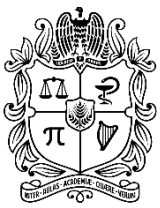

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Evaluación de herramientas Bioinformáticas útiles en la tipificación de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* a partir de datos de secuenciación de genomas completos

Oscar Eduardo Carabali Mosquera

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C, Colombia
2023

Evaluación de herramientas Bioinformáticas útiles en la tipificación de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* a partir de datos de secuenciación de genomas completos

Oscar Eduardo Carabali Mosquera

Trabajo de investigación presentado como requisito parcial para optar al título de:
Magister en Bioinformática

Director (a):
PhD. Emiliano Barreto Hernández

Línea de Investigación:
Bioinformática funcional y estructural
Grupo de Investigación:
Bioinformática – Instituto de Biotecnología

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C., Colombia
2023

A quienes creen que es posible…

Como bien dice Rubén Blades en su canción plástico "…Estudia, trabaja, sé gente primero Allí está la salvación…"

.

Declaración de obra original

Yo declaro lo siguiente: He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor.

Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores. Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto). Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

Firma

Fecha 01/06/2023

Agradecimiento

Una tesis de maestría es el resultado de entender, aprender, planificar y coordinar ideas para posteriormente plasmarlos en un trabajo escrito de investigación ejecutado con mucho esfuerzo. Un esfuerzo que, en tiempos de pandemia ha puesto a prueba la capacidad que tiene la sociedad de recuperarse, de levantarse cada día con la esperanzana de trabajar por futuro mejor creado nuevas oportunidades.

Deseo expresar un agradecimiento a Dios porque ha sido mi guía en tiempos de pandemia y en el arduo camino que he recorrido, llenándonos de sabiduría, tolerancia, fortaleza y amor por esta profesión.

De igual manera, ofrezco un agradecimiento al profesor Emiliano Barreto por compartir sus conocimientos, tener paciencia, dedicación y por permitir formar parte de su grupo de investigación.

Agradezco de manera especial, a mi madre Vilma Inés Mosquera, mi hija Isabela Carabali Zabala, a la familia Zabala y Cuellar, hermanos, familiares y amigos, por todo su apoyo en este proceso. sin su apoyo social, emocional y económico no hubiera encontrado la fuerza y energía para continuar este proyecto en tiempos tan difíciles como el que estamos cursando.

Finalmente agradezco a la Universidad Nacional de Colombia por brindarme todos los espacios académicos y de bienestar, los cuales permitieron adquirir conocimiento durante estos años a través de sus excelentes profesores.

Resumen

La Secuencia del Genoma Completo se obtiene mediante las tecnologías de secuenciación, especialmente las de próxima generación (NGS). Gracias a su alto poder discriminatorio, simplicidad, precisión, velocidad y flexibilidad, la aplicación de la Secuenciación de Genoma Completo (WGS) se ha convertido en una herramienta que aporta el nivel más alto hasta el momento de discriminación de cepas bacterianas para la investigación de brotes, permitiendo identificar estructura y composición de genes, variantes genéticas y reordenamientos del genoma entre otros (Kwong et al., 2015). Por tal motivo, esta herramienta es de gran utilidad para la investigación epidemiológica ya que proporciona información más detallada y precisa para la toma oportuna de medidas de control derivado de la identificación, tipificación, dinámica de transmisión, procedencia de infección y posibles patrones de propagación de brotes con bacterias como *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* causantes de Infecciones Asociadas a la Atención en Salud (IAAS), esto dado que estas bacterias poseen un alto grado de adaptabilidad fisiológica y elevados niveles de resistencia frente a numerosos agentes antimicrobianos, por lo que son patógenos con una elevada incidencia de morbilidad y mortalidad (Saharman et al., 2019)(Moradigaravand et al., 2017).

Dentro del uso más frecuente de los datos de Secuenciación de Genoma Completo (WGS) se encuentra la tipificación molecular bacteriana. Por la cual, se han desarrollado varios métodos que se basan principalmente en análisis derivados de *Ribosomal Multilocus Sequence Typing* (rMLST), *Core genome multilocus sequence typing* (cgMLST), *Whole genome multi locus sequence typing* (wgMLST), *core genome single nucleotide polymorphism* (cgSNP), *whole-genome single nucleotide polymorphism* (wgSNP) y *pangenome* (Coll et al., 2020)(Anani et al., 2020).

Estos métodos pueden variar en su resolución e idoneidad dependiendo de las especies. Sin embargo, debido al variado número de herramientas bioinformáticas útiles para tipificar y a la falta de consenso de evaluación comparativa de las herramientas los investigadores se pueden enfrentan con dificultades en la elección de herramientas indicada para sus actividades. Es por ello, que se hace necesario realizar una evaluación del desempeño de las herramientas útiles para tipificar, con el fin de informar al usuario sobre las mejores herramientas bioinformáticas disponibles actualmente que brinden información precisa y relevante.

Palabras clave: Secuenciación completa del genoma, *Klebsiella pneumoniae* y *Pseudomonas aeruginosa*, Tipificación molecular bacteriana, Herramientas Bioinformáticas, *Benchmarking*.

Abstract

Evaluation of Bioinformatics tolos useful in typing *Klebsiella pneumoniae* and *Pseudomonas aeruginosa* from sequencing data complete genomes

Whole Genome Sequence is obtained by sequencing technologies, especially next-generation sequencing (NGS). Due to its high discriminatory power, simplicity, precision, speed, and flexibility, the application of Whole Genome Sequencing (WGS) has become a tool that provides the highest level of discrimination of bacterial strains for outbreak investigation to date, allowing to identify both, structure and composition of genes, genetic variants, and rearrangements of the genome among others (Kwong et al., 2015). For this reason, this tool is highly useful for epidemiological research since it provides more detailed and precise information for the timely taking of control measures derived from the identification, classification, transmission dynamics, the origin of infection, and possible patterns of spread of outbreaks with bacteria such as Klebsiella pneumoniae and Pseudomonas aeruginosa that cause Health Care Associated Infections (IAAS), these bacteria have a high degree of physiological adaptability and high levels of resistance against numerous antimicrobial agents, which constitutes them as a pathological with a high incidence of morbidity and mortality (Saharman et al., 2019) (Moradigaravand et al., 2017).

One of the most frequent uses of data from Whole Genome Sequencing (WGS) is bacterial molecular typing. Consecuently, several methods have been developed are uptoday mainly based on analyzes derived from Ribosomal Multilocus Sequence Typing (rMLST), Core genome Multilocus Sequence Typing (cgMLST), Whole genome multilocus sequence typing (wgMLST), core genome Single Nucleotide Polymorphism (cgSNP), whole-genome single nucleotide polymorphism (wgSNP) and pangenome (Coll et al., 2020) (Anani et al., 2020).

These methods vary in their resolution and suitability depending on the species. However, due to the varied number of helpful bioinformatics tools for typing, and the lack of consensus benchmarking tools, researchers may face difficulties in choosing the right tools for their activities. For this reason, it is necessary to perform an evaluation of the performance of the tools for typing, to inform the user about the best bioinformatics tools currently available that provide accurate and relevant information.

Keywords: **Whole genome sequencing, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*, Bacterial molecular typing, Bioinformatics tools, Benchmarking**

Este Trabajo Final de maestría fue calificado en abril de 2023 por el siguiente evaluador:

Luis Fernando Niño Vásquez PhD.
Profesor Facultad de Ingeniería.
Universidad Nacional de Colombia.

Contenido

Lista de figuras

Pág.

Lista de Tablas

Pág.

Lista de Símbolos y abreviaturas

Abreviaturas

Abreviatura	Término
cgMLST	Tipificación de Secuencias Multilocus del Genoma Central
cgSNP	Polimorfismo de Nucleótido Único del Genoma Central
DNA	Ácido Desoxirribonucleico
DLST	Tipificación de Secuencia de Doble Locus
IAAS	Infecciones Asociadas a la Atención en Salud
INS	Instituto Nacional de Salud
MLST	Tipificación de Secuencias Multilocus
MLVA	Análisis de Repetición en Tándem de Número Variable de Múltiples Locus
NCBI	Centro Nacional de Información de Biotecnología
NGS	Secuenciación de Nueva Generación
PFGE	Electroforesis en Gel de Campo Pulsado
PCR	Reacción en Cadena de la Polimerasa
TRF	*Tandem Repeats Finder*
RAM	Resistencia a los Antimicrobianos
rMLST	Tipificación de Secuencia Multilocus Ribosoma
RNA	Ácido Ribonucleico
SNP	Polimorfismo de un Solo Nucleótido
WGS	Secuenciación de Genoma Completo
wgMLST	Tipificación de Secuencias Multilocus de Genoma Completo

Introducción

En la epidemiología clínica de enfermedades infecciosas, el seguimiento desempeña un papel importante para las investigaciones de brotes antiguos y emergentes. Es por ello, que en la última década se han tenido avances importantes en la epidemiología clínica basadas en métodos moleculares de enfermedades infecciosas permitiendo explicar los procesos biológicos básicos de transmisión, la diversidad microbiana y dinámica poblacional entre otros procesos (Mirande et al., 2018) (Riley, 2018). Sin embargo, la diversidad genética de las bacterias permite la rápida evolución dinámica y ecológicas de algunos agentes patógenos, creando así la necesidad de evaluar sus genes y la variación de nucleótidos, dando paso al surgimiento de a epidemiología clínica basada en métodos moleculares genotípicos, la cual proporciona información más detallada y precisa para la toma oportuna de medidas de control de brotes con bacteria como *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* causantes de Infecciones Asociadas a las Atenciones Sanitarias (IAAS). (Payne et al., 2020)(Moradigaravand et al., 2017).

Mediante los métodos basados en la secuenciación de ADN del genoma completo (WGS), se tiene una oportunidad tecnología poderosa para la investigación de brotes, cadenas de transmisión, propagación de clones y estructuras poblacionales de enfermedades asociadas con microorganismos a nivel local o global. Por ello, se requiere de la tipificación como método para caracterizar de manera genotípica las bacterias como los son wg/cgMLST, SNP y Pangenoma, los cuales utilizan las bandades de WGS (Jolley & Maiden, 2014). Lo que lleva a que se desarrollen cada día más un número importante de herramientas bioinformáticas como lstdbNet, SAMtools, Roary, goeBURST, GATK4, PanOCT, BIGSdb, Parsnp, PanACEA, SRST2, BactSNP, PIRATE, Seqsphere, kSNP3, stringMLST, Harvest, canowgMLST_BacComare, chewBBACA, MentaLiST útiles para realizar análisis de tipificación utilizando nucleótidos o genoma entre otros.

Sin embargo, dado al gran número de métodos, herramientas bioinformáticas y falta de conceso en la herramienta especifica, la elección de la herramienta indicada de acuerdo con los análisis requeridos por los investigadores, se ha convertido en un verdadero desafío (Escalona et al., 2016). Por lo que, en el presente trabajo realizó una evaluación comparativa del desempeño de las herramientas útiles para tipificar, con el fin de informar al usuario sobre las mejores herramientas Bioinformáticas disponibles actualmente y que brinden información precisa y relevante.

1. Capítulo 1.

1.1 Marco teórico.

La epidemiologia clínica ha adoptado las bondades de los datos de Secuenciación del Genoma Completo (WGS) de microorganismos, pasando rápidamente a la época de la epidemiología genómica. Lo que ha permitido avances importantes sobre variaciones y relaciones genéticas, al igual que difusiones epidémica de los principales patógenos bacterianos que causan infecciones y aumento de la morbimortalidad a nivel mundial (Michael Dunne et al., 2018). La comunidad científica ha Identificado con el acrónimo *ESKAPE,* a un grupo de diferentes especies de bacterias caracterizadas por ser cada día más resistentes a los antibióticos disponibles, entre ellas se encuentran las bacterias sujeto de este estudio *Klebsiella pneumoniae y Pseudomonas aeruginosa* de gran importancia para la epidemiologia y la salud pública (De Rosa et al., 2015). Debido a su alta resistencia a muchos antimicrobianos y su alta patogenicidad, que las han convertido en unos de los principales patógenos responsables de brotes y muerte de paciente derivados de infecciones asociadas a la atención en salud (IAAS) (Anani et al., 2020)(Royer et al., 2020). Entre las patologías más comunes producidas por estos, se encuentran las neumonías, abscesos hepáticos, infecciones del tracto urinario, endocarditis, bacteriemia entre otras infecciones (J. Liu et al., 2018)(van Dorp et al., 2019).

Por esta razón, la organización mundial de la Salud (OMS) ha catalogado la resistencia a los antimicrobianos como un problema de salud pública, recomendando una serie de estrategias que buscan disminuir la diseminación de organismos resistentes (INS, 2018). Dentro de estas estrategias se encuentra el seguimiento a las dinámicas de transmisión, establecer las relaciones genéticas, evolución genómica entre otras, las cuales se pueden realizar a través de la identificación y tipificación molecular (Moradigaravand et al., 2017) (Martin Vestergaard et al., 2016).

1.1.1 Identificación y Tipificación.

La identificación y tipificación microbiana son fundamentales particularmente para el diagnóstico, tratamiento y vigilancia epidemiológica de las infecciones, ya que por medio de estos se puede identificar y caracterizar rápidamente las bacterias en brotes de enfermedades y realizar pruebas de resistencia antimicrobiana (Uelze et al., 2020). Algunos métodos de identificación y tipificación microbiana se han convertido en procedimientos microbiológicos de rutina en hospitales e institutos de investigación reportados en la literatura desde la década de 1850 y 1880, en donde se realizaron importantes avances en las técnicas de aislamiento y cultivo de organismos bacterianos, lo que permitió a los microbiólogos diferenciar claramente las bacterias impulsando así el desarrollo de la taxonomía procariota. Inicialmente, las propiedades fisiológicas, bioquímicas y otras fenotípicas sirvieron como marcadores para la identificación de especies. En la década de 1930, la serotipificación fue uno de los primeros enfoques para diferenciar bacterias basándose en reacciones antígeno-anticuerpo a nivel de especie y subespecie y a principios de la década de 1980 se desarrollaron otros métodos de

tipificación de cepas bacterianas basados en Ácidos Nucleicos como moléculas marcadoras. Desde entonces, se han desarrollado muchos métodos de tipificación basados en secuencias para la caracterización de patógenos (Larsen et al., 2014)(Uelze et al., 2020).

Como concepto, la identificación es la utilización práctica de un esquema de clasificación para determinar la identidad de un aislamiento como miembro de un taxón específico o como un miembro de una especie que no había sido identificada previamente. Siendo el último paso en la sistemática de la taxonomía, ya que es a la vez el acto y el resultado de establecer si las cepas pertenecen a una taxa establecida y publicada en donde las bacterias se identifican de forma rutinaria mediante pruebas morfológicas y bioquímicas, complementadas según sea necesario con pruebas especializadas como la serotipificación y los patrones de inhibición de antibióticos. Las nuevas técnicas moleculares permiten identificar las especies por sus secuencias genéticas, a veces directamente de la muestra clínica (A. K. Gupta, 1996)(Zhi et al., 2011)(Bou et al., 2011).

En cuanto a la tipificación bacteriana, se podría definir como un método sistemático que permite caracterizar de manera fenotípica o genotípica más detalladamente las bacterias con el propósito de determinar la similitud genómica entre dos o más cepas de la misma especie, la cual la clasifica en tipos para describir al grupo cuyos integrantes exhiben ciertas características similares y posibilita su distinción frente a ejemplares de especies diferentes (Uelze et al., 2020)(Marcos-Zambrano et al., 2014)(W. Li et al., 2009). La tipificación fenotípica permite determinar la morfología de las colonias en varios medios de cultivo, pruebas bioquímicas, serología entre otros (W. Li et al., 2009). Sin embargo, no son lo suficientemente variables como para discriminar entre cepas estrechamente relacionadas. Por tal motivo, la tipificación de microorganismos se ha desplazado hacia los métodos genotípicos, ya que presentan mayor resolución en el momento de realizar la discriminación precisa y confiable entre cepas bacterianas en función de su contenido genético (Pérez-Losada et al., 2018).

En la actualidad, la tipificación genómica de microorganismos se utiliza ampliamente para dilucidar dinámicas evolutivas, relaciones genómicas, relaciones filogenéticas, genética de poblaciones de microorganismos y discriminación entre genotipo (Payne et al., 2020). Abarcando varias áreas de la investigación microbiológica aplicada. (Van Belkum et al., 2001) (Vaz et al., 2014).

1.1.2 Métodos de tipificación.

La tipificación de microrganismos, se pueden dividir en fenotípicos y genotípicos, siendo los más utilizados en la actualidad por su mayor resolución los métodos genotípicos en el momento de realizar la discriminación y clasificación de cepas relacionadas, dinámica poblacional, diversidad genómica, seguimiento de brotes específicos y vías de transmisión de enfermedades entre otras aplicaciones (Mirande et al., 2018) (Van Belkum et al., 2001).

Un método de tipificación ideal debe tener validez, discriminabilidad y reproducibilidad. Adicionalmente, debe ser rápido, rentable y fácil de realizar (W. Li et al., 2009). Sin embargo, los métodos de tipificación disponibles en la actualidad no son universales para todos los microorganismos, porque cada uno tiene diferente poder discriminatorio, costo, equipo, facilidad de uso y experiencia necesaria lo que genera ventajas y desventajas para el estudio de las cepas. (Uelze et al., 2020).

1.1.2.1 Métodos de tipificación fenotípicos.

Los métodos fenotípicos se basan en la determinación de características bioquímicas y fisiológicas, siendo una herramienta útil en la identificación de microorganismos que se basan en la determinación de actividades enzimáticas, capacidades metabólicas, determinantes antigénicos, propiedades bioquímicas y metabólicas entre otros parámetros. Debido a su bajo costo, este método se utiliza con frecuencia en laboratorios de microbiología como métodos de diagnósticos ya que permite el aislamiento del microorganismo, el estudio de sensibilidad a los antimicrobianos y facilita la aplicación de marcadores epidemiológicos. Sin embargo, estos métodos no identifican genes, polimorfismo o mutaciones que determinen la expresión de las características de los microrganismos (Merchán et al., 2017) (Bou et al., 2011).

1.1.2.2 Métodos de tipificación genotípicos.

Los métodos genotípicos se podrían utilizar para investigar brotes, identificar cadenas de transmisión, propagación de clones y estructuras poblacionales de enfermedades asociadas con microorganismos a nivel local o global. Actualmente, los métodos de genotipado se puede clasificar de la siguiente manera: I) Métodos basados en patrones de bandas de ADN, que clasifican las bacterias según el tamaño de los fragmentos generados, II) Métodos basados en secuenciación de ADN, que estudian el polimorfismo de las secuencias de ADN, III) Métodos basados en hibridación de ADN que utilizan sondas nucleotídicas. Estos métodos se basan en la detección del material genético del microrganismo siendo posible obtener resultados con mayor poder de resolución, sensibilidad y especificidad en comparación con los métodos fenotípicos (Merchán et al., 2017) (Pérez-Losada et al., 2018) (Pérez-Losada et al., 2018)(W. Li et al., 2009).

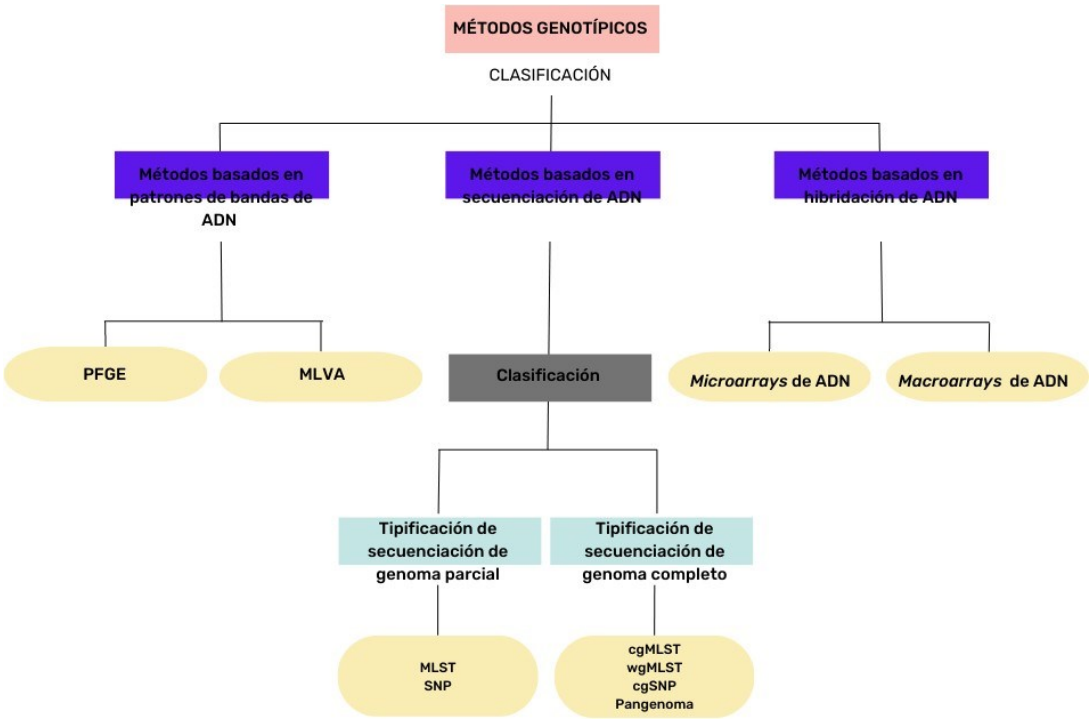

Figura 1 Métodos de tipificación genotípicos.

