequence containment estimation for genome discovery. *Genome Biology*, *20*(1), 232. https://doi.org/10.1186/s13059-019-1841-x
50. Blin, K. (2023). *Ncbi-genome-download* (Versión 0.3.3) [Software]. Zenodo. https://doi.org/10.5281/zenodo.8192486
51. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015
52. Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J., & Harris, S. R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, *43*(3), e15. https://doi.org/10.1093/nar/gku1196

53. Katz, L. S., Griswold, T., Morrison, S. S., Caravas, J. A., Zhang, S., Bakker, H. C. den, Deng, X., & Carleton, H. A. (2019). Mashtree: A rapid comparison of whole genome sequence files. *Journal of Open Source Software*, *4*(44), 1762. https://doi.org/10.21105/joss.01762

54. Sanchez-Herrero, J. F., & mjsull. (2020). *spaTyper: Staphylococcal protein A (spa) characterization pipeline* (Versión v0.3.1) [Software]. Zenodo. https://doi.org/10.5281/zenodo.4063625

55. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

56. Raghuram, V., Alexander, A. M., Loo, H. Q., Petit, R. A., Goldberg, J. B., & Read, T. D. (2022). Species-Wide Phylogenomics of the Staphylococcus aureus Agr Operon Revealed Convergent Evolution of Frameshift Mutations. *Microbiology Spectrum*, *10*(1), e01334-21. https://doi.org/10.1128/spectrum.01334-21

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

58. III, R. A. P. (s. f.). *Introduction—Bactopia*. Recuperado 5 de noviembre de 2024, de https://bactopia.github.io/v3.1.0/

59. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics (Oxford, England)*, *34*(13), i142-i150. https://doi.org/10.1093/bioinformatics/bty266

60. J. Sola, P. (2020, febrero 24). *Common mash reference*.

61. Shen, W. (2024). *Shenwei356/csvtk* [Go]. https://github.com/shenwei356/csvtk (Obra original publicada en 2016)