Métodos basados en patrones de bandas de ADN.

Estos métodos, permiten caracterizar las cepas de las bacterias en función de las diferencias en el tamaño de las bandas de ADN (fragmentos) generadas por amplificación de ADN genómico o por escisión de ADN utilizando enzimas de restricción (RE) (Uelze et al., 2020). Entre los enfoques más usados de acuerdo con la literatura publicada desde el año 2015 se encuentran:

- **Tipificación por electroforesis en gel de campo pulsado (PFGE).**

La PFGE es un método que permite el fraccionamiento de ADN de alto peso molecular entre 10 Kb a 10 Mb, es útil en tipificación de bacterias para el rastreo, investigación y seguimiento de brotes. La tipificación de bacterias se realiza por medio de patrones de similitud de cepas utilizando enzimas de restricción que cortan el ADN genómico con una baja frecuencia produciendo perfiles simples de bandeo entre 10-20 bandas cuando son separadas por electroforesis (Noller et al., 2003).

Estos patrones de bandas se comparan mediante herramientas bioinformáticas que permiten identificar similitudes entre aislados y determinar si tienen la misma procedencia evolutiva para cada representación en el dendograma. Si dos patrones no difieren en sus bandas, se considera que las cepas pertenecen al brote; si dos patrones difieren de 2 a 3 bandas, se consideran estrechamente relacionados y las cepas posiblemente hacen parte del brote; si dos patrones difieren de 4 a 6 bandas, se consideran posiblemente relacionados y las cepas posiblemente hacen parte del brote; y si dos patrones difieren en 7 o más bandas, se consideran diferentes y por lo tanto no hacen parte del brote (Tsai et al., 2017) (Botes et al., 2003)(W. Li et al., 2009).

Dentro de las herramientas bioinformáticas útiles en este método se encuentran el programa NTSYS-pc (*Numerical Taxonomy and Multivariate Analysis System*) para el análisis de los patrones en el dendograma, UPGMA (*Unweighted Pair-Group Method Using Arithmeti Average*) para realizar el coeficiente de similitud de Jaccard y el agrupamiento (Tenover et al., 1995)(Neoh et al., 2019). Sin embargo, este método tiene menor poder discriminatorio que el MLST y presenta grandes limitaciones ya que se requiere personal altamente capacitado para llevar a cabo la actividad, no se puede reproducir fácilmente, no discrimina los aislados no relacionados, algunas cepas no pueden ser tipificadas por PFGE y la interpretación de los patrones de banda puede ser subjetiva, entre otros parámetros (Salipante et al., 2015).

- **Tipificación por repetición en tándem de número variable de múltiples locus (MLVA).**

Es un método que permite subtipificación de microorganismos de acuerdo con el número de copias variables de repeticiones en tándem (VNTR) por locus. Esta repetición se determina mediante la amplificación por PCR al utilizar cebadores en las secuencias, las cuales generan fragmento de ADN de diferente tamaño de acuerdo con el número de unidades repetidas, generando un perfil de MLVA, el cual se define de acuerdo con el número de repeticiones de los loci VNTR. A este perfil se le asigna un código de un dígito múltiple llamado número de tipo MLVA, los cuales se almacenan en bases de datos para realizar la comparación de cepas y estudios epidemiológicos (Pérez-Losada et al., 2018)(Noller et al., 2003)(Saharman et al., 2019).

Dentro de sus ventajas se encuentra la capacidad para manejar volúmenes altos de muestras, la similitud de las agrupaciones, la fácil interpretación de los resultados, bajo costo, y requerir poco tiempo para su ejecución (Singh et al., 2020). Sin embargo, la diferencias en la elección de los loci, la nomenclatura, el tamaño del ampliación debido a las diferencias de cebador, plataforma y / o química, y la interpretación de repeticiones parciales o incompletas, falta de estándares para el desarrollo, validación y control de calidad, generan problemas en la comparación e interpretación de los resultados (Nadon et al., 2013).

El análisis de tipificación se realiza a través de diferentes herramientas bioinformáticas como la base de datos VNTRDB, la cual registra información completa sobre locus TRF, información genética relacionada e información comparativa entre secuencias altamente conservadas (Chang et al., 2007). MLVA-NET contiene datos de perfiles y aislados bacterianos, combina TRF con comparaciones de múltiples genomas para la identificación de loci VNTR que son polimórficos entre los genomas, y para *Pseudomonas aeruginosa,* cuenta con aproximadamente 100 tipos de VNTR (Guigon et al., 2008)(Youenou et al., 2014).

Métodos basados en secuenciación de ADN.

Estos métodos generan secuencias de nucleótidos y permiten la discriminación entre cepas bacterianas directamente de los genes o polimorfismos en su ADN, los cuales se pueden dividir en dos grupos; I) Tipificación de secuenciación de genoma parcial, el cual utiliza fragmentos del genoma secuenciado para poder hacer la tipificación. Entre los enfoques más comunes se encuentran la tipificación de *Multilocus Sequence Typing* (MLST), *single nucleotide polymorphisms* (SNP) y *doublé-locus sequence typing* (DLST); II) Tipificación de secuenciación de genoma completo, el cual ofrece la máxima resolución para la detección de nuevas variantes genéticas y estructurales, entre los enfoques más utilizados se encuentran el *Ribosomal Multilocus Sequence Typing* (rMLST), *Core genome multilocus sequence typing* (cgMLST), *Whole genome multilocus sequence typing* (wgMLST), *core genome single nucleotide polymorphism* (cgSNP), *whole genome single nucleotide polymorphism* (wgSNP). Adicionalmente, la tipificación por pangenoma permite usar los dos grupos con el fin de establecer relaciones filogenéticas y correlaciones del perfil genómico con el perfil fenotípico ofreciendo distintas alternativas para la investigación epidemiológica (Carattoli et al., 2014) (K. Zhou et al., 2016)(Jolley et al., 2012) (Perrin & Rocha, 2021)(Tsai et al., 2017).

Tipificación de secuenciación de genoma parcial.

▪ Tipificación de secuencias multilocus (MLST).

La tipificación de secuencias multilocus (MLST), es un enfoque molecular útil para estudios epidemiológicos globales o en análisis de la estructura poblacional de microorganismos, ya que permite establecer relaciones clonales entre cepas derivadas de brotes por medio de datos de población con varios algoritmos heurísticos que incluyen arboles *neighbor joining* y expansión mínima entre otros. El método evalúa la variación de nucleótidos en secuencias de fragmentos internos de siete genes de mantenimiento (loci) de aproximadamente 400-600 pb cada uno (Y. Y. Liu et al., 2019). Los alelos en cada uno de los siete loci definen el perfil alélico o el tipo de secuencia (ST) caracterizados por una serie de siete números enteros, que corresponden a los alelos en cada uno de los siete loci de mantenimiento. La discriminación entre aislamientos se determina a través de una

matriz de distancias que representan el número de diferencias de locus entre cada par de muestras (Jolley & Maiden, 2014) (Maiden, Rensburg, et al., 2013).

L as principales ventajas de los esquemas multilocus se relacionan con el hecho de que los datos de secuencia no generan ambigüedades, por lo que permiten se puedan comparar con perfiles alélicos en base de datos centrales de origen diverso, se puedan utilizar para investigar las relaciones clonales entre aislamiento, tanto a nivel global como local siendo fácilmente reproducibles y escalables (Platt et al., 2006) (Larsen et al., 2014).

▪ Polimorfismo de un solo nucleótido (SNP).

El SNP es una variación en un solo nucleótido que se produce en una posición específica del genoma y puede ocurrir cada 100 a 300 pb. Debido a su distribución, estas variaciones se pueden encontrar en cualquier parte de la estructura de los genes y el genoma. A su vez, los SNP se pueden clasificar en tres categorías de acuerdo con su función sobre los niveles de expresión génica los cuales se denominan; I) SNP reguladores (rSNP), los cuales se pueden encontrar en los genes que sintetizan proteínas y su variación puede afectar la expresión génica; II) SNP estructurales (srSNP) se pueden encontrar en transcritos que contienen intrones y transcritos que ya no contienen intrones, variación que puede a afectar la estructura y función de los ARN, la regulación de la traducción de los ARNm a proteínas, la funcionalidad de las proteínas y la estabilidad de los ARNm entre otros parámetros; y III) SNP codificantes (cSNP) se ubican en los exones y se pueden clasificar en sinónimos y no sinónimo de acuerdo a la afectación que la traducción tenga de nucleótidos y aminoácidos (Hall, 2014) (Altmann et al., 2012)(Blanc et al., 2020).

Este método se ha convertido actualmente en una herramienta útil para genotipificar microorganismos basado en la secuenciación de ADN, en donde la secuencias se alinean con un genoma de referencia y se determinan las diferencias de nucleótidos en las regiones codificantes y no codificantes, considerando que las posiciones de referencia son aquellas que están cubiertas por todos los genomas de consulta (Sahl et al., 2016) (Miro et al., 2020). Por otra parte, la filogenias basadas en distancias SNP permite identificar diferencias de nucleótido único entre genomas de aislamientos proporcionando así marcadores estables de cambio evolutivo entre genomas (Gardner & Hall, 2013) (Blanc et al., 2020).

La detección de SNP se considera un método importante para discriminar microorganismos, ya que permite la identificación de marcadores genéticos específicos de la cepa y de esta manera se puede estudiar la relación evolutiva y la dinámica de las poblaciones (G. H. Zhou et al., 2005) (Magalhães et al., 2020). Sin embargo, La no elección de un genoma de referencia representativo, la limitada estandarización de los enfoques, la variación en los elementos genéticos móviles como los plásmidos, afectan la resolución y el análisis de SNP, los cuales deben utilizar más del 95% del genoma secuenciado para tener una adecuada cobertura (Seth-Smith et al., 2019)(Kozyreva et al., 2017) (Timme et al., 2017).

- **Tipificación de secuencia de doble locus (DLST).**

DLST es un enfoque rápido y simple basado en secuencia de ADN para tipificar bacterias a nivel local, Este método se basa en la secuenciación de aproximadamente 500 pb de dos loci altamente variables. Para cada locus, se asigna un número arbitrario a cada alelo y la combinación de ambos alelos constituye el tipo DLST (Basset & Blanc, 2014). Sin embargo, este método no es suficientemente discriminatorio para distinguir entre brotes recientes y resolver grupos de cepas pertenecientes al mismo clon, no está disponible para todas las bacterias y se requiere del uso de otros métodos de tipificación (Basset & Blanc, 2014) (Tissot et al., 2016).

Tipificación de genoma completo.

- **Tipificación de secuencia multilocus ribosomal (rMLST).**

rMLST es un enfoque de genotipificación que abarca toda la diversidad bacteriana, desde el dominio hasta la cepa. Este enfoque indexa 53 genes que codifican las subunidades de proteínas del ribosoma bacteriano denominados como genes rps (*ribosome protein subunits*). Para discriminar microrganismos, emplea secuencias de referencia seleccionadas para identificar variantes de genes (loci rps) ya que están presentes en todas las bacterias, distribuidos en todo el cromosoma y codifican proteínas que son consideradas como conservadas (Jolley et al., 2012)(Larsen et al., 2014). Dado a que el enfoque rMLST integra métodos taxonómicos y de tipificación en un esquema MLST, permite identificar la distribución en varias localizaciones cromosómicas, interpretar la diversidad de las bacterias con una variedad de modelos evolutivos y proporciona información de tipificación definitiva y completa, útil en la investigación epidemiológica (Pérez-Losada et al., 2018) (Alikhan et al., 2018)(Maiden, Van Rensburg, et al., 2013).

Adicionalmente, este enfoque tiene algunas ventajas sobre otros métodos de tipificación basados en secuencias del genoma completo, dado a que no requiere genomas de referencia, es escalable, se puede utilizar para analizar aislamientos muy divergentes (Pérez-Losada et al., 2018). Sin embargo, el rMLST presenta desventajas para combinar genes suficientemente variables en un esquema de tipificación para mapear las relaciones filogenéticas, es costoso y requiere mucho tiempo para su ejecución (Alikhan et al., 2018)(Jolley & Maiden, 2014).

- **Tipificación de secuencia multilocus de genoma central (cgMLST).**

El cgMLST es un enfoque gen por gen que compara genomas utilizando más de 1000 loci de genes conservados o fijos en todo el genoma específicos para cada especie. En la actualidad, este enfoque se puede dividir en esquemas que proporcionan una nomenclatura pública ampliable y ad hoc (Pérez-Losada et al., 2018)

(van Beek et al., 2019)(Kimura, 2018). Este método utiliza el llamado de alelo, busca cada locus en el ensamblaje y coincidencias con una secuencia de alelos existente, asignando un número de alelo. En el caso de que no se contenga una secuencia de referencia de alelos, se puede crear un número de alelo y agregarlo a la secuencia de alelos para consultas futuras (Lüth et al., 2021)(Uelze et al., 2020).

Los análisis cgMLST se pueden realizar utilizando filogenia a través de agrupaciones jerárquica de enlace único, árboles de unión de vecinos (NJ) o de expansión mínima (MS). En estos la similitud entre genomas se obtiene de sus perfiles de alelos y se calcula el número total de alelos diferentes. Las diferencias se determinan por pares mediante comparación cruzada para todas las muestras, con lo que se genera una matriz de distancia, a partir de la cual se puede calcular un árbol filogenético (Deneke et al., 2021)(van Beek et al., 2019). El cgMLST ha demostrado ser un esquema útil para la identificación de brotes y relaciones clonales a nivel local y global, proporcionando datos de alta resolución en un grupo de aislados relacionados, pero no idénticos. Sin embargo, este enfoque presenta algunas limitaciones dado a que no está disponible para todas las bacterias y requiere una base de datos centralizada de alelos para hacer comparaciones entre laboratorios (Deneke et al., 2021)(Papić et al., 2021).

- **Tipificación de secuencia de múltiples locus de genoma completo (wgMLST).**

El enfoque wgMLST adicional al cgMLST, utiliza normalmente 1500 a 4000 loci centrales y accesorios proporcionando una resolución alta para grupos estrechamente relacionados. Dado a que la matriz de distancias se calcula en un conjunto más grande de loci, se pueden identificar genes que presentan diferencias entre las cepas comparadas y de esta manera, se proporciona información sobre funciones bacterianas críticas, como la patogenicidad y la virulencia (Y. Y. Liu et al., 2019) (Tadee et al., 2018). La tipificación por wgMLST identifica diferencias de nucleótidos (SNP, VNTR e INDEL) para cada marco de lectura abierto (ORF) de un organismo, lo que permite comparaciones de todo el genoma (Kingry et al., 2016). Los estudios demuestran que los resultados derivados de los enfoques wgMLST y cgMLST suelen ser similares, ya que no demostraron diferencias estadísticamente significativas en su capacidad discriminatoria (Blanc et al., 2020). Para la aplicación práctica, se puede utilizar un primer análisis cgMLST en un conjunto de datos diverso de una especie seguido de wgMLST para cepas estrechamente relacionadas.(Kingry et al., 2016)(Papić et al., 2021).

Sin embargo, no es útil usar un enfoque wgMLST para tipificación de bacterias con genomas abiertos, esto dado a que los genes accesorios son extremadamente volátiles debido a la adquisición repetida y la pérdida de ADN dentro de las cepas individuales. Por lo que se puede afectar la discriminación (Y. Y. Liu et al., 2016)(Martínez-Carranza et al., 2020).

- **Polimorfismo de nucleótido único del genoma central (cgSNP).**

El enfoque cgSNP se podría definir como secuencias ortólogas conservadas en todos los genomas alineados a partir de los genomas completos ensamblados. El análisis se realiza por medio de un árbol basado en cgSNP utilizando el método de unión de vecinos, en donde las distancias se calculan sobre la base de la porción del genoma de referencia que correspondía al genoma central de la especie, tomado en consideración solo las posiciones donde un nucleótido está presente en todos los aislamientos del conjunto de datos del genoma central (Yoshimura et al., 2019)(Coll et al., 2020).

- **Polimorfismo de nucleótido único de genoma completo (wgSNP).**

wgSNP es un enfoque de genotipado de cepas bacterianas a partir de lecturas sin procesar de WGS y el genoma de referencia. Para obtener una adecuada resolución, el genoma de referencia debe ser lo más cercano posible a las secuencias de muestra. Para calcular la distancia se utiliza un gráfico de escala multidimensional (MDS) o un árbol de unión de vecinos (NJ) después de la extracción de los elementos genéticos móviles para generar una buena presentación de los grupos de brotes (Coll et al., 2020)(Tsai et al., 2017).

- **Tipificación por Pangenoma.**

El pangenoma hace referencia a un conjunto de todos los genes de todas las cepas de una especie, los cuales pueden presentar variación de contenido genético entre cepas estrechamente relacionadas, el cual puede definir como abierto o cerrado según la capacidad de la especie para adquirir genes exógenos (Rouli et al., 2015). Este se puede dividir en tres componentes: I) El genoma central, II) Genes accesorios, III) Genes únicos, que están presentes solo en una sola cepa. Los cuales permiten estudiar diferentes características como el resistoma, mobiloma, el metabolismo global, redefinir las especies y clasificar según su contenido genómico entre otros parámetros (Anani et al., 2020)(Vernikos, 2020). El enfoque del pangenoma es una herramienta importante que se centran en identificar la presencia o ausencia de genes, y de esta manera permite estudiar las diferencias genéticas entre distintas cepas de las mismas especies (Rouli et al., 2015)(Perrin & Rocha, 2021).

Métodos basados en la hibridación de ADN.

Este método se basa en la hibridación de las cadenas complementarias de ácidos nucleicos para formar complejos de las cepas bacterianas por medio de oligonucleótidos o sondas marcadas con radioisótopos, enzimas o quimioluminiscentes. La discriminación bacteriana se realiza analizando la hibridación de su ADN con sondas de secuencias conocidas en donde intervienen dos elementos principales, sondas y dianas. Las sondas

son fragmentos de ADN de secuencias conocidas, mientras que las dianas son los ácidos nucleicos libres cuya identidad y abundancia se detectan mediante hibridación con sondas marcadas con fluorescencia en función de su complementariedad. Sin embargo, este método carece de una adecuada sensibilidad para la detección de microorganismos directamente de muestras clínicas, ya que requieren de una carga microbiana mínima de 104 unidades del genoma buscado. Entre los enfoques más usados se encuentran los microarrays y macroarrays de ADN (W. Li et al., 2009)(Bogaerts et al., 2019).

1.1.3 Tecnologías de secuenciación de genoma completo (WGS).

Los avances tecnológicos iniciales se centraron en mejorar el método de secuenciación de terminación de cadena publicado por Sanger en 1977 el cual genera fragmentos de secuencia de menos de 1000 pb. Por tal motivo, la búsqueda de nuevos métodos más eficientes para secuenciar lecturas largas y complejas de ADN se ha vuelto una prioridad para los investigadores, utilizando métodos que incluyen etiquetado fluorescente de moléculas, utilización de instrumentos basados en capilares y la automatización de estos procesos para permitir el análisis de múltiples muestras en paralelo (Kwong et al., 2015).

Recientemente, las tecnologías de secuenciación de próxima generación (NGS) como *Roche* 454, *Ion Torrent*, *Illumina*, *Pacific Biosciences* y *Oxford Nanopore*, tienen la capacidad de llegar secuenciar más 100 genomas bacterianos en una sola ejecución, lo que puede tardar entre 1 y 3 días (Kwong et al., 2015)(Pérez-Losada et al., 2018). Esta secuenciación paralela produce grandes cantidades de datos y con algunas tecnologías como *Pacific Biosciences* y *Oxford Nanopore*, se obtienen lecturas más largas (> 5000 pb), lo que facilita el ensamblaje de los genomas microbianos, permitiendo mejorar la secuenciación de genoma completo (WGS) de los microorganismos, disminuyendo costo y siendo más eficiente en comparación a los primeros métodos de secuenciación (Seth-Smith et al., 2019), los que ha impulsado el uso la genómica en microbiología clínica y salud pública. Los datos derivados de la secuenciación del genoma completo (WGS) se ha vuelto una poderosa herramienta, ya que por medio de esta se puede hacer comparación de aislamientos en el análisis de brotes, tipificación de bacterias, determinar clonalidad, diversidad poblacional, análisis filogenético, entre otros parámetros, siendo de gran utilidad para la vigilancia epidemiológica (Allard, 2016).

1.1.4 Herramientas Bioinformáticas.

Las herramientas bioinformáticas pueden definirse como cualquier software, procesador, algoritmos, base de datos, lenguaje de programación y navegadores entre otros, en especial aquellas herramientas utilizadas para adquirir, almacenar, organizar, archivar, analizar y visualizar dichos datos, útiles para la gestión y el análisis de datos biológicos y de salud (Misra et al., 2019).

Gracias a los avances en las tecnologías de secuenciación de próxima generación (NGS), se ha transformado la investigación y el análisis en las ciencias ómicas en los últimos años (Escalona et al., 2016). Siendo un área con desarrollos constantes, las herramientas bioinformáticas son de gran utilidad para la práctica científica en cualquiera de las ciencias ómicas computacionales. Hasta el 2014 se han reportado el desarrollo de más de 4400 herramientas computacionales para el análisis de datos de secuenciación del genoma, las cuales incluyen bases de datos, interfaces, herramientas de software para hacer control de calidad, alineación, ensamblaje, mapeo, filogenia entre otras (Henry et al., 2014) . Sin embargo, la creciente dependencia de los científicos a estas poderosas herramientas podría estar generando dificultades en el momento de escoger las herramientas más adecuadas para tareas analíticas específicas y tipos de datos con los que se esté trabajando (Mangul et al., 2019).

2. Capítulo 2.

2.1 Estado del arte.

2.1.1 Tipificación.

De acuerdo con la revisión bibliográfica de los últimos 5 años, se podría decir que los métodos más utilizados para tipificar *Klebsiella pneumoniae y Pseudomonas aeruginosa* son cgMLST y wgMLST seguido de la combinación de enfoques como cgMLST/cgSNP en menor proporción, MLST y SNP solos o combinados los cuales no utilizan datos de WGS. En cuanto a pangenoma, la proporción de uso para tipificar es más baja que los anteriores.

Finalmente, todos los métodos que utilizan los datos de WGS usan agrupamientos para definir tipos dentro de un conjunto de cepas analizar, los cuales proporcionan conclusiones similares en cuanto la tipificación (Pérez-Losada et al., 2018). Sin embargo, esta podría variar de acuerdo con su resolución e idoneidad para cada especie, esto dado a que los enfoques cgMLST/wgMLST cuentan con nomenclatura de alelos organizados en bases de datos públicas para la asignación de secuenciotipos, la cual hace que sus resultados sean comparables entre laboratorios, a diferencia de los métodos cgSNP/wgSNP y pangenoma. Los cuales, no cuentan con una base de datos de secuenciotipos por la cual solo se podría utilizar localmente y la caracterización de microorganismo se basarían en criterios de los usuarios (Uelze et al., 2020) .

2.1.2 Tipificación de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* basadas en WGS.

Los brotes de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* representan una serie de amenazas para los pacientes hospitalizados a nivel mundial, particularmente los

pacientes susceptibles a infecciones (Seth-Smith et al., 2019) (Martin et al., 2017). Comprender la transmisión a través de métodos de tipificación basados en datos de WGS, proporciona una herramienta importante para la gestión de brotes, identificación de propagación génica clonal y el análisis de estructuras de plásmidos de las bacterias sujetos de este estudio con una mayor resolución que permite comparar con los métodos de tipificación molecular convencionales incluidos los no basados en ADN (Chen et al., 2016) (Ruppé et al., 2017).

El género *Klebsiella* cuenta con 10 especies aproximadamente, dentro de ellas se encuentra; I) *Klebsiella pneumoniae*, II) *Klebsiella ozaenae*, III) Klebsiella terrigena, IV) *Klebsiella rhinoscleromatis*, V) *Klebsiella oxytoca*, VI) *Klebsiella planticola* VII) *Klebsiella ornithinolytica*, VIII) *Klebsiella aerogenes*, IX) *Klebsiella granulomatis*, X) *Klebsiella mobilis.* Para género *Pseudomonas,* este cuenta con 7 tipos de especies aproximadamente, dentro de ellas se encuentra*; I) P. aeruginosa, II) P. chlororaphis III) P. fluorescens, IV) P. pertucinogena, V) P. putida, VI) P. stutzeri VII) P. syringae.* Algunas de ellas, se han encontrado en muestras clínicas tomadas en humanos relacionadas con infecciones asociadas a la atención en salud (IAAS) como *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* (Jonas et al., 2004)(Peix et al., 2018)(Peix et al., 2009). Encontradas entornos como dispositivos médicos y superficies hospitalarias, causando aproximadamente el 14-20% de las infecciones relacionadas con el tracto respiratorio, conducto biliar inferior, heridas quirúrgicas y tracto urinario(Wang et al., 2020).

En Colombia, según el Instituto Nacional de Salud, la frecuencia de infección por *Klebsiella pneumoniae* en UCI adulto es del 17,4%, en UCI pediátrica 13,4% y en UCI neonatal 21,2%. En el servicio de hospitalización adulto las infecciones causadas por esta bacteria fueron el 28,4% del total de reporte de todos los microorganismos causantes de IAAS en el 2018, representando una amenaza grave y urgente para la salud pública debido a su alta resistencia a los antimicrobianos (INS, 2018).

Hasta la fecha, se han registrados en *taxonomy browser* del NCBI 29,579 genomas de la especie Klebsiella pneumoniae, los cuales 2117 cuentan con secuencias de genoma completo (Benson et al., 2018)(Nguyen et al., 2019) y 15,402 genomas para la especie *Pseudomonas aeruginosa*. De estos, 716 cuentan con secuencias de genoma completo (Das et al., 2015)(Benson et al., 2017).

El análisis de tipificación basado en WGS para *Klebsiella pneumoniae* y *Pseudomonas aeruginosa*, se puede realizar mediante tres principales estrategias: I) Métodos basados en alelos: Entre ellos se encuentran los enfoques rMLST, cgMLST y wgMLST, los cuales se basan en la consideración de las designaciones de secuenciotipos (ST) y alelos para estimar la relación entre los aislados. Encontrando registros de 4342 Perfiles ST en 7 genes y 78050 rST en 52 genes para la especie *Pseudomonas aeruginosa* en la base PubMLST(http://www.pubmlst.org); 5417 registros de scgST en 629 loci para la especie *Klebsiella pneumoniae* en la base Pasteur MLST (www.pasteur.fr/cg/mlst). 8899 tipos complejos (CT) en 2358 genes para la especie *Klebsiella pneumoniae* y 3409 tipos complejos (CT) en 3867 genes para la especie *Pseudomonas aeruginosa* registrados en

la base cgMLST.org (https://www.cgmlst.org/ncs). II) Métodos basados en nucleótidos. Entre ellos se encuentran los enfoques cgSNP y wg*SNP*, los cuales se basan en la aplicación directa de secuencias de nucleótidos para estimar la variación y los parámetros de población. En la búsqueda de información de este trabajo, no se encontraron secuenciotipos basados en SNP ni bases de datos de nomenclatura pública. III) Métodos basados en el pangenoma. Agrupan los genomas en función de la presencia o ausencia de genes permitiendo un análisis sensible y detallado de los aislados (W. Li et al., 2009)(Pérez-Losada et al., 2018)(Benson et al., 2017) (Deneke et al., 2021).

2.1.3 Tipificación de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* basadas en WGS.

De acuerdo con la literatura consultada, los métodos más usados para tipificar estas bacterias utilizando datos de secuenciación de genoma completo en los últimos 5 años son cgMLST, wgMLST, cgSNP y wgSNP.

Tabla 1 Esquemas y herramientas bioinformáticas útiles para tipificar que utilizan datos de secuenciación de genoma completo.

Herramientas para esquemas MLST	Herramientas para esquemas SNP	Herramientas para esquemas Pangenoma
lstdbNet	SAMtools	Roary
goeBURST	GATK4	PanOCT
BIGSdb	Parsnp	PanACEA
SRST2	BactSNP	PIRATE
Seqsphere	kSNP3	
stringMLST	Harvest	
canowgMLST_BacComare		
chewBBACA:		
MentaLiST		

Para esquemas MLST se registran 9 herramientas en un conjunto de datos comunes de genomas de procariotas como:

- mlstdbNet: Base de datos de tipificación de secuencias de múltiples locus (MLST), que usa un software que permite a múltiples bases de datos aisladas consultar las bases de datos PubMLST que contiene perfiles alélicos y definiciones de secuencias. Este permie que los laboratorios individuales establezcan bases de datos aisladas, cada una personalizada según las necesidades del proyecto en particular y con las restricciones de acceso apropiadas, mientras se mantienen los

beneficios de una única fuente definitiva de información de perfil y secuencia. mlstdbNet utiliza partes de una implementación temprana del software de base de datos MLST y se ejecuta en sistemas Linux utilizando la base de datos PostgreSQL y el servidor web Apache, tiene licencia GNU GPL y se puede consultar en http://pubmlst.org/software/database/mlstdbnet/ (Jolley et al., 2004).

- goeBURST: Es un algoritmo que utiliza el software PHYLOViZ con el fin de identificar patrones alternativos de descendencia para varias especies bacterianas mediante MLST. Utilizando un enfoque gráfico que asegura una solución óptima para la ubicación de enlaces entre tipos de secuencia. Esta implementación utiliza un marco de visualización, permite la división de un conjunto de datos en varios grupos de aislamientos relacionados, denominadas complejos clonales, mediante la implementación de un modelo simple de expansión y diversificación clonal. El software está disponible en http://goeBURST.phyloviz.net, no se ejecuta con línea de comando en sistemas Linux (Francisco et al., 2009).
- BIGSdb: Es un software diseñado para el almacenamiento, análisis y distribución de datos de tipificación de secuencias multilocus (MLST) de cepas bacterianas. Se puede vincular cualquier número de secuencias que pueden ser pequeños contigs ensamblados hasta genomas completos, donde se pueden definir un gran número de loci con alelos asignados por referencia en la base de datos PubMLST y MLST Pasteur. El software cuenta con licencia pública general (GNU), disponible en https://bigsdb.readthedocs.org, se puede usar de manera local mediante sistema operativo Linux (Francisco et al., 2009)
- SRST2: Es un programa diseñado para hacer la detección rápida y precisa de genes, alelos y tipos de secuencias de múltiples locus (MLST) a partir de datos de WGS, basados en lecturas obtenidas por metodología Illumina o en una base de datos de secuencias de genes como PubMLST o cgMLST.org. Cuenta con licencia de derechos de autor del software (licencia BSD), se puede ejecutar con sistema operativo LINUX disponible en https://github.com/katholt/srst2#installation (Inouye et al., 2014).
- Seqsphere: Es un Software de acceso cerrado diseñado ´para realizar la tipificación bacteriana con datos de secuenciación de genoma completo (WGS). El software trabaja directamente con los archivos FASTQ de lectura corta, el cual se conecta directamente al Servidor de nomenclatura alélica cgMLST.org, permitiendo

hacer investigación de brotes y vigilancia en tiempo real(Miro et al., 2020)(Seth-Smith et al., 2019).

- MentaLiST: Es un algoritmo basado en un algoritmo de votación k-mers para determinar los tipos de secuencia para esquemas MLST, cgMLST y wgMLST, a partir de datos de secuenciación de genoma completo. Es rápido, eficiente en memoria y no requiere ningún paso de preprocesamiento como ensamblaje, mapeo o construcción de árboles. Además, es robusto en situaciones de baja cobertura y la presencia de otras cepas en la muestra. MentaLiST abre la posibilidad de determinar los tipos MLST de cientos de cepas escribiendo miles de genes, en cuestión de horas. La construcción de la base de datos requiere un archivo de definición de perfil para el esquema de tipificación junto con secuencias de alelos para cada locus en el esquema; el usuario puede crear este archivo o recuperarlo de la base de datos PubMLST o cgMLST.org. Cuenta con licencia MIT que solo requiere la preservación de los derechos de autor y avisos de licencia. Se puede usar de manera local mediante línea de comando en sistema operativo LINUX y está disponible en https://github.com/WGS-TB/MentaLiST (Feijao et al., 2018).
- stringMLST: Es una herramienta de código abierto rápida basada en *k-mers* para la tipificación de secuencias de múltiples locus MLST, rMLST, cgMLST. Predice el ST de un aislado de una manera completamente libre de ensamblaje y alineación. El proceso de descubrimiento de ST puede dividirse conceptualmente en tres etapas: filtrado, recuento e informes. En la etapa de filtrado, stringMLST descarta una secuencia leída si el *k-mers* situado en el medio de la secuencia leída no tiene una coincidencia en la base de datos stringMLST. Las lecturas de secuencia cuyos k-mers del medio tienen una coincidencia se k-merizan en la etapa de conteo. Luego, se busca cada *k-mers* en la base de datos de PudMLST y cgMLST.org, para cada coincidencia, se registran los alelos y los loci correspondientes; se incrementa un contador para cada alelo cuyo constituyente k-mers coincidió. Una vez que se han procesado todas las secuencias, stringMLST identifica el alelo en cada locus con el valor máximo del contador para generar un perfil alélico y la correspondiente llamada ST. Cada secuencia de alelos específicos de locus se registra el alelo y loci correspondientes para cada k-mer. stringMLST es gratis para usuarios académicos y requiere permiso antes de cualquier uso comercial. Se puede usar de manera local mediante línea de comando en sistema operativo LINUX y está

disponible en https://github.com/jordanlab/stringMLST (Dalsass et al., 2019)(A. Gupta et al., 2017).

- cano-wgMLST_BacCompare: Es un servidor que emplea dos procesos principales, permitiendo la extracción de esquemas de genoma completo y refinamiento de loci discriminatorios, utilizando un algoritmo de importancia de características. Integrando varios módulos funcionales como la anotación de contig, creación de perfiles wgMLST, construcción de árbol de parentesco genético. El enfoque emplea dos características de dos capas para filtrar loci. En la primera capa, se extrae el esquema de genoma completo para las secuencias del genoma cargadas por el usuario. En la segunda capa, se aplica el "algoritmo de importancia de la característica" para seleccionar los loci más críticos que poseen una capacidad de distinción definida. No se puede usar de manera local y está disponible en http://baccompare.imst.nsysu.edu.tw/ (Y. Y. Liu et al., 2019).
- chewBBACA: Es un *pipeline* integral que incluye un conjunto de funciones para la creación y validación de esquemas de tipificación de secuencia múltiple de genoma completo y genoma central (wgMLST cgMLST). Utiliza un algoritmo de llamada de alelos basado en *Blast Score Ratio* que se puede ejecutar en configuraciones de multiprocesador y un conjunto de funciones para visualizar y validar la variación alélica en los loci. chewBBACA realiza la creación de esquemas y llamadas de alelos en genomas completos o en *draft,* resultantes de ensambladores de novo. El software realiza llamadas de alelos en cuestión de segundos a minutos por cepa en una computadora portátil, es fácilmente escalable para el análisis de grandes conjuntos de datos de cientos de miles de cepas utilizando opciones de multiprocesamiento, siendo una solución de código abierto eficiente en Python 3. Con licencia pública general (GNU), se encuentra disponible en https://github.com/B-UMMI/chewBBACA (Mamede et al., 2021)(Mamede et al., 2021).

Para los enfoques SNP, en la actualidad se han desarrollado numerosas herramientas bioinformáticas basadas en genomas de referencia para microorganismos, entre las más utilizadas de acuerdo con la revisión bibliográfica del presente trabajo se encuentran las siguientes seis:

- Snippy ref.: Es una herramienta ampliamente utilizada en estudios epidemiológicos para llamada de variantes de SNP a través de un genoma de referencia y secuencias de interés, en la cual se realiza una alineación de SNP central que se

pueden usar para construir una filogenia de alta resolución (https://github.com/tseemann/snippy). Adicionalmente, Snippy permite minimizar los falsos positivos o llamadas falsas dentro del proceso, esto dado a que realiza una profundidad de lectura mínima de 10 y una calidad de llamada de variante de 100, mejorando así el rendimiento (Bush, 2021).

- SAMtools: Es un paquete software que permite calcular la probabilidad del genotipo y llamar variantes mediante ecuaciones para muestras diploides. La identificación de los SNP se da cada vez que una lectura mapeada muestra una diferencia con el genoma de referencia. Incorpora diferentes tipos de información, como el número de lecturas diferentes que comparten una falta de coincidencia con la referencia, los datos de calidad de secuencia y las tasas de error de secuenciación esperadas. Cuenta con licencia MIT por la cual se otorga permiso, sin cargo, a cualquier persona que obtenga una copia de este software. Permite trabajar en línea de comando de manera local y está disponible en https://github.com/samtools/samtools (H. Li et al., 2009)(H. Li, 2011)(Coll et al., 2020).
- GATK4: Es una colección de herramientas de línea de comandos para analizar datos de secuenciación de alto rendimiento con un enfoque principal en el descubrimiento de variantes. Las bases de código GATK y Picard en un marco simplificado, permiten que las herramientas seleccionadas se ejecuten de forma masivamente paralela en *clústeres* locales o en la nube utilizando Apache Spark. Este repositorio es de código abierto con licencia Apache 2.0. necesita Java 8 y Python 2.6 o superior para poder ejecutar. Disponible en https://github.com/broadinstitute/gatk (Friedman et al., 2020)(Bathke & Lühken, 2021).
- Parsnp: Es una herramienta diseñada para alinear el genoma central de genomas bacterianos en poco tiempo utilizando genoma de referencia. Los archivos de entrada pueden ser ensamblajes genomas terminados y la salida incluye archivos de Árbol SNP del genoma central con formato Newick, SNP utilizados para inferir filogenia, Archivo binario con formato Gingr, alineación múltiple con formato XMFA. Gracias a las alineaciones múltiples (Treangen et al., 2014).

- BactSNP: Es una herramienta para identificar SNP entre cepas bacterianas. Puede detectar SNP con precisión y sensibilidad, genera archivo de salida TSV simple con información de SNP, así como un archivo FASTA. BactSNP utiliza lecturas de cada aislado y un genoma de referencia como entrada. Primero, las lecturas son ensambladas de novo por *Platanus* para cada aislamiento, y luego los *contigs* ensamblados se alinean contra el genoma de referencia por NUCmer. En segundo lugar, se determina el nucleótido correspondiente al genoma de referencia en cada sitio. Las variables alelo e *indel* representan el número de alelos alineados en el sitio y la distancia desde el *indel* más cercano al sitio, respectivamente. En cada sitio del genoma de referencia, el alelo correspondiente se determina como el alelo alineado. Actualmente, BactSNP solo está disponible en LINUX y se pueden descargar como paquete binario, paquete fuente o paquete RPM fuente. Todos los paquetes están disponibles en https://github.com/IEkAdN/BactSNP (Yoshimura et al., 2019).
- kSNP3: Es un software que permite identificar los SNP pangenoma en un conjunto de secuencias del genoma y estima árboles filogenéticos basándose en esos SNP. El descubrimiento de SNP se basa en el análisis de *k-mers* y no requiere alineamientos múltiples ni la selección de un genoma de referencia, por lo que kSNP puede tomar cientos de genomas microbianos como entrada. kSNP puede analizar tanto genomas completos (terminados) como genomas no terminados en contigs ensamblados o lecturas sin ensamblar. Los genomas terminados y no terminados se pueden analizar juntos, y kSNP puede descargar automáticamente archivos GenBank de los genomas terminados e incorporar la información de esos archivos en la anotación SNP. kSNP3 se puede implementar en LINUX y Mac OS X bajo la licencia BSD de código abierto y está disponible en https://sourceforge.net/projects/ksnp/files (Gardner & Hall, 2013)(Hall, 2014)(Gardner et al., 2015).
- Harvest: Es una herramienta adecuada para el análisis de genomas microbianos. Alberga tres módulos; Parsnp para análisis del genoma central, Gingr para visualización de salida y HarvestTools para metaanálisis. La suite Harvest es de código abierto y está disponible gratuitamente en https://github.com/marbl/harvest (Treangen et al., 2014).

La tipificación por medio del Pangenoma para bacterias es un método importante para establecer relaciones clonales entre cepas, la cual se puede realizar a través de las siguientes cuatro herramientas:

- Roary: Es una herramienta rápida para extraer pangenoma completos, conjuntos de genes centrales o diferencias entre genomas de referencia. Permite construir grandes Pangenomas produciendo resultados precisos atribuible a la utilización del contexto de la información conservada de la vecindad de genes. La entrada a Roary es un ensamblaje anotado por muestra en formato GFF3 como el producido por Prokka (Page et al., 2015). Las regiones codificantes se extraen y se convierten en secuencias de proteínas, se filtran para eliminar secuencias parciales y se agrupan de forma iterativa. Esto da como resultado un conjunto sustancialmente reducido de secuencias de proteínas, que se comparan de todas contra todas con BLASTP, , para agruparlas cuando tienen entre ellas un porcentaje de identidad de secuencia definido por el usuario. Roary se puede trabajar de manera loca de a través de línea de comandos en un sistema LINUX, cuenta con licencia pública general GNU, disponible en https://github.com/sanger-pathogens/Roary (Sitto & Battistuzzi, 2020) (Page et al., 2015).
- PanOCT: Es una herramienta escrita en PERL que realiza agrupación de ortólogos basada en gráficos para el análisis de Pangenoma de genomas de procariotas estrechamente relacionadas. Utiliza información de vecindad de genes conservados y homología para ubicar las proteínas en grupos ortólogos por medio de Blast+, generando así un Pangenoma consenso. La herramienta se puede ejecutar de manera local a través de línea de comandos en un sistema LINUX. Adicionalmente, cuenta con licencia pública general GNU, disponible en https://github.com/JCVenterInstitute/PanGenomePipeline (Fouts et al., 2012)(Clarke et al., 2018).
- PanACEA: Es una herramienta que permite visualizar regiones centrales y variables del pangenoma, que consiste en visualizaciones jerárquicas de varios niveles que se extienden desde Pancromosomas a regiones centrales y variables a genes individuales. Las regiones y los genes se anotan funcionalmente para permitir la búsqueda rápida y la identificación visual de las regiones de interés con

la opción de incorporar filogenias genómicas y metadatos. Es una herramienta de código abierto y se puede usar localmente mediante línea de comandos en un sistema LINUX, cuenta con licencia pública general GNU disponible en https://github.com/JCVenterInstitute/PanACEA (Clarke et al., 2018)(Vernikos, 2020).

- PIRATE: con sus siglas en Iterative *Refinement and Threshold Evaluation*, es un software escalable útil para construir Pangenoma a partir de formatos de entrada GFF3, en donde se evalúa y clasifica la diversidad genética dentro del pangenoma. PIRATE proporciona medidas de divergencia de secuencia y diversidad alélica dentro de la muestra, las cuales permiten identificar y clasificar familias de genes ortólogos en pangenoma bacterianos en una amplia gama de umbrales de similitud de secuencia. PIRATE está implementado en Perl y se puede ejecutar localmente mediante línea de comandos en un sistema LINUX. Adicionalmente, está disponible gratuitamente bajo una licencia de código abierto GNU GPL 3 en https://github.com/SionBayliss/PIRATE (Bayliss et al., 2019).

Si bien no son las únicas herramientas útiles para crear Pangenoma, son las que más uso presentan en los últimos años de acuerdo con la revisión bibliográfica de este trabajo (Rouli et al., 2015)(Vernikos, 2020).

2.2 Contexto de la evaluación comparativa o *Benchmarking.*

La evaluación comparativa se considera un proceso en el cual se crean, recopilan, analizan y comparan productos, servicios, indicadores entre otros, con el fin de medir su rendimiento e identificar un punto de referencia (Robinson & Vitek, 2019). El concepto de *Benchmarking* fue implementado por la empresa Xerox en un esfuerzo por reducir sus costos de producción. En la década de 1980 el concepto se extiende por todo el sector industrial y se empieza a usar como una herramienta de autoevaluación y apoyo para la toma de decisiones (Ettorchi -Tardy et al., 2012).

En biología computacional y ciencias ómicas, los estudios de evaluación comparativa tienen como objetivo comparar rigurosamente el rendimiento de diferentes métodos utilizando conjuntos de datos de referencia bien caracterizados, para determinar las fortalezas de cada método o para proporcionar recomendaciones con respecto a las opciones adecuadas de métodos para un análisis (Altmann et al., 2012) (Weber et al., 2019). En los trabajos previos, se identifican varios *Benchmarking* para identificar errores en herramientas computacionales que realizan ensamblaje del genoma (Mangul et al., 2019) alineamiento, identificación de variantes, anotación de variantes y visualización,

presentando una descripción general de la funcionalidad, las características y los requisitos específicos de las herramientas a nivel individual (F. Li et al., 2019).

En la búsqueda de las herramientas bioinformáticas mediante la revisión de artículos científicos referentes a tipificación de diferentes aislados de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa*, no se encontró hasta el momento información referente a *Benchmarking* de herramientas bioinformáticas útiles para tipificación. Sin embargo, no se descarta que durante el desarrollo del trabajo se publiquen evaluaciones de herramientas bioinformáticas específicas utilizadas para tipificar, acorde con los métodos moleculares anteriormente descritos.

2.3 Justificación.

La genómica computacional se ha vuelto esencial para la investigación biológica moderna presentando muchos desafíos en el campo de la bioinformática, esto dado a las nuevas tecnologías que producen datos con mayor tamaño y complejidad, las cuales estimulan avances masivos en todas las áreas de la ciencia de datos (Robinson & Vitek, 2019). De igual manera, la creciente dependencia de los científicos de estas poderosas herramientas que permiten caracterizar microrganismos patógenos (Sczyrba et al., 2017), hacen que se desarrollen nuevos enfoques computacionales a partir de datos de secuenciación de genomas completos para manejar conjuntos de datos grandes, complejos y ruidosos (Mangul et al., 2019). Dado a que se plantean diferentes alternativas para resolver algunos problemas, siempre están apareciendo múltiples herramientas para un mismo propósito, las cuales pueden afectar los análisis biológicos posteriores e incluso producir resultados no conformes (Robinson & Vitek, 2019)(Buchka et al., 2021).

Sin embargo, el rápido crecimiento de las ciencias computacionales hace que sea difícil elegir herramientas útiles para tipificar microrganismos adecuados con el fin desarrollar conocimientos científicamente rigurosos (Zheng, 2017). Es por ello, que se hace necesario realizar una evaluación sistemática de las herramientas computacionales con el fin de verificar la exactitud, precisión, especificidad y desempeño entre otros parámetros, usando un conjunto de datos de genomas completos de referencias, dado a que cada herramienta tiene requisitos funcionales de entrada y salida diferentes. Adicionalmente, el rápido desarrollo y publicaciones de nuevos métodos dicta la necesidad de una evaluación continua (Zheng, 2017) (Weber et al., 2019), que de no realizarse, podría favorecer una inadecuada utilización de las herramientas por parte de los investigadores en el momento de analizar secuencias y obtener información biológicamente relevante.

De acuerdo con lo anterior, se realizó este trabajo de *Benchmarking* para lograr seleccionar una o varias herramientas de análisis útiles en la tipificación de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa*, con el fin de informar al usuario sobre las mejores herramientas bioinformáticas disponibles de acuerdo a diferentes enfoques descritos previamente, de manera que se disponga de información precisa y relevante para la selección de dichas

herramientas, lo que redundará en una mejor toma de decisiones por parte del equipo de salud y epidemiologia, en la detección de brotes y selección de tratamientos.

Con base en lo anterior se planteó la siguiente pregunta de investigación.

¿Cuáles son las diferencias en el desempeño de las herramientas bioinformáticas utilizadas para tipificación de aislamientos de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* utilizando datos de secuenciación de genomas completos, en función de su exactitud, precisión y especificidad?

3. Capítulo 3

3.1 Objetivos.

3.1.1 General.

Evaluar el desempeño de las herramientas bioinformáticas disponibles para la tipificación de aislamientos de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* utilizando datos de secuenciación de genomas completos.

3.1.2 Específicos.

- Identificar las herramientas bioinformáticas reportadas en los último cinco años en la literatura especializada, útiles para la tipificación de *K. pneumoniae* y *P. aeruginosa* a partir de datos de secuenciación de genomas completos, que puedan ser implementadas en un servidor local con sistema operativo LINUX.
- Implementar las herramientas bioinformáticas seleccionadas en un servidor local con sistema operativo LINUX.
- Comparar el desempeño de las herramientas bioinformáticas implementadas de acuerdo con su exactitud; precisión y especificidad.

4. Capítulo 4

4.1 Diseño metodológico.

El presente estudio es de tipo descriptivo dado a que se especifican las características cualitativas de las herramientas recolectadas, con un diseño experimental compuesto por tres fases que se realizaran de manera sistemática, cuya estrategia para el desarrollo de las actividades es de carácter mixto dado a que se realizaran mediciones de las métricas de funcionamiento de cada herramienta (Luis, 2012).

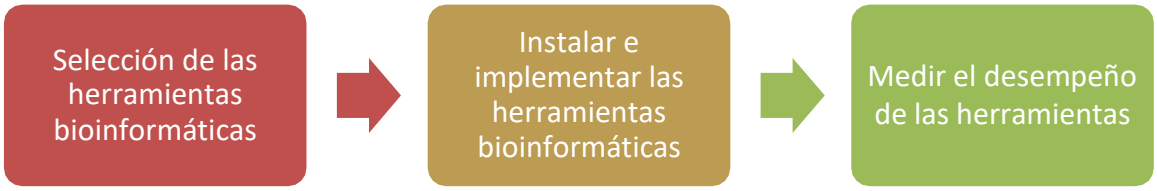

Figura 2 Fase metodológicas del proyecto.

4.1.1 Fase 1. Selección de las herramientas bioinformáticas.

De acuerdo con la figura 2, la primera fase metodológica se realizará en dos etapas, iniciando con la selección de las diferentes herramientas bioinformáticas útiles para la tipificación de microorganismos bacterianos descritas en el estado del arte y en la figura 3, las cuales están disponibles en la literatura especializada como revistas, y páginas web (Tabla 1).

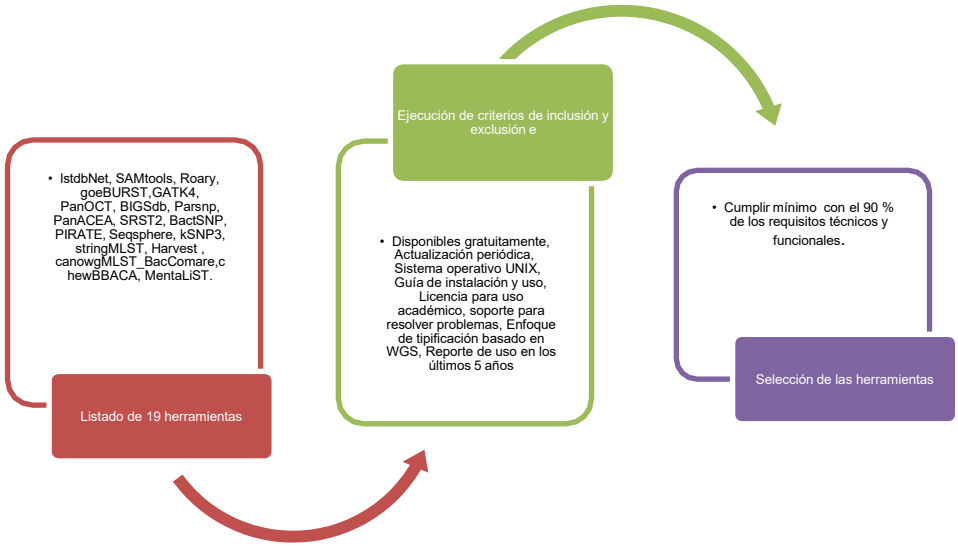

Figura 3 Búsqueda de la información.

útiles para la tipificación de microorganismos bacterianos Se escogerán las herramientas descritas en el estado del arte y la figura 3 para aplicar los 9 criterios de inclusión, en donde a través de una clasificación binaria se le asigna el numero 1 a la herramienta cuando cumpla algún criterio y el numero 0 cuando la herramienta no cumpla algún criterio, con el fin de determinar las herramientas más útiles para desarrollar el estudio.

Criterios de inclusión y exclusión de las referencias.

- Disponibles gratuitamente.
- Actualización periódica.
- Disponibles para el sistema operativo LINUX.
- Guía de instalación y uso.
- Cuenta con licencia para uso académico o de investigación.
- Cuenta con soporte para resolver problemas no reportados en el repositorio.
- Utiliza enfoque de tipificación basado en WGS.
- Reporte de uso en los últimos 5 años.
- Método de tipificación utilizado.

Etapa 2. Selección de las herramientas.

Una vez aplicado los criterios de inclusión y exclusión, se seleccionarán las herramientas bioinformáticas que cumplan mínimo el 90% de los requisitos técnicos y funcionales anteriormente nombrados para su ejecución.

4.1.2 Fase 2. Instalar e implementar las herramientas bioinformáticas.

La implementación de las herramientas bioinformáticas seleccionadas se realizará siguiendo los instructivos de instalación y ejecución de los programas operativos con los cuales funcionan las herramientas. En caso tal que la herramienta presente errores en su instalación o no funcione adecuadamente en el servidor local será excluida de la lista de seleccionadas.

4.1.3 Fase 3. Medir el desempeño de las herramientas bioinformáticas.

Esta fase metodológica se realizará en dos etapas, en donde se evaluará y se comparará los resultados de las herramientas seleccionadas contra los resultados de una herramienta bioinformática de referencia como Snippy, con el fin de realizar una clasificación de acuerdo con las métricas establecidas de evaluación, mediante métodos cuantitativos y cualitativos de agrupamiento que ayudan a identificar sus fortalezas (Weber et al., 2019). Para ello, se utilizara un conjunto de secuencias de referencia que contengan genomas completos y *reads* de bacterias sujetos de este estudio disponibles en la base de datos de

genomes del NCBI disponible en http://www.ncbi.nlm.nih.gov/genbank/ (Benson et al., 2017).

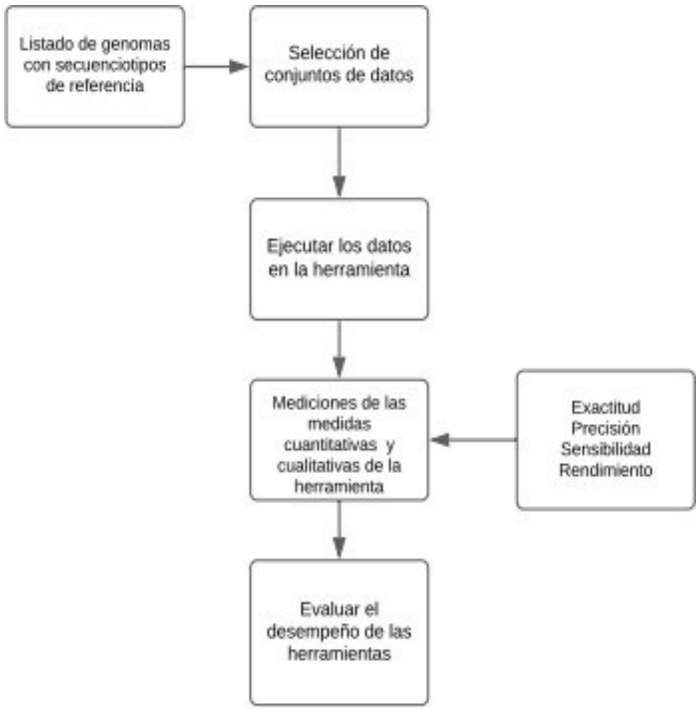

Figura 4 Flujograma de evaluación de desempeño de las herramientas bioinformáticas.

Etapa 1. Selección de conjuntos de datos.

De acuerdo con la figura 4, se seleccionaron un conjunto de datos de prueba con ensamble descrito como completo en formato FASTA de la base de datos *genomes* del NCBI, del cual también se conozca los secuenciotipos de las bacterias sujeto de este estudio, las cuales serán almacenadas en un servidor local de la Universidad Nacional de Colombia para su procesamiento. Esta selección se realizó siguiendo el siguiente paso:

- Listado de genomas completos con secuenciotipos de referencia conocidos para *Klebsiella pneumoniae* y *Pseudomonas aeruginosa,* disponibles en bases de datos de aislados PubMLST, MLST.org y MLSTPausteur entre otras (Anexo 1)

- Listado de genomas completos con secuenciotipos de especies cercanos y lejanos evolutivamente a *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* (Anexo 1)

Etapa 2. Comparar las herramientas

En esta etapa, se contrasta las herramientas considerando las medidas cuantitativas (métricas) y cualitativas del grado en que la herramienta posee atributos dado a la

exactitud, precisión, sensibilidad y tasa de error. Cabe mencionar que en la literatura se evidencia varias métricas o medidas para hacer evaluación (Weber et al., 2019).

Sin embargo, solo se tendrán en cuenta las medidas anteriormente nombradas de desempeño para este estudio.

Evaluación de la exactitud de las herramientas.

Para evaluar la Exactitud (*Accuracy*), se mide el porcentaje de casos que la herramienta ha acertado, con el fin de comprobar qué tan exactas son las mediciones en todo el rango esperado (Bogaerts et al., 2019).

Fórmula de exactitud

$$\frac{TP + TN}{TN + FN + TP + FP} \times 100 \qquad (4.1)$$

- Verdaderos positivos (TP): La herramienta predice que la muestra es positiva y, en realidad, sí que lo es (Lindgreen et al., 2016)
- Falsos positivos (FP): La herramienta predice que la muestra es positiva, pero, en realidad, no lo es (Lindgreen et al., 2016)
- Verdaderos negativos (TN): La herramienta predice que la muestra es negativa y, en realidad, sí que lo es (Lindgreen et al., 2016).
- Falsos negativos (FN): La herramienta predice que la muestra es negativa, pero, en realidad, no lo es (Lindgreen et al., 2016).

Evaluación de la precisión de las herramientas.

Para valorar la precisión, se utilizó un estudio de reproducibilidad con el fin de medir las variaciones cuando se usa la misma secuencia en las herramientas muchas veces bajo las mismas condiciones, y variaciones cuando se utilizan diferentes secuencias en las herramientas muchas veces bajo las mismas condiciones (Bogaerts et al., 2019).

Fórmula de precisión

$$\frac{TP}{TP + FP} \times 100 \qquad (4.2)$$

Evaluar la Sensibilidad de las herramientas.

Para medir la Sensibilidad (*Recall*) de la herramienta, se utilizó los *sets* de datos de secuencias anteriormente nombrada, con el fin de identificar la probabilidad de que la herramienta tipifique correctamente las secuencias de referencias. (Bogaerts et al., 2019).

Fórmulas de Sensibilidad

$$\frac{TP}{TP + FN} \times 100 \qquad (4.3)$$

Evaluar el rendimiento de las herramientas.

Se evaluó otras medidas para cada herramienta seleccionada como facilidad de uso, cantidad de memoria utilizada y velocidad de ejecución entre otros parámetros.

- **Facilidad de uso:** Hace referencia a la capacidad de la herramienta bioinformática para ser entendida y asimilada con facilidad bajo las condiciones especificadas. Su evaluación, se realizará a través de cinco criterios los cuales se le asignará una puntuación definida de la siguiente forma:

  0 = Muy difícil

  1 = Difícil

  2 = Normal

  3 = Fácil

  4= Muy fácil
- **Cantidad de memoria utilizada:** Se medió la cantidad de recursos que utiliza cada herramienta al ser ejecutada en la unidad de almacenamiento mediante la unidad de medida *Megabyte* (MB) o *Gigabyte* (GB) .
- **Velocidad de ejecución**: Se medió el tiempo en horas o minutos que tarda la ejecución de la herramienta al procesar los datos de entrada y generar un resultado mediante pruebas de rendimiento de cargas con el mismo número de secuencias. Con el fin de conocer las herramientas que requieren más tiempo.

5. Capítulo 5

5.1 Resultados y Discusión.

5.1.1 Selección de herramientas bioinformáticas útiles para tipificar.

De acuerdo con el modelo para identificación de las herramientas bioinformáticas descritas en el estado del arte, se realiza una evaluación de acuerdo con los criterios de selección anteriormente nombrados a las 19 herramientas bioinformáticas útiles para tipificar, dividida en tres enfoques; I) Nueve herramientas que utilizan enfoques multilocus; II) Seis herramientas que utilizan enfoques SNP; III) Cuatro herramientas que utilizan enfoques basados en pangenoma.

Selección de herramientas bioinformáticas que utilizan enfoques multilocus.

Al aplicar los criterios de selección, se evidencia que las herramientas chewBBACA y MentaLiST cumplen con el 100% de todos los requerimientos, seguida de SRST2, Seqsphere, stringMLST, BIGSdb las cuales cumplen solamente con el 87.5 de los requerimientos, goeBURST con un 75 %, canowgMLST BacComare con un 62.5% y lstdbNet 37.5% (Tabla2). Dentro de los criterios que menos cumplen las herramientas se encuentran la no disponibilidad para el sistema operativo LINUX, enfoque de tipificación diferente al WGS.

Cabe mencionar que la herramienta BIGSdb, cuenta con guía de instalación en un servidor local y uso en sistema operativo LINUX. Sin embargo, no cuenta con guía de uso para línea de comando específica para tipificación. Esto dado a que es un software que al descargarlo crea bases de datos PostgreSQL, se conectan al servidor de BIGSdb Web por medio de un *proxy* de API, lo que conlleva realizar el trabajo en la Web y no en el servidor local.

Tabla 2: Evaluación de herramientas bioinformáticas que utilizan enfoques basados multilocus.

Herramienta	Disponibles gratuitamente.	Actualización periódica.	Disponibles para el sistema operativo LINUX	Guía de instalación y uso.	Cuenta con licencia para uso académico o de investigación	Cuenta con soporte para resolver problemas	Utiliza enfoque de tipificación basado en WGS	Reporte de uso en los últimos 5 años.	Enfoque de tipificación utilizado	Total
lstdbNet	1	0	1	0	1	0	0	0	MLST	3
goeBURST	1	1	0	1	1	1	0	1	rMLST MLST	6
BIGSdb	1	1	1	0	1	1	1	1	MSLT rMLST cgMLST	7
SRST2	1	1	1	1	0	1	1	1	MLST	7
Seqsphere	0	1	1	1	1	1	1	1	MSLT rMLST cgMLST	7
stringMLST	1	1	1	1	1	1	0	1	MLST rMLST cgMLST	6
canowgMLST_BacComare	1	1	0	0	0	1	1	1	wgMLST	5
chewBBACA:	1	1	1	1	1	1	1	1	wgMLST cgMLST	8
MentaLiST	1	1	1	1	1	1	1	1	MLST wgMLST cgMLST	8

De acuerdo con los resultados obtenidos (Tabla 2), se escogen a chewBBACA y MentaLiST, dos herramientas bioinformáticas que esquemas cgMLST y wgMLST, y cumplieron con mínimo el 90% de criterios establecidos.

chewBBACA requiere de genomas ensamblados para el proceso, para evitar sesgos por genoma fragmentados, marco de lectura desplazado, longitudes de genomas muy grandes o pequeñas entre otros (A. Gupta et al., 2017). Por su parte, MentaLiST se basa en *k-mers,* lo que le permite utilizar lecturas cortas de genomas sin ensamblar (A. Gupta et al., 2017).

Dependiendo del funcionamiento de cada herramienta, los resultados finales y las conclusiones de un análisis de multilocus de genoma central, dependen también de la base de datos de nomenclatura de referencia utilizada (BIGSdb-Pasteur, PubMLST y cgMLST.org). Las cuales tienen diferentes esquemas de tipificación obtenidos a partir de conjuntos de secuencias de genes de referencias.

Selección de herramientas bioinformáticas que utilizan enfoques basados en SNP.

Para la evaluación de las seis herramientas bioinformáticas identificadas que emplearon enfoque basados en SNP, se utilizaron los mismos criterios mencionados previamente. Se encontró que las herramientas SAMtools, Parsnp, GATK4 y kSNP3 cumplen con el 100% de todos los requerimientos, seguida de BactSNP la cual cumple solamente con el 87.5 de los requerimientos y Harvest con un 75 % (Tabla 3). Dentro de los criterios que menos cumplen las herramientas basadas en SNP se encuentran el soporte para resolver problemas.

Tabla 3 Evaluación de las herramientas bioinformáticas que utilizan enfoque SNP.

Herramienta	Disponibles gratuitamente.	Actualización periódica.	Disponibles para el sistema operativo LINUX	Guía de instalación y uso.	Cuenta con licencia para uso académico	Cuenta con soporte para resolver problemas	Utiliza enfoque de tipificación basado en WGS	Reporte de uso en los últimos 5 años.	Enfoque de tipificación utilizado	Total
SAMtools	1	1	1	1	1	1	1	1	SNP	8
GATK4	1	1	1	1	1	1	1	1	SNP	8
Parsnp	1	1	1	1	1	1	1	1	SNP	8
BactSNP	1	1	1	1	1	0	1	1	SNP	7
kSNP3	1	1	1	1	1	1	1	1	SNP	8
Harvest	1	0	0	1	1	1	1	1	SNP	6

De las herramientas bioinformáticas que utilizan enfoque de tipificación basados en nucleótidos con esquemas cgSNP o wgSNP, se escogen cuatro herramientas las cuales cumplieron con mínimo el 90% de criterios establecidos (Tabla 4).

Tabla 4 Herramientas seleccionadas que utilizan enfoque SNP.

Herramientas	Enfoque de tipificación utilizado	Puntuación	Porcentaje de aceptación
SAMtools	SNP	8	100%
GATK4	SNP	8	100%
Parsnp	SNP	8	100%
kSNP3	SNP	8	100%

El desarrollo continuo de herramientas bioinformáticas útiles para hacer llamada de SNP basados en WGS, también podrían afectar el análisis para identificar polimorfismos de un solo nucleótido entre los aislamientos, esto dado a que cada herramienta proporciona información diferente al hacer inferencia filogenética u otros análisis.

Las herramientas SAMtools, GATK, Parsnp, son herramientas basadas en ensamblaje y mapeo por referencia para evitar falsos positivos (Yoshimura et al., 2019). Mientras que la herramienta kSNP3 utiliza *K-mers* para identificación de SNP y el análisis filogenético, ya que no utiliza alineación del genoma y genomas de referencia, lo que podría favorecer a la aparición de falsos positivos(Gardner et al., 2015).

Selección de herramientas bioinformáticas que utilizan enfoques basados en el pangenoma.

Para evaluar las suites que emplean enfoque basados en Pangenoma, se utilizó la revisión del mismo modo que enfoques multilocus de genoma central y SNP. Se evidenció que las herramientas Roary y PIRATE cumplen con el 100% de todos los requerimientos, seguida de PanOCT y PanACEA las cuales cumplen solamente con el 75% de los requerimientos (Tabla 5). Entre los criterios que menos cumplen estas herramientas se encuentran la no Actualización periódica y soporte para resolver problemas.

Tabla 5 Evaluación de herramientas bioinformáticas que utilizan enfoques pangenoma.

Herramienta	Disponibles gratuitamente.	Actualización periódica.	Disponibles para el sistema operativo LINUX	Guía de instalación y uso.	Cuenta con licencia para uso académico	Cuenta con soporte para resolver problemas	Utiliza enfoque de tipificación basado en WGS	Reporte de uso en los últimos 5 años.	Enfoque de tipificación utilizado	Total
Roary	1	1	1	1	1	1	1	1	Pangenoma	8
PanOCT	1	0	1	1	1	0	1	1	Pangenoma	6
PanACEA	1	0	1	1	1	0	1	1	Pangenoma	6
PIRATE	1	1	1	1	1	1	1	1	Pangenoma	8

Para este enfoque se seleccionan preliminarmente las herramientas Roary y PanACoTA las cuales cumplieron con el 100% de la serie de criterios establecidos en comparación con las otras herramientas bioinformáticas (Tabla 6).

Tabla 6 Selección de herramientas útiles para realizar pangenoma.

Herramientas	Enfoque de tipificación utilizado	Puntuación	Porcentaje de aceptación
Roary	Pangenoma	8	100%
PIRATE	Pangenoma	8	100%

Si bien, la aplicación del pangenoma no contempla secuenciotipos definidos para caracterizar bacterias. Mediante la clasificación de ausencia y presencia de genes, se puede agrupar cepas de acuerdo a homologías de vecindad de genes y de esta manera, asignar características a diferentes secuencias.(Page et al., 2015)(Perrin & Rocha, 2021)(Bayliss et al., 2019).

5.1.2 Instalación de las herramientas bioinformáticas seleccionadas.

Todas las herramientas bioinformáticas seleccionadas, se instalaron y ejecutaron de acuerdo con sus guías de uso en un servidor de marca Dell, referencia PowerEdge T640 con sistema operativo OpenSuse v.15.2 LINUX, perteneciente a la Universidad Nacional de Colombia. El tiempo de instalación aproximado fue de 2 minutos por herramienta excepto MentaLiST y kSNK las cuales requieren instalación manual. El 91 % de los códigos fuente de las herramientas se encuentran almacenados y publicados en GitHub. Por otro lado, el 75% de las herramientas se pueden instalar a través de gestores de paquetes como Anaconda o Pip y solo un 25% se recomienda instalar manualmente (Tabla 7).

Dado a que todas las funciones de las herramientas desarrolladas se implementan en lenguajes como Perl, julia, java y Python respectivamente, se deben instalar algunos paquetes y módulos complementarios para la adecuada ejecución de las herramientas, las cuales contienen diferentes archivos comprimidos, código fuente y paquetes binarios complementarios entre otros.

Tabla 7 Instalación de las herramientas seleccionadas.

Herramientas	Paquetes para instalación	Gestores de paquetes usado	Instalabilidad	Instalado correctamente
chewBBACA	https://github.com/gregorylburgess/Chewbacca/blob/master/install.sh	Anaconda	Fácil	Si
MentaLiST	https://github.com/WGS-TB/MentaLiST	Manual	No fácil	Si
SAMtools	https://github.com/samtools/samtools	Anaconda	Fácil	Si
GATK4	https://github.com/broadinstitute/gatk	Anaconda	Fácil	Si
Parsnp	https://github.com/marbl/parsnp	Anaconda	Fácil	Si
kSNP3	https://github.com/cdeanj/kSNP3	Manual	No fácil	Si
Roary	https://github.com/sanger-pathogens/Roary	Anaconda	Fácil	Si
PIRATE	https://github.com/SionBayliss/PIRATE	Anaconda	Fácil	Si

La percepción de instalabilidad, se consideró fácil y sencilla de interpretar para aquellas herramientas que utilizan gestores de paquetes, esto dado a que no se requiere un conocimiento avanzado en programación en sistemas operativos LINUX. Adicionalmente, anaconda contempla aplicaciones, librerías y conceptos diseñados para eliminar posibles problemas al momento de instalar dependencias de paquetes y control de versiones.

Sin embargo, aquellas herramientas que no se ejecutaron a través de gestores de paquetes, son propensas a presentar problemas en el proceso de la instalación de dependencias, esto dado a que se requiere tener un poco más de conocimiento en programación para poder configurarlas.

5.1.3 Evaluar el desempeño de las herramientas bioinformáticas.

5.1.3.1 Selección del conjunto de datos.

La construcción del *dataset* para la especie *Pseudomonas aeruginosa* se realizó utilizando secuencias de genomas completos en archivos en formato FASTA obtenido de la base de datos *Genomes* del NCBI el 25 de enero de 2021. Se descargaron 53 secuencias de genomas en total, de las cuales 48 pertenecen a la especie *P.aeruginosa,* tomando como referencia la secuencia del genoma de la cepa PAO1(NC_002516). 2 secuencias de *Salmonella entérica* y 3 secuencias de *Klebsiella pneumoniae* evolutivamente lejanas con el fin de verificar si las herramientas identifican falsos positivos. Adicionalmente, se determinaron los secuenciotipos MLST, rMLST y cgMLST de los sitios web PubMLST (https://pubmlst.org/) y SeqSphere (http://www.cgmlst.org/) (Anexo 1) .

De igual manera, se construye el *dataset para la especie K. pneumoniae, en donde* se descargaron 47 secuencias de genomas en archivos en formato FASTA de la base de datos *Genomes* del NCBI el 26 de enero de 2021, de las cuales 41 pertenecen a *Klebsiella*

pneumoniae tomando como referencia la secuencia del genoma KP-1(NZ_CP012883.1), 2 a *Pseudomonas aeruginosa,* una a *Burkholderia pseudomallei, 2* a *Listeria monocytogenes* evolutivamente lejanas y una a *Klebsiella oxytoca* evolutivamente cercana. Adicionalmente, se determina los secuenciotipos MLST y rMLST del sitio web PubMLST (https://pubmlst.org), cgMLST del sitio web SeqSphere (http://www.cgmlst.org) y cgMLST del sitio web BIGSdb-Pasteur (https://bigsdb.pasteur.fr) (Anexo 1).

5.1.3.2 Ejecución de las herramientas.

La comparación del desempeño de las herramientas bioinformáticas seleccionadas, se realiza cotejando el árbol filogenético obtenido utilizando Snippy (Figura 5) (Figura 6) (Bush, 2021), elegido como *gold standard.* Esto dado a que es una herramienta ampliamente utilizada en estudios genómicos para identificación de SNP, y se ha demostrado que es preciso y minimiza las llamadas de falsos positivos (Coll et al., 2022)(Brilhante et al., 2021). Adicionalmente, se evidencia su uso en algunos estudios de evaluaciones comparativas de herramientas bioinformáticas (Bush et al., 2020)(Labbé et al., 2021).

El árbol filogenético de máxima verosimilitud (ML) estándar se enraizó con la herramienta iTOL, la cual permite hacer una visualización de forma gratuita (Letunic & Bork, 2019). Claramente se muestran dos *clústers,* uno que agrupa todas las especies diferentes a *P. aeruginosa* y el otro agrupando los genomas de *P. aeruginosa* (Figura 5). En el árbol se puede identificar seis *clústeres* conformados de la siguiente manera; I) *Clúster* número uno lo conforma los ST19 de la especie *S.enterica* y ST65, ST15, ST66 de la especie *K.pneumoniae*; II) El siguiente *clúster* lo conforma tres cepas del ST253 y una cepa de ST234, ST237 de la especie *P.aeruginosa*; III) El clúster número tres, está conformado por cuatro cepas del ST782 de la especie *P.aeruginosa*; IV) El *clúster* número cuatro, está conformado por tres cepas ST549 de la especie *P.aeruginosa*; V) El *clúster* número cinco, está conformado por una cepa del ST82 y una cepa del ST198 de la especie *P.aeruginosa*; VI) El *clúster* número seis, está conformado por cinco cepas del ST146 y una cepa de del ST308 de la especie *P.aeruginosa.*

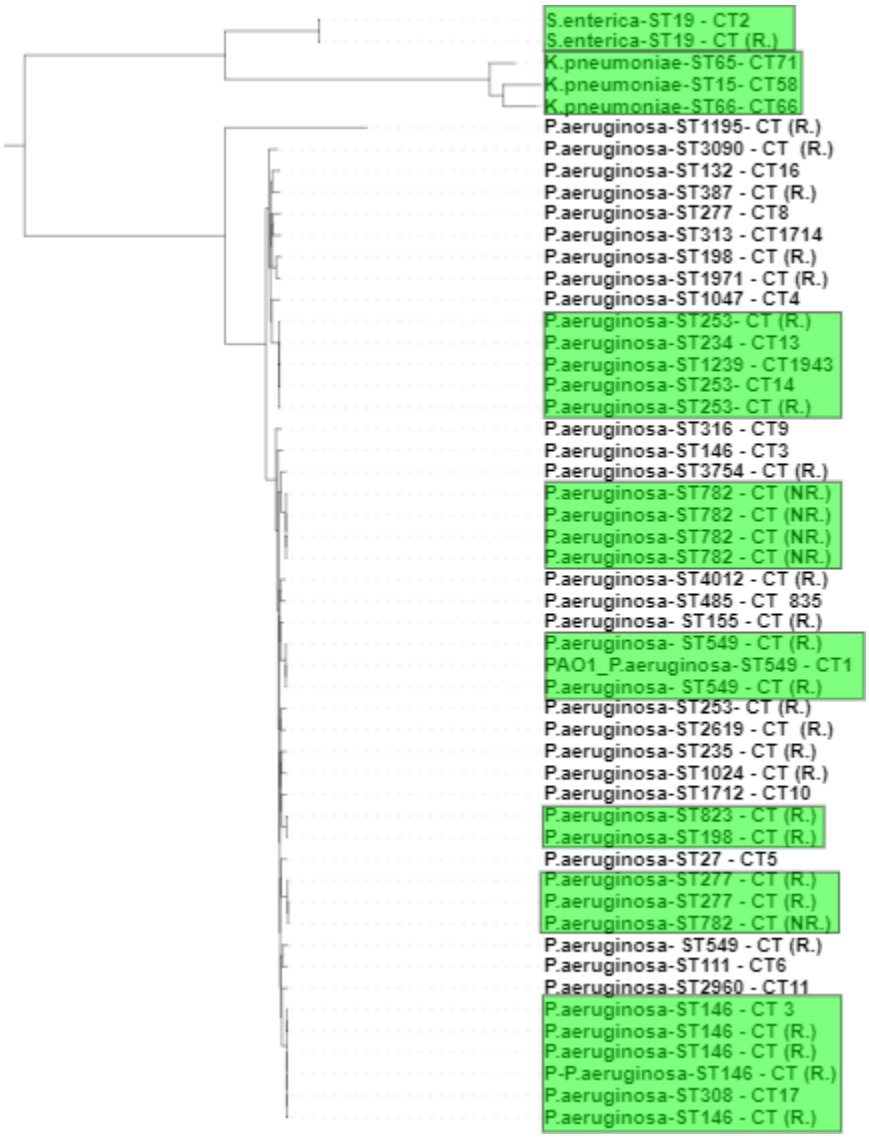

Figura 5 Árbol de máxima verosimilitud (ML) con escala 01. *Pseudomonas aeruginosa* Snippy.

Al igual que con *Pseudomonas aeruginosa*, el árbol filogenético de máxima verosimilitud (ML para la especie *K. pneumoniae* se enraizó con la herramienta iTOL (Figura 6). El árbol incluye siete *clústeres* conformados de la siguiente manera; I) El *clúster* número uno lo conforma dos ST2 de la especie *L.monocytogens* y dos ST65, ST27 de la especie *P.aeruginosa*; II) El siguiente *clúster* lo conforma ocho cepas del ST15 de la especie *K.pneumoniae*; III) El *clúster* número tres, está conformado por tres cepas del ST340 y dos cepa del ST11 de la especie *K.pneumoniae*; IV) El *clúster* número cuatro, está conformado por cuatro cepas ST11 de la especie *K.pneumoniae*; V) El *clúste*r número cinco, está conformado por cuatro cepa del ST86 de la especie *K.pneumoniae*; VI) El *clúster* número seis, está conformado por cinco cepas del ST 29 y una cepa del ST277 y una cepa de ST 782 de la especie *K.pneumoniae.*; VII) El clúster número seis, está conformado por cinco cepas del ST 29 y una cepa del ST17 de la especie *K.pneumoniae.*

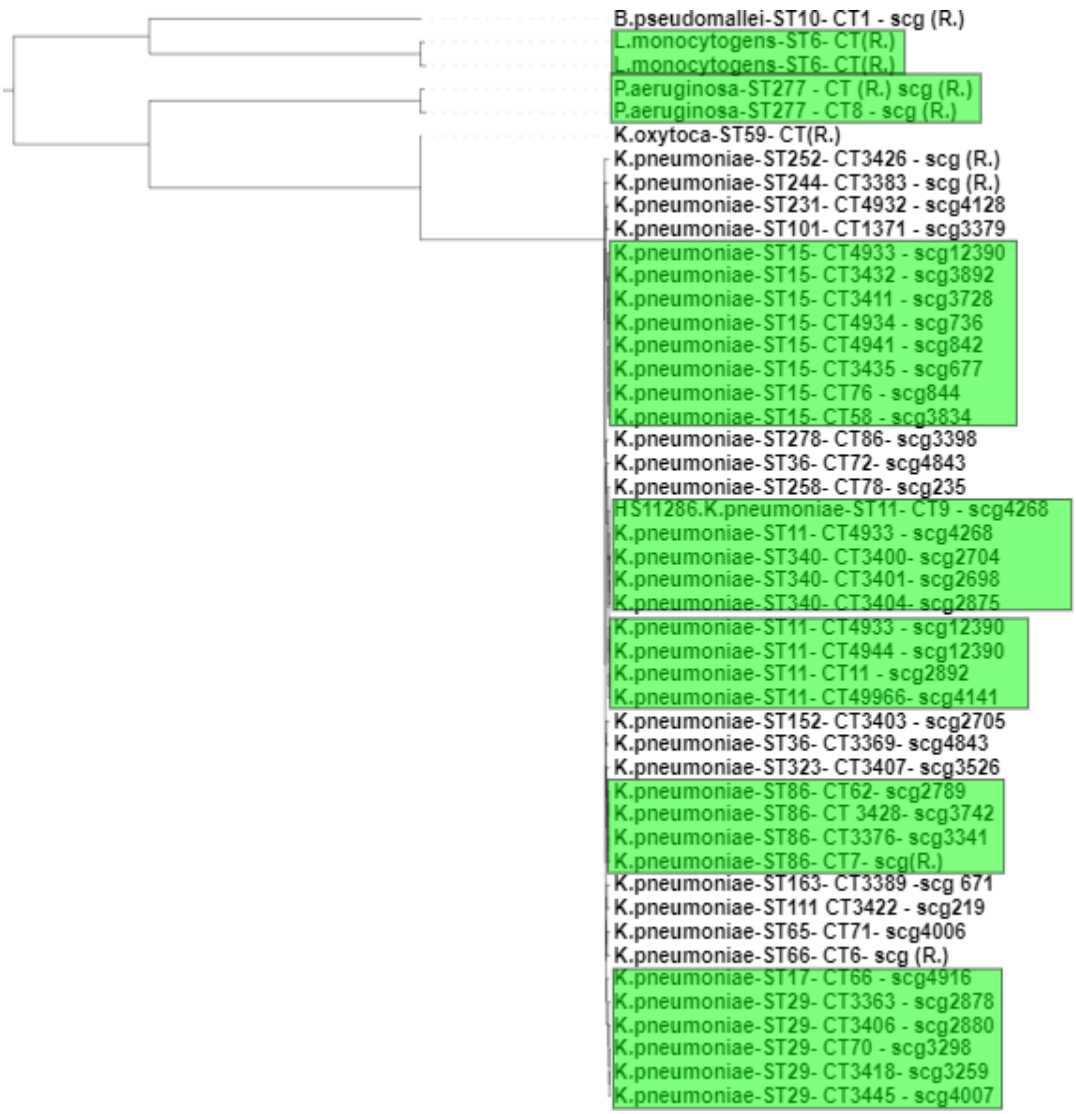

Figura 6 Árbol de máxima verosimilitud (ML) con escala 0.1 *Klebsiella Pneumoniae* Snippy.

5.1.3.2.1 Ejecución de herramientas bioinformáticas que utilizan enfoques multilocus.

Herramienta chewBBACA .

Este flujo de trabajo inicia con la adopción de diferentes esquemas de tipificación (Tabla 8) descargados de las bases de datos de nomenclatura alélica publica, los cuales tienen diferentes conjuntos de loci de las especies sujetas de este estudio disponibles en el sitio web del servidor de nomenclatura cgMLST.org http://www.cgmlst.org/, BIGSdb-Pasteur https://bigsdb.pasteur.fr/ y ChewBBACA https://github.com/B-UMMI/chewBBACA.

Tabla 8 Esquemas para tipificación seleccionadas.

Esquema	Especie	Genes adoptados	Método de tipificación	Genoma de referencia
chewBBACA	*P. aeruginosa*	13586	cgMLST	PAO1(NC_002516)
SeqSphere	*P. aeruginosa*	3867	cgMLST	PAO1(NC_002516)
SeqSphere	*K. pneumoniae*	2358	cgMLST	KP1(NZ_CP012883.1)
BIGSdb-Pasteur	*K. pneumoniae*	626	cgMLST	KP1(NZ_CP012883.1)

Se utilizó el algoritmo de entrenamiento Prodigal para predecir rápidamente los genes presentes en la secuencia FASTA de referencia para *P. aeruginosa* PAO1(NC_002516) y *K. pneumoniae* HS11286 (NZ_CP012883.1) con un tiempo de ejecución de 20 segundos cada uno. Cabe mencionar que esta actividad se debe realizar antes de adoptar o crear un esquema de tipificación (cgMLST) ya que permite garantizar la reproducibilidad de la llamada alélica.

Para la adopción de esquemas externos (wgMLST/cgMLST) se utilizó el algoritmo *PrepExternalSchema* de la herramienta chewBACCA. Se adoptaron esquemas de SeqSphere para la especie *P. aeruginosa*, que usa un total de 3867 genes, de los cuales se invalidaron 58 alelos y cuyo tiempo de ejecución fue 8 minutos 30 segundos, usando 4 *cores* y el esquema chewBACCA 13586 genes, esquema del que se invalidaron 70 alelos procesados en 13 minutos 44 segundos, utilizando 4 *cores*. Esta información está disponible en la carpeta *P.aeruginosa*_Chewbbaca y *P.aeruginosa*_SeqSphere, archivos *invalid_alleles.txt*, *genes_novalidos.txt* y *summary_stats.tsv* (Anexo 5).

De igual manera, se adoptan los esquemas cgMLST de SeqSphere y BIGSdb-Pasteur para la especie *K. pneumoniae*. Con un total de 2358 genes para el esquema SeqSphere, con 20840 alelos inválidos con un tiempo de ejecución de 6 minutos 51 segundos, utilizando 4 *core,* mientras que para el esquema BIGSdb-Pasteur, fueron 626 genes, 8 alelos inválidos con un tiempo de ejecución de 1 minuto 43 segundos, usando 4 *cores*. Esta información, se puede consultar en la carpeta KPC_Ridom y KPC_pasteur, archivos *invalid_alleles.txt*, *genes_novalidos.txt* y *summary_stats.tsv (*Anexo 5).

Cabe mencionar que a medida que se aumenta el número de *cores*, aumenta el porcentaje de velocidad de procesamiento. Por consiguiente, se obtendrá un procesamiento de los datos con menor tiempo de ejecución al que se reporta en el presente trabajo. Adicionalmente, los esquemas adoptados y usados en la herramienta chewBACCA https://github.com/B-UMMI/chewBBACA, se ejecutaron con parámetros predeterminados con una medida de similitud basada en la proporción de puntuaciones BLAST (BLAST Score Ratio de 0.6), el cual permite aceptar secuencias de genes de cualquier longitud. De igual manera, debido a la variabilidad en el tamaño de los alelos identificados por la herramienta se usa un umbral de +/-20 %.

En el proceso de llamada de alelo, se utiliza el algoritmo *AlleleCall en* donde *se* definieron los esquemas wgMLST creados a partir de genes adoptados de chewBACCA y BIGSdb-Pasteur para los 52 genomas completos del *dataset* para la especie *Pseudomonas aeruginosa* y 47 genomas completos del *dataset* para la especie *Klebsiella pneumoniae* con el fin de establecer los perfiles alélicos conocidos, nuevos y loci no encontrado de cada cepa.

Adicionalmente, para establecer los perfiles alélicos de las secuencias usadas en el algoritmo *AlleleCall*, se proporcionó el archivo de entrenamiento de Prodigal para cada una de las especies *P. aeruginosa* PAO1 y *K. pneumoniae* KP1. Se crea una base de datos usando BLASTp con los genes identificados por Prodigal de los genomas de referencia, en donde se comparan los locus presentes con una identidad del 100 % entre todos los genomas con cada una de las bases de datos de alelos de locus identificado con la abreviatura EXC. Para aquellos locus que no coinciden exactamente, se utiliza la relación de puntuación BLAST (BSR) con un valor predeterminado por la herramienta de 0.6 con el fin de identificar los locus que tienen una identidad entre el 70% y 80% con la abreviatura INF.

Asimismo, la herramienta chewBBACA utiliza una nomenclatura para identificar los locus no encontrados con la abreviatura LNF. Para los locus parálogos y coincidencias exactas en el alelo, se identifica con abreviatura NIPH y NIPHEM, los cuales se deben eliminar del análisis por generar múltiples copias o duplicación de genes que indican secuencias mal ensambladas o de mala calidad. Cabe mencionar que hasta que no se eliminen estas secuencias, el algoritmo *AlleleCall* no permite realizar el llamado de alelo. Esta información se puede descargar del archivo *results_statistics.tsv* disponible en las carpetas *P.aeruginosa*_Chewbbaca, *P. aeruginosa* Ridom, KPC_Pasteur, KPC_Ridom (Anexo 5).

Tabla 9 Parámetros de llamada de alelo.

Esquema adoptado	Especie	Numero de loci usados	% de coincidencia exacta	Loci parálogos	Tiempo de ejecución
chewBACCA	*P. aeruginosa*	4698	88,03	63	32 minutos 20 segundos
SeqSphere	*P. aeruginosa*	3809	88,96	50	22 minutos 40 segundos
BIGSdb-Pasteur	*K. pneumoniae*	626	80.86	2	2 minutos 34 segundos
SeqSphere	*K. pneumoniae*	2358	83,78	12	12 minutos 47 segundos

El tiempo de ejecución de la llamada es proporcional al número de genomas utilizados (Tabla 11). Adicionalmente, se puede concluir que el número de loci usados por la herramienta solo está presente en el 88 % aproximadamente de los genomas de *P. aeruginosa* y el 81% aproximadamente de los genomas de K. *pneumoniae* (Tabla 11). Los alelos nuevos, se agregan al esquema aumentando el número de alelos en el esquema. La información se puede descargar del archivo *results_statistics.tsv* disponible en las carpetas *P.aeruginosa*_Chewbbaca, *P. aeruginosa*_Ridom, KPC_Pasteur, KPC_Ridom (Anexo 5).

Debido a que el algoritmo de llamada de alelo *AlleleCall* imprime una lista de loci parálogos los cuales generan posible fluctuación en la asignación de alelo, se debe realizar la remoción de estos utilizando el algoritmo *RemoveGenes* obteniendo un nuevo archivo de loci llamado *results_alleles_NoParalogs.tsv* el cual contiene todos los genes que se usaron wgMLST para definir el esquema cgMLST (Anexo 5).

Selección y evaluación del esquema cgMLST .

Con la nueva lista de loci registrados en el archivo *results_alleles_NoParalogs.tsv* se determinó el conjunto de loci en el genoma central (cgMLST) utilizando la operación *TestGenomeQuality* como prueba de calidad para determinar el umbral de presencia de genes al 100 %, 99,5 %, 99 % y el 95 % de los genomas con buena calidad analizados (Tabla 10). De igual manera, se excluyen los loci que están por debajo del 95% de los genomas en análisis.

Tabla 10 Resultados de TestGenomeQuality.

Esquema adoptado	Especie	Numero de genomas	Genes precedentes en el 100%	Genes precedentes en el 99,5%	Genes precedentes en el 99%	Genes precedentes en el 95%
chewBACCA	*P.aeruginosa*	52	145	145	145	318
SeqSphere	*P.aeruginosa*	52	152	152	152	335
BIGSdb-Pasteur	*K.pneumoniae*	47	3	3	3	32
SeqSphere	*K.pneumoniae*	47	14	14	14	120

El archivo de salida se puede encontrar en la carpeta *TestGenomeQuality* /*GenomeQualityPlot*.html con los loci que componen el cgMLST (95%). Adicionalmente, se extrae el conjunto de loci en el genoma central al 95 % ejecutando el módulo *ExtraCgMLST* el cual dura aproximadamente 15 segundos y cuyo archivo de salida contiene una matriz de presencia y ausencia de alelos (Archivo *Presence_Abscence.tsv*) y el listado loci del esquema cgMLST (Archivo *cgMLSTschema.tx*). Por último, la matriz de perfiles alélicos para cgMLST(Archivo *cgMLST.tsv*).

De acuerdo con la matriz de perfiles alélicos (*cgMLST.tsv)* de cada uno de los esquemas, se construyeron los árboles de unión de vecinos (NJ) y expansión mínima (MST) usando el software en línea de comando GrapeTree (Z. Zhou et al., 2018) . Estos árboles se

pueden visualizar en la herramienta en línea iTOL (Letunic & Bork, 2019) y PHYLOViZ (Francisco et al., 2012). Esta información se puede consultar en la carpeta tree_Chewbbaca https://github.com/ocarabali/Anexos.git.

Representación gráfica de los resultados de cgMLST: Arboles de unión de vecinos (NJ) *P.aeruginosa.*

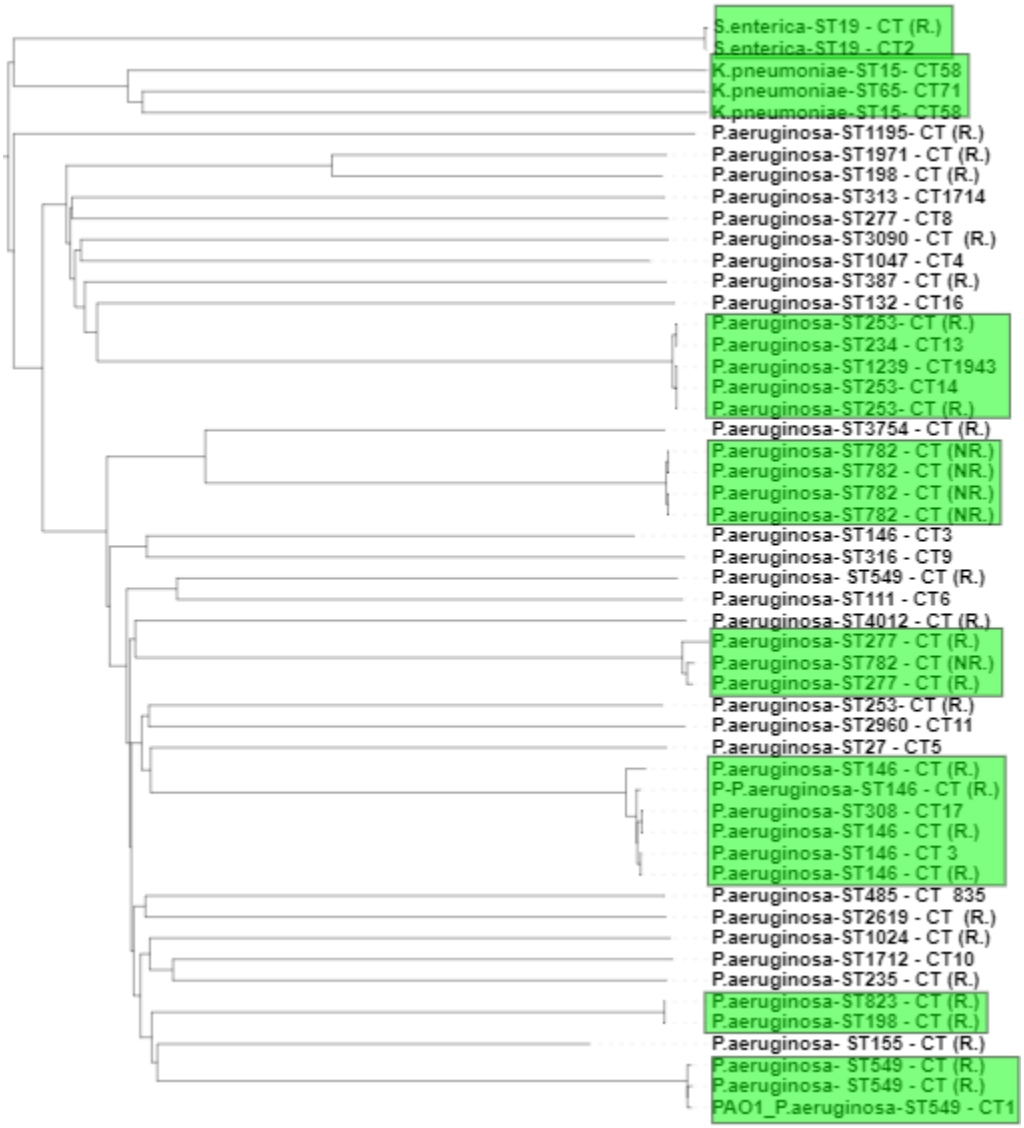

Figura 7 Árbol de unión de vecinos (NJ) cgMLST con escala 0.1 basado en perfiles del esquema *P. aeruginosa*_ SeqSphere.

.

El árbol de unión de vecinos (NJ) de perfiles alélicos de *P. aeruginosa* (Figura 7), agrupó 52 genomas con un tamaño de perfil de 355 loci. Al realizar la comparación con el árbol *P. aeruginosa* estándar (Figura 5), se evidencia un árbol con una topología diferente. Sin embargo, el esquema SeqSphere utilizado en la herramienta chewBACCA agrupó

correctamente las cepas estrechamente relacionadas con un máximo de 20 diferencias alélicas de distancia. De igual manera, los genomas que pertenecen a especies diferente se agruparon correctamente. Este esquema se ejecutó en un tiempo de 2 horas aproximadamente utilizando 4GB de memoria.

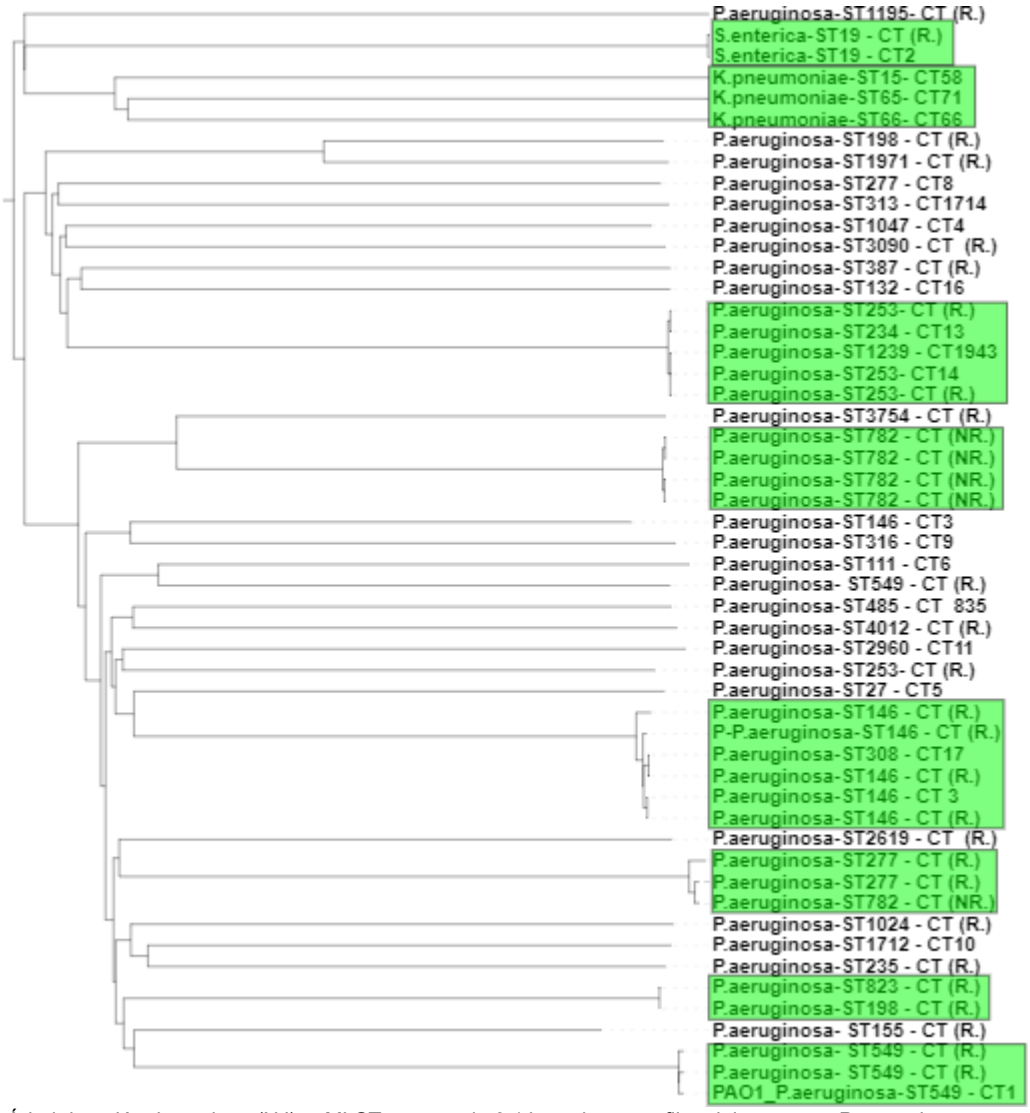

Figura 8 Árbol de unión de vecinos (NJ) cgMLST con escala 0.1 basado en perfiles del esquema *P. aeruginosa* _Chewbbaca

Al igual que con el esquema SeqSphere, el árbol de unión de vecinos (NJ) de perfiles alélicos de *P. aeruginosa* de esquema chewBACCA (Figura 8) agrupó 52 genomas con un tamaño de perfil de 319 loci. Al realizar la comparación con el árbol *P. aeruginosa* estándar(Figura 5), se evidencia un árbol con una topología diferente, pero similar al árbol de la figura 7 del esquema SeqSphere, en donde se agrupo correctamente las cepas estrechamente relacionadas con un máximo de 20 diferencias alélicas de distancia, al igual

que los genomas que pertenecen a especies diferentes. Este esquema se ejecutó en un tiempo de 2 horas 37 minutos aproximadamente utilizando 4GB de memoria.

Representación gráfica de los resultados de cgMLST: Arboles de unión de vecinos (NJ) *K.pneumoniae.*

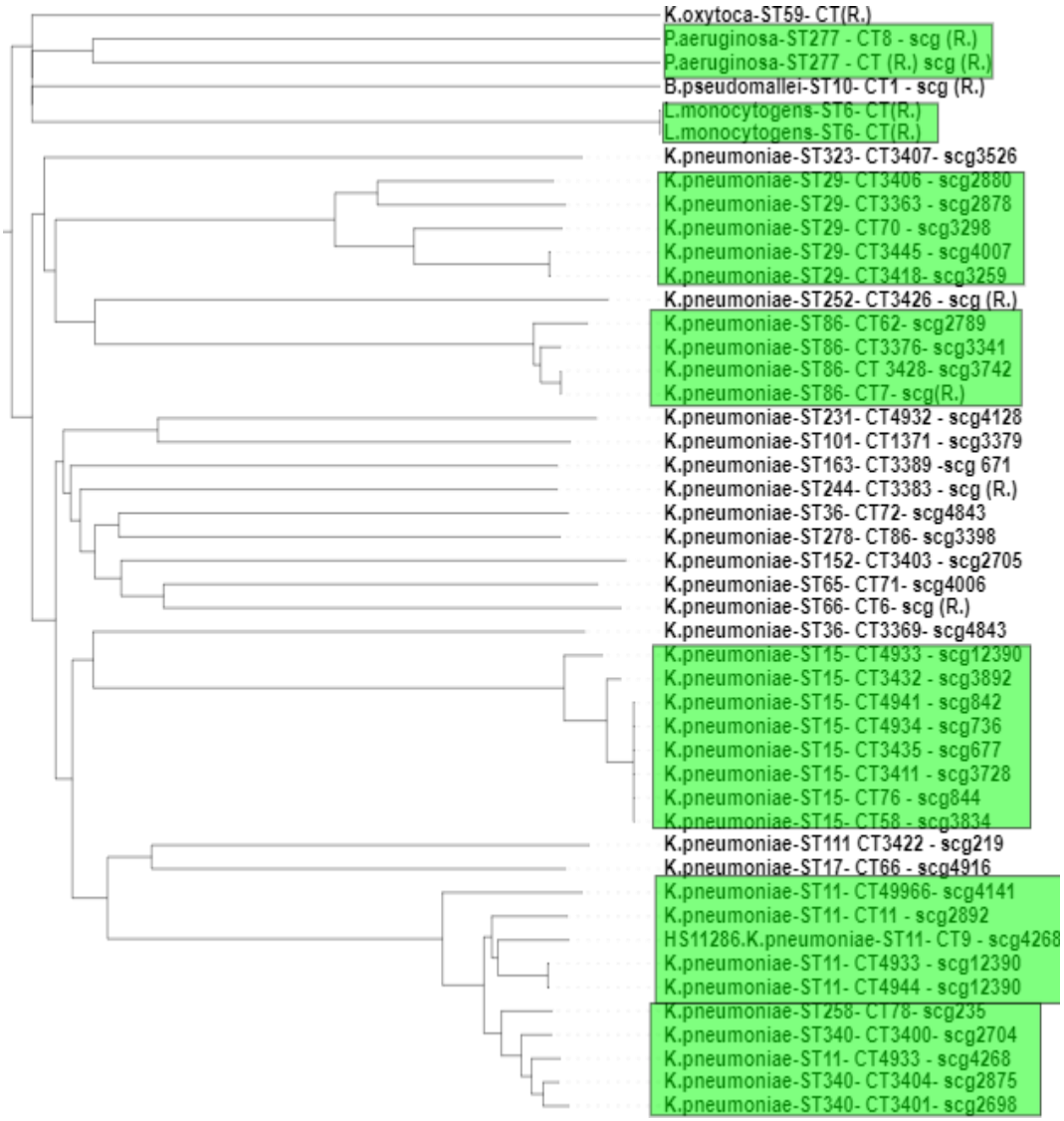

Figura 9 Árbol de unión de vecinos (NJ) cgMLST con escala 0.1 basado en perfiles del esquema *K. pneumoniae* _Pasteur.

El árbol de unión de vecinos (NJ) de perfiles alélicos de *K. pneumoniae* del esquema Pasteur (Figura 9) agrupó 47 genomas con un tamaño de perfil de 32 loci. Al realizar la comparación con el Árbol *K. pneumoniae* estándar (Figura 6), se evidencia un árbol con una topología diferente en donde se agrupo correctamente la mayoría de las cepas estrechamente relacionadas con un máximo de 20 diferencias alélicas de distancia. Excepto, ST258 el cual está presente en el grupo 3, ST11 CT9 presente en el grupo 4. Por

el cual se toman como falsos positivos y ST17 clasificándolo como falso negativo. Esto dado porqué la herramienta los agrupa con ST a los cuales no tienen cercanía. La herramienta agrupa adecuadamente las cepas que pertenecen a especies diferentes. Este esquema se ejecutó en 47 minutos aproximadamente utilizando 6GB de memoria.

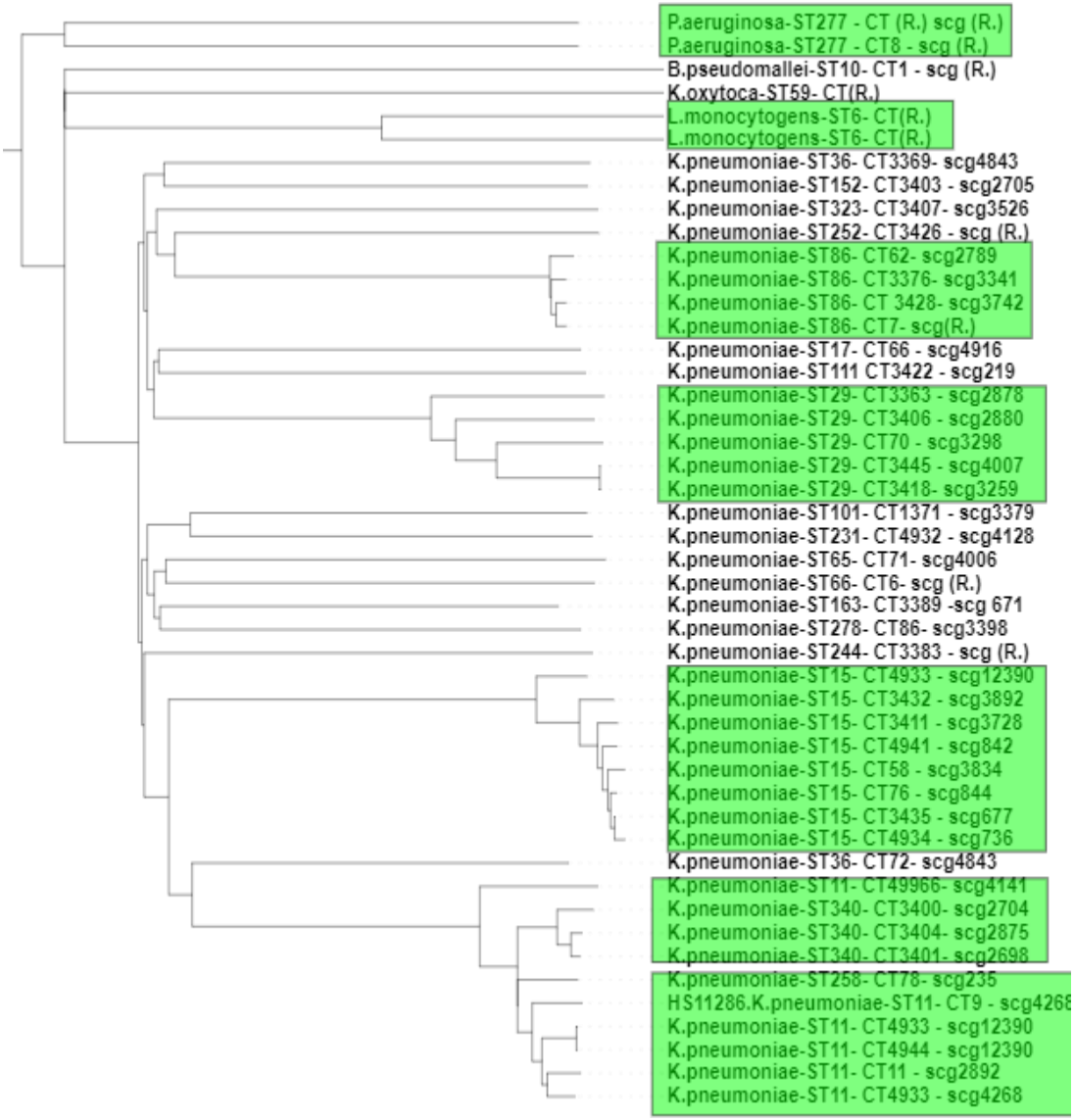

Figura 10 Árbol de unión de vecinos (NJ) cgMLST con escala 0.1 basado en perfiles esquema *K. pneumoniae* _ SeqSphere.

El árbol de unión de vecinos (NJ) de perfiles alélicos de *K. pneumoniae* del esquema SeqSphere (Figura 10), agrupó 47 genomas con un tamaño de perfil de 120 loci. Al realizar la comparación con el Árbol *K. pneumoniae* estándar (Figura 6), se evidencia un árbol con una topología diferente pero similar al del esquema Pasteur pese a utilizar menos genes que el esquema SeqSphere. en donde se agrupo correctamente la mayoría de las cepas

estrechamente relacionadas con un máximo de 20 diferencias alélicas de distancia. Excepto ST258 el cual está presente en el grupo 3, ST11 CT9 presente en el grupo 4. Por el cual se toman como falsos positivos y ST17 clasificándolo como falso negativo. Esto dado porqué la herramienta los agrupa con ST a los cuales no tienen cercanía. La herramienta agrupa adecuadamente las cepas que pertenecen a especies diferentes. Este esquema se ejecutó en 1:00 hora aproximadamente utilizando 6GB de memoria.

Herramienta MentaLiST.

Para la evaluación de MentaLiST, se utilizaron los *datset* anteriormente nombrados de las especies *P. aeruginosa* y *K. pneumoniae* con el fin de medir el desempeño, tiempo de ejecución y recursos computacionales.

El flujo de trabajo de *MentaLiST*, inicia con la adopción e instalación del esquema SeqSphere para la especie *K. pneumoniae* basado en 2361 genes y *P. aeruginosa* basado en 3869 genes disponible en https://www.cgmlst.org/ncs y continua con la construcción de una base de datos de *k-mers* de longitud predeterminada (k=31), en donde se calculan los *k-mers* de todos los alelos del locus y se crea una Tabla *hash* que vincula cada *k-mers* con los alelos donde está presente (Feijao et al., 2018). Este paso duro aproximadamente 3 minutos por cada *dataset.*

Posteriormente, se realiza el llamado de alelo por triplicado para cada esquema utilizando en la base de datos de *k-mers* anteriormente nombrada, obteniendo como resultado cuatro archivos por cada especie (*kpc_call.txt* y *P_call.txt*) en los cuales se relaciona la matriz de perfiles alélicos del enfoque wgMLST. (*kpc_call.txt.novel.fa* y *P_call.txt.novel.fa*) relacionan las secuencias de los alelos nuevos, (*kpc_call.txt, coverage.txt* y *P_call.txt.coverage.txt*) relacionan los alelos con mayor puntuación y cobertura del 100% o debajo del umbral del 50%. Por último (*kpc_call.txt.novel.txt* y *P_call.txt.novel.txt*), contiene la matriz en donde se relacionan los alelos nuevos, mutaciones y sus descripciones (Anexo 5).

En el proceso de llamado de alelo para las 2 especies, se utiliza el algoritmo MentaLiST *call* a partir de todos los genes adoptados para los genomas del *dataset* de la especie *Klebsiella pneumoniae y* de *P.aeruginosa.* La duración de este proceso fue de 55 minutos aproximadamente y uso de memoria RAM de 2 Gb por cada una de las réplicas.

Representación gráfica de los resultados de wgMLST: Arboles de unión de vecinos (NJ) *K.pneumoniae.*

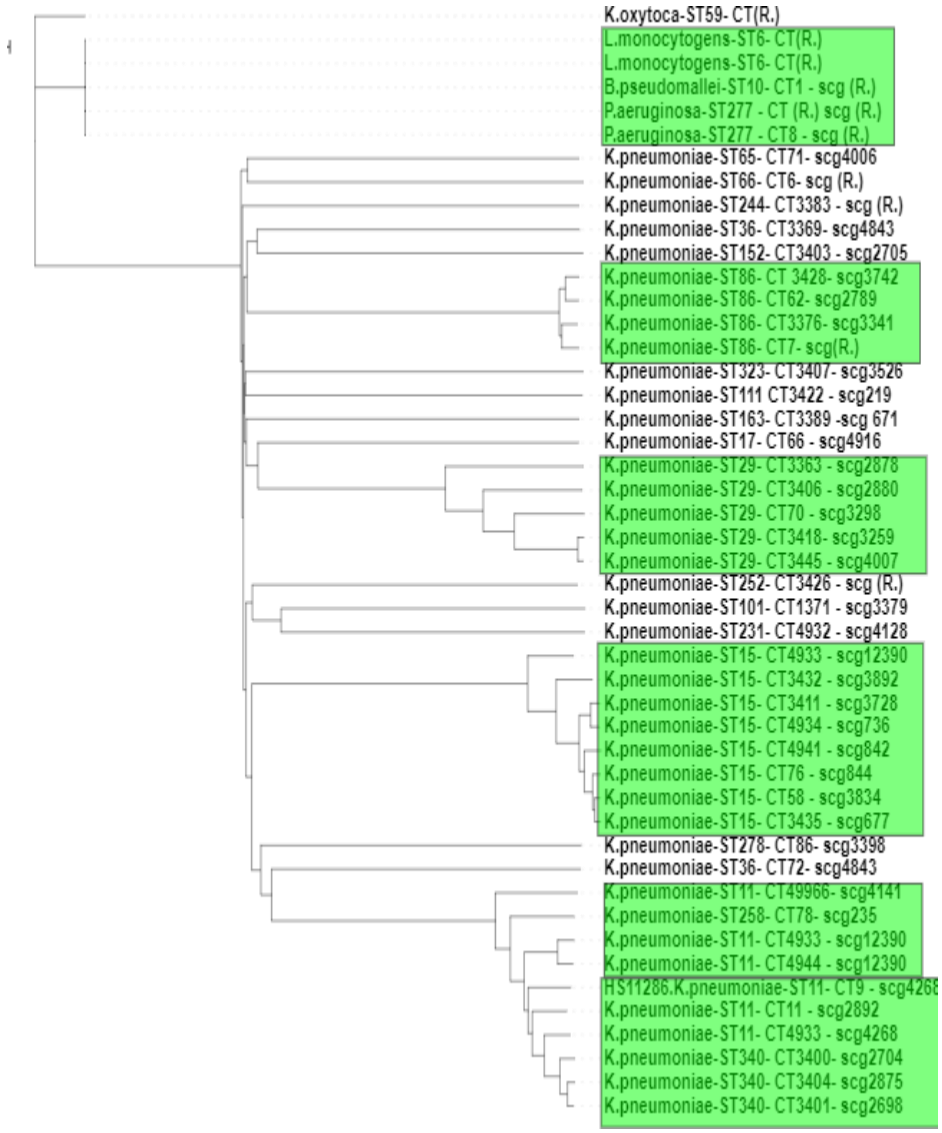

Figura 11 Árbol de unión de vecinos (NJ) basado en enfoque wgMLST con escala de 0.1 *K. pneumoniae* _ SeqSphere.

El árbol filogenético de unión de vecinos (NJ) utilizando el esquema SeqSphere con perfiles alélicos de *K. pneumoniae* en la herramienta MentaLiST (Figura 11), agrupó 47 genomas con un tamaño de perfil de 2360 loci. Al realizar la comparación con el árbol *K. pneumoniae* estándar (Figura 6), se evidencia un árbol con una topología diferente. Sin embargo, agrupa correctamente las mayorías de las cepas estrechamente relacionadas con un máximo de 20 diferencias alélicas de distancia. Excepto, las cepas del ST17, ST10, ST6, ST11 generando un falso positivo y el ST258 generando un falso negativo. Esto dado porqué la herramienta los agrupa con ST a los cuales no tienen cercanía.

Representación gráfica de los resultados de wgMLST: Arboles de unión de vecinos (NJ) *P.aeruginosa.*

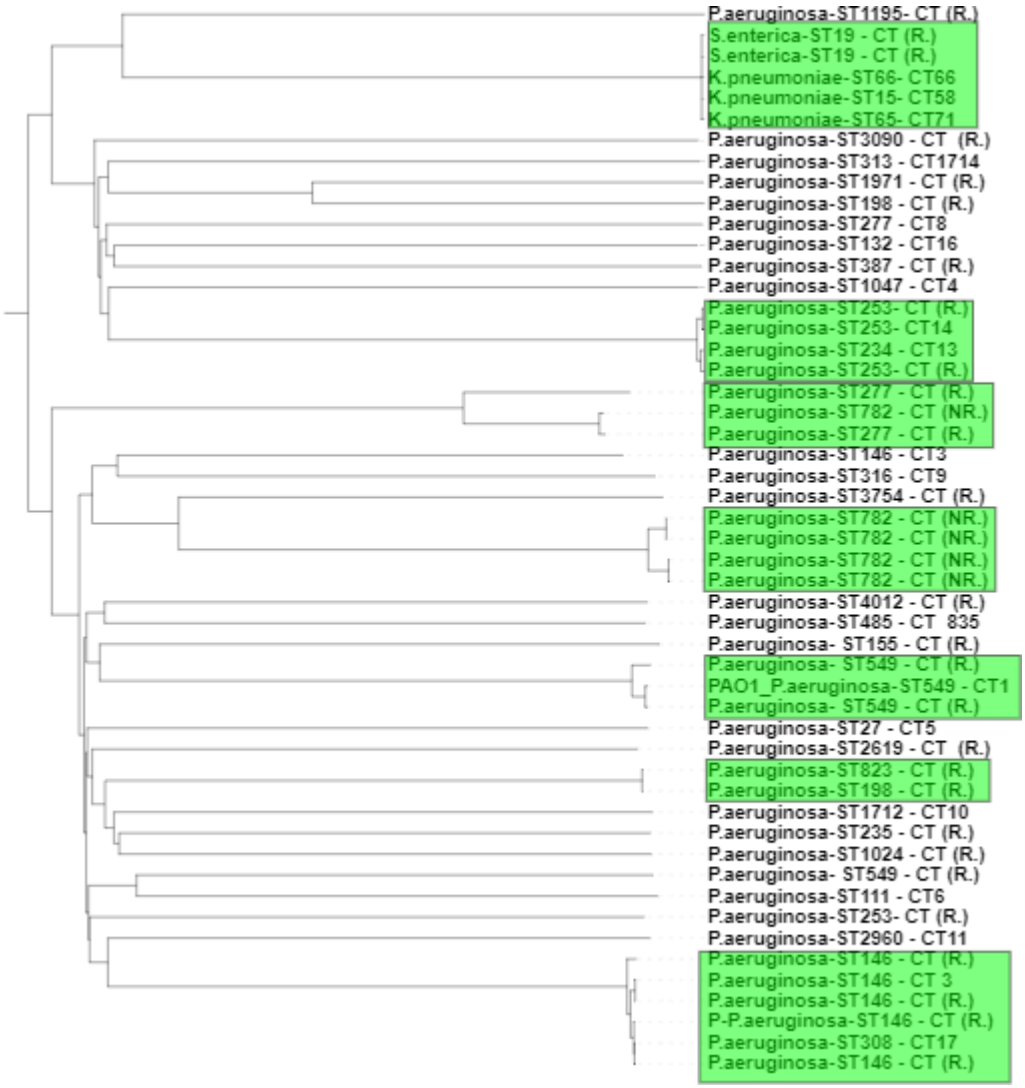

Figura 12 Árbol de unión de vecinos (NJ) mínima basado en enfoque wgMLST con escala 0.1 P.aeruginosa_ SeqSphere.

El árbol filogenético de unión de vecinos (NJ) utilizando el esquema SeqSphere con perfiles alélicos de *P. aeruginosa* en la herramienta MentaLiST (Figura 12), agrupó 52 genomas con un tamaño de perfil de 2869 loci. Al realizar la comparación con el árbol *P. aeruginosa* estándar (Figura 5) y el árbol *P. aeruginosa* obtenido con chewBACCA (Figura 7), se evidencia un árbol con una topología diferente. Sin embargo, agrupa correctamente las mayorías de las cepas estrechamente relacionadas con un máximo de 20 diferencias alélicas de distancia. excepto en las cepas del ST19 generando un falso positivo y el ST1239 generando un falso negativo. Esto dado porqué la herramienta los agrupa con ST a los cuales no tienen cercanía.

5.1.3.2.2 Ejecución de la herramienta bioinformáticas basada en nucleótidos.

Herramienta SAMtools.

El flujo de trabajo para genotipificar con la herramienta SAMtools, inicia con la alineación de los *reads* contra el genoma de referencia de *Pseudomonas aeruginosa* (NZ_PAO1.fasta) y *Klebsiella pneumoniae* (HS11286.fasta) utilizando el software de alineamiento BWA el cual acepta datos en formato FASTQ (Wingett & Andrews, 2018), y el software de alineamiento pareado Minimap2, el cual acepta formatos FASTA y FASTQ como entrada, permitiendo hacer análisis de genomas completos de cepas estrechamente relacionados (Hallgren et al., 2021). Sin embargo, no fue posible convertir alineaciones al archivo SAM/BAM para realizar el llamado e identificar SNPs entre otros. Esto dado a que los *dataset* utilizados por cada especie, son ensamblaje de genomas completos en formato FASTA y no *reads* como lo exige la guía de la herramienta. Por lo anterior, se excluyó la herramienta SAMtools del análisis bioinformático.

Herramienta GATK4.

El kit de herramientas *Genome Analysis Toolkit* (GATK), se actualizo de versión a partir de diciembre del año 2022 pasando del software GATK3 al GATK4. Este flujo de trabajo requiere la alineación de un genoma de referencia en formato FASTA (.fa.gz,) de *Pseudomonas aeruginosa* (NZ_PAO1.fasta) y *Klebsiella pneumoniae* (HS11286.fasta) contra las secuencias analizar en formato FASTQ (.fastq.gz) ejecutada con el alineador BWA para generar un archivo SAM y luego convertir al archivo BAM con la herramienta SAMtools, para posteriormente realizar la llamada variantes a través de GATK (Bathke & Lühken, 2021). Sin embargo, no fue posible realizar el trabajo en esta herramienta porque los *dataset* utilizados por cada especie son ensamblaje de genomas completos en formato FASTA y no en FASTQ. Por lo anterior, se excluyó la herramienta GATK4 del análisis bioinformático.

Herramienta Parsnp.

Parsnp es una herramienta rápida el cual utiliza genoma de referencia para identificar las posiciones de SNP gracias a MUMi, el cual ayuda a identificar rápidamente genomas estrechamente relacionados. Para análisis de SNP, se utilizaron los archivos (*parsnp.tree*) para visualizar el árbol filogenético esto dado a que los archivos (*parsnpAligner.log*, *parsnp.xmfa* y *parsnp.ggr*) no muestran información relevante para análisis de datos como él % de los genomas alineados, tamaño de alineación del genoma centra o posición SNP (Anexo 5).

Por otro lado, la herramienta no ejecutó al utilizar el *dataset* de la especie *K. pneumoniae* debido a que la distancia entre el genoma de referencia y cada uno de los genomas del *dataset* es MUMi = 0,1, lo que la herramienta Parsnp interpreta como demasiada divergencia entre los genomas y no arroja resultados, aunque se disminuya MUMi = 0,05.

La llamada de variantes (SNP*)* y las alineaciones múltiples para el *dataset* de la especie *Pseudomonas aeruginosa* el cual contiene 52 genomas, se realizó en un tiempo de 6 minutos aproximadamente utilizando 8 CPU.

Análisis filogenético del SNP de la especie *P. aeruginosa* con datos de la herramienta Parsnp.

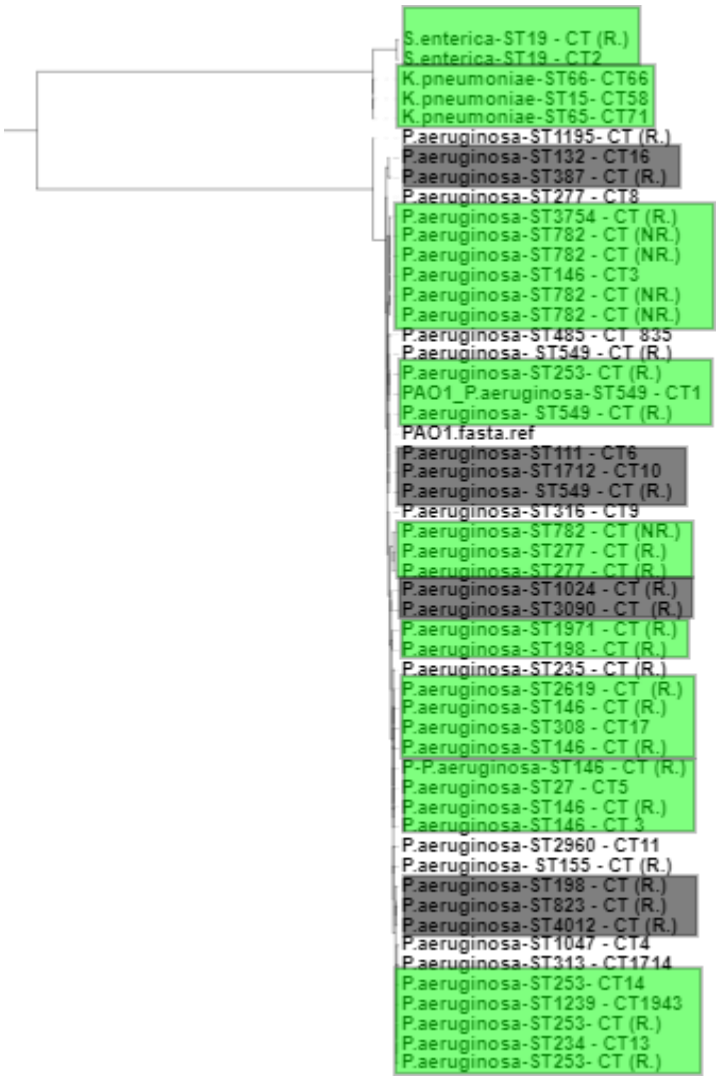

Figura 13 Árbol de máxima verosimilitud (ML) con escala 0.1 basado en enfoque SNP *P. aeruginosa* _Parsnp.

El árbol filogenético de máxima verosimilitud (ML) de la herramienta Parsnp construido a partir de alineaciones de genoma central de SNP (Figura 13), agrupó 52 genomas y muestra una topología diferente a la del árbol *P. aeruginosa* estándar (Figura 5). Generó 11 agrupamiento estrechamente relacionadas. Sin embargo, agrupo inadecuadamente los ST313, ST2619, ST27 en *clústers* a los cuales no tienen cercanía. Por lo tanto, se toman como falsos positivos. Esto puede deberse porqué Parsnp está diseñado para realizar alineamientos de genoma central de manera moderada, como consecuencia es menos sensible y precisa (Treangen et al., 2014). Cabe mencionar, que la herramienta ejecuto todo el proceso en 6 minutos utilizando 5GB.

Herramienta kSNP3.

La herramienta kSNP basada en *k-mers* es útil para identificar, anotar y generar árboles filogenéticos de unión de vecinos (NJ), máxima verosimilitud (ML) y parsimonia basados en SNP sin utilizar genoma de referencia (Gardner et al., 2015). A partir de diciembre del año 2022, la herramienta cambia de versión pasando de kSNP3 a KSNP4, una versión significativamente mejorada.

Los árboles filogenéticos se pueden visualizar en la herramienta MEGA X (Kumar et al., 2018) sugerida por los desarrolladores de kSNP o en su defecto por la herramienta iTOL (Letunic & Bork, 2019).

El flujo de trabajo inicia con la conversión del *dataset* de las especies *Pseudomonas aeruginosa* y *Klebsiella pneumoniae* utilizando el programa MakeKSNP4infile el cual genera un archivo que da la ruta a cada secuencia que contiene el *dataset* y un nombre para cada genoma, el cual se puede ver en los archivos (kpc*_file* y *Pseudomonas_file)* (Anexo5).

Posteriormente, el programa *genomeNames4* utiliza los archivos (kpc*_file* y *Pseudomonas_file*) para realizar la anotación de cada SNP en un genoma utilizando información recuperada del NCBI (Gardner et al., 2015). Como resultado se generan los archivos(*annotated*PSEUDO)para la especie *Pseudomonas aeruginosa* y (*annotatedGenomes*kpc) para *Klebsiella pneumoniae* (Anexo 5).

Para finalizar, el programa Kchooser4 utiliza los archivos de las especies (kpc*_file* y *Pseudomonas_file*) para determinar el valor óptimo de *k-mers* que KSNP4 identifica en todas las secuencias (Gardner et al., 2015), dando como resultado una longitud de *k-mers* a utilizar para *Pseudomonas aeruginosa de* 21 y *Klebsiella pneumoniae* de 17.

Una vez anotadas e identificadas las distancia *k-mers* para cada una de las especies, la herramienta KSNP4 realizó un análisis de SNP del genoma completo para evaluar la relación filogenética entre las cepas de cada *datase.*

La herramienta identificó un total de 345.652 SNP, de los cuales 123 SNP eran centrales, 345.529 SNP no centrales y al menos 184.414 SNP tenían una fracción del 0,75 en los genomas de *Pseudomonas aeruginosa.* En cuanto a los SNP de *Klebsiella pneumoniae*, KSNP4 identificó un total de 841.496 SNP, de los cuales 21 SNP son centrales, 841.475 SNP son no centrales y al menos 283.749 SNP tenían una fracción del 0,75 en los genomas. Esta información se puede recuperar de los archivos (*COUNT_SNPs* y *COUNT_coreSNPs*) para cada especie (Anexo 5).

Análisis filogenético SNP de las especies *Pseudomonas aeruginosa* y *Klebsiella pneumoniae* con datos de la herramienta KSNP4.

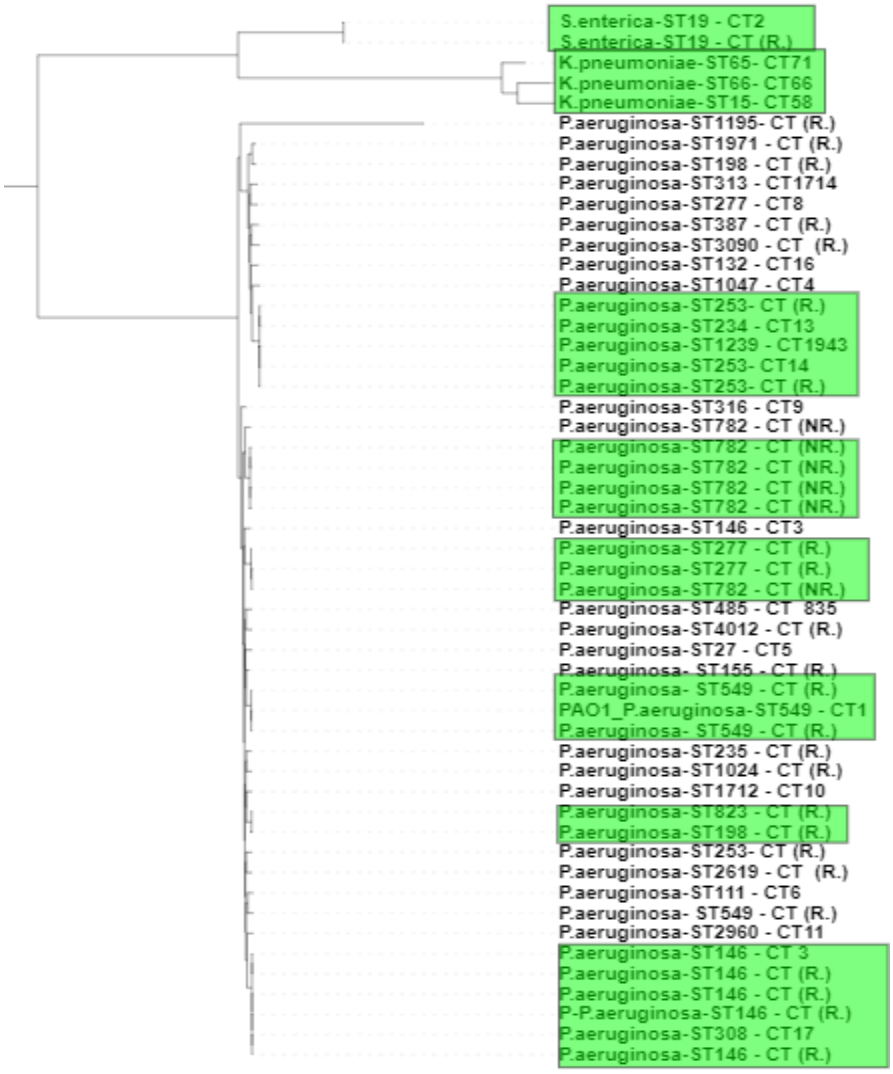

Figura 14 Árbol de máxima verosimilitud (ML) con escala 0.1 basado en enfoque SNP Pseudomonas aeruginosa kSNP.

El árbol filogenético de máxima verosimilitud (ML) (Figura 14) muestra una topología diferente al árbol *P. aeruginosa* estándar (Figura 5), agrupando 52 genomas con *k-mers* = 21. Sin embargo, se evidencia correctamente agrupados los genomas en los 7 *clústers* estrechamente relacionados incluidos los genomas diferentes a *Pseudomonas aeruginosa*. Cabe mencionar, que los genomas en cada uno de los *clústers* están en diferente posición con respeto al árbol estándar. Este proceso se ejecutó en 3 horas 22 minutos utilizando 7GB de memoria RAM.

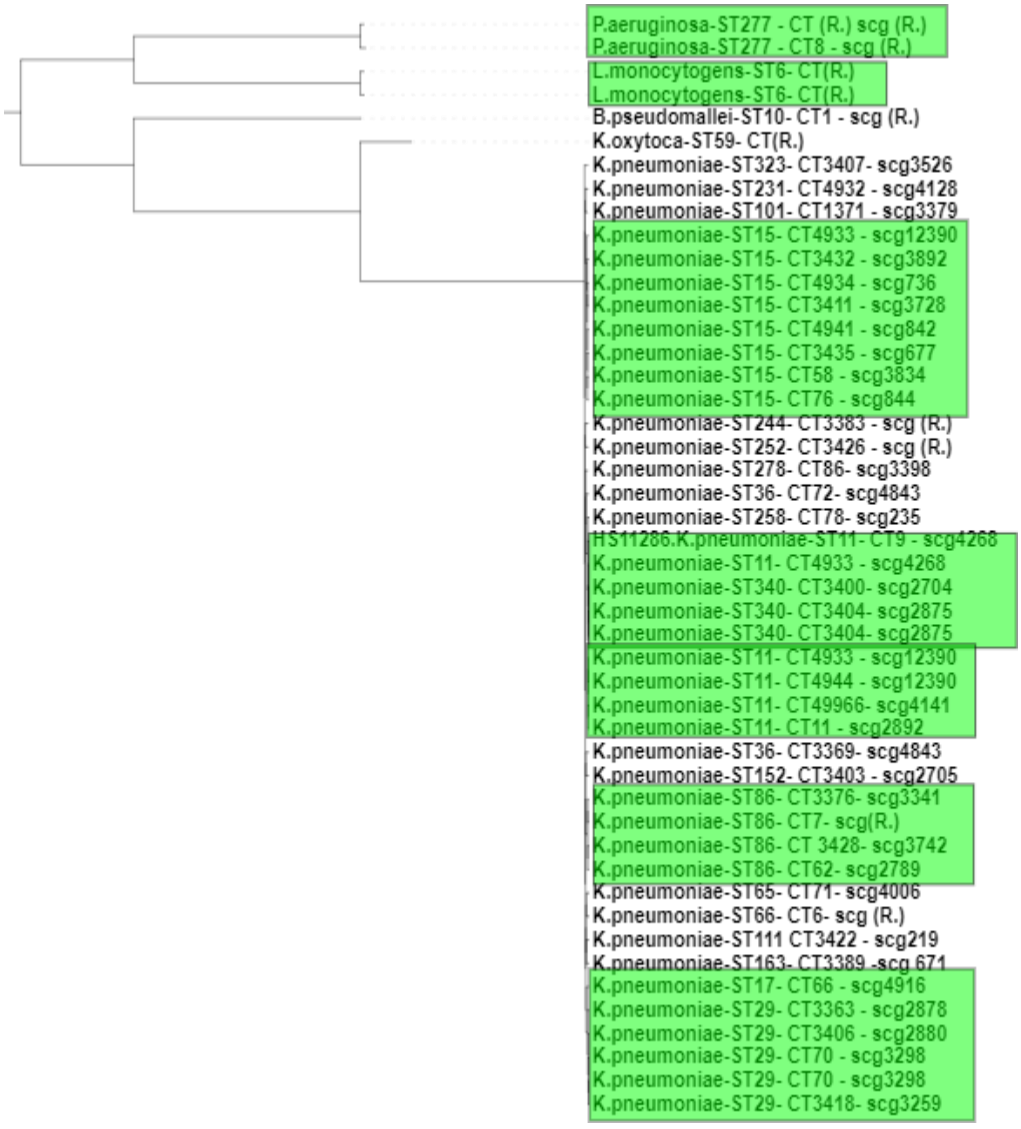

Figura 15 Árbol de máxima verosimilitud (ML) con escala 0.1 basado en enfoque SNP *Klebsiella Pneumoniae* kSNP.

Al igual que los árboles filogenéticos de *Pseudomonas aeruginosa* (Figura 14)*,* el árbol de *Klebsiella pneumoniae* (Figura 15) también muestra una topología diferente al árbol de *Klebsiella pneumoniae* estándar (Figura 6). Sin embargo, se evidencia correctamente.

agrupados los genomas en los 7 *clústers* estrechamente relacionados incluidos los genomas diferentes a *Pseudomonas aeruginosa*. Cabe mencionar, que los genomas en cada uno de los *clústers* están en diferente posición con respeto al árbol estándar. Este proceso se ejecutó en 2 horas 37 minutos utilizando 7GB de memoria RAM.

5.1.3.2.3 Ejecución de la herramienta bioinformáticas basada en pangenoma.

Herramienta Roary.

Roary es un software útil para construir pangenoma. Entendiendo como pangenoma a el número total de genes que están presentes en las cepas de una especie, los cuales pueden ser centrales, accesorio y específicos (Guimarães et al., 2015). El flujo de trabajo de Roary, inicia con la generación del formato de entrada GFF3 el cual se obtiene a través de la herramienta Prokka convirtiendo los archivos de los *datasets* que estaban en formato FASTA a archivos de anotación GFF3 (Page et al., 2015).

En la ejecución del proceso anterior, se generaron 17 archivos de salida para cada una de las especies que se están analizando en el presente trabajo. De los cuales, el archivo (*accessory_binary_genes.fa.newick*) se utilizó como entrada en la plataforma ITOL para visualizar el árbol filogenético de unión de vecinos (NJ*)* y el archivo (*summary_statistics.txt*) para identificar la totalidad de genes presentes en el pangenoma (Anexo 5).

Análisis filogenético del pan-genoma de las especies *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* con datos de la herramienta Roary.

Para el análisis Pangenómico, se estableció el parámetro del 95% de identidad de secuencia para definir la homología de las dos especies. obteniendo como resultado 29234 genes de los cuales 5387 genes hacen parte del genoma central para la especie *Klebsiella pneumoniae* y 24957 genes de los cuales 6324 genes hacen parte del genoma central para la especie *Pseudomonas aeruginosa.* Por lo tanto, se procede hacer el análisis filogenético.

El árbol filogenético de máxima verosimilitud (ML) la especie *Klebsiella pneumoniae* (Figura 16), construido a partir de un pan-genoma de 29234 genes, genero 8 grupos estrechamente relacionados. Sin embargo, al comparar con el árbol de *Klebsiella pneumoniae* estándar (Figura 6) se evidencia una topología diferente. Adicionalmente, la herramienta no agrupo correctamente las cepas (ST340 y ST11) las cuales comparten cercanía genómica con los genomas del grupo 3, lo que podría interpretarse como un falso positivo. Este proceso se ejecutó en 4 horas 40 minutos utilizando 3 GB de memoria RAM.

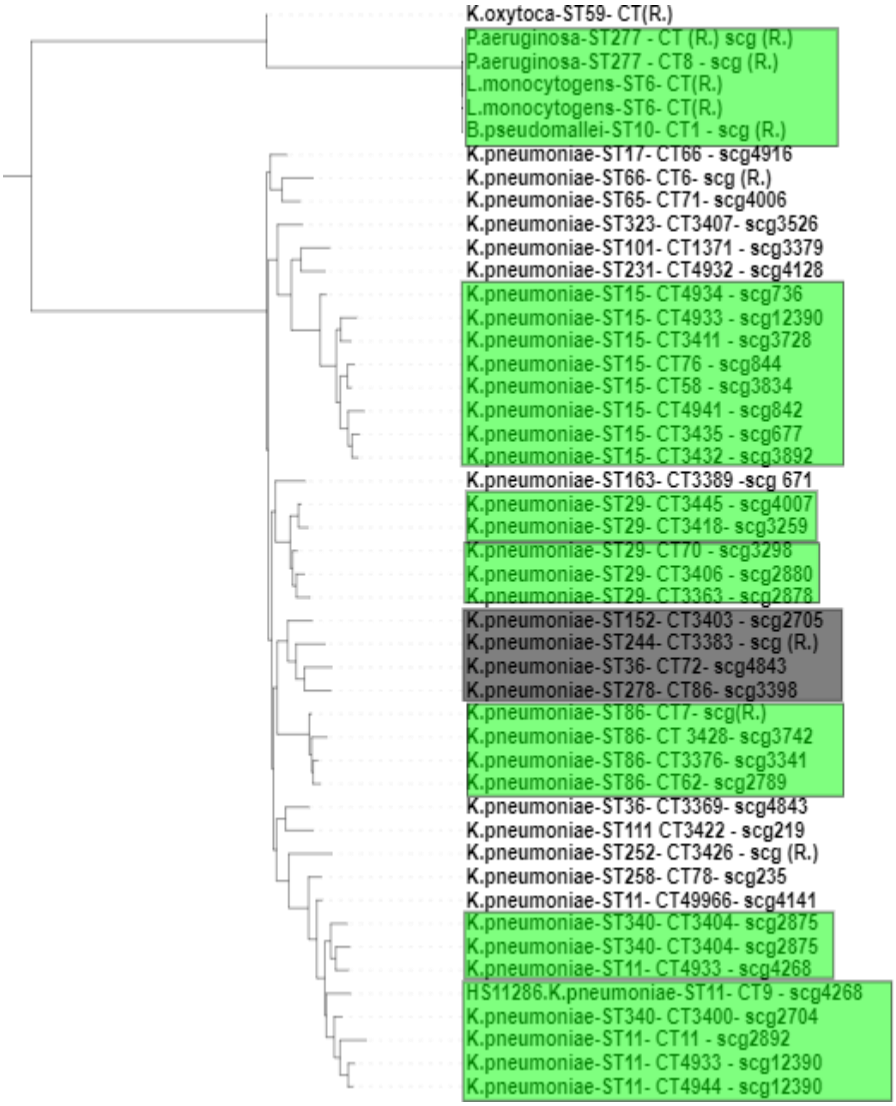

Figura 16 Árbol de máxima verosimilitud (ML) con escala 0.1 basado en enfoque pangenoma *Klebsiella Pneumoniae* Roary.

La construcción del árbol filogenético de *Pseudomonas aeruginosa* (Figura 17) se realizó con un pan-genoma de 24957 genes, generando 6 grupos filogenéticamente relacionados. Sin embargo, al comparar con el árbol estándar (Figura 5) se evidencian diferencias en la topología del árbol, esto dado a que la herramienta agrupa a *K. pneumoniae* y *S.enterica* en un solo *clústers* como si pertenecieran a una misma especie, lo que indica un error en la clasificación de los genes centrales debido al bajo porcentaje de similitud, lo que conlleva a variabilidad entre los genes de las especies.

Adicionalmente, no agrupo correctamente el ST 277 el cual debería de estar dentro del grupo 6 en donde su genoma tiene cercanía. Por la cual, se interpreta como un falso negativo. Este proceso se ejecutó en 5 horas 30 minutos utilizando 3 GB de memoria RAM.

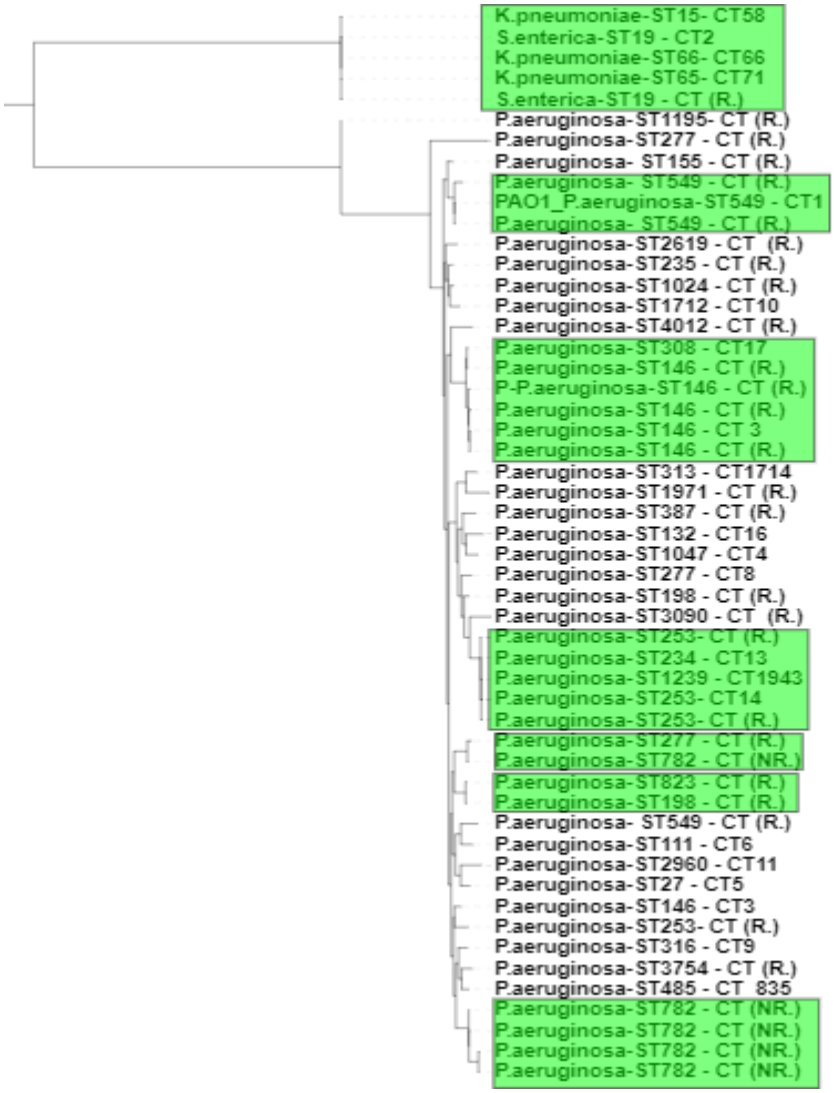

Figura 17Árbol de máxima verosimilitud (ML) con escala de 0.1 basado en enfoque pangenoma *Pseudomonas aeruginosa* Roary.

Herramienta PIRATE.

El flujo de trabajo para la creación del pangenoma para las especies *Pseudomonas aeruginosa* y *Klebsiella pneumoniae*, inicia con la adopción del Formato de Característica General GFF3 generado anteriormente la herramienta Prokka (P.gff y kpc.gff) para posteriormente construir el pangenoma.

Se ejecuto adecuadamente el *dataset* de las dos especies anteriormente nombradas, con múltiples umbrales de identidad de (50,60,70,80,90,95,100) para construir el pangenoma. Obteniendo así 19426 genes en 47 genomas para la especie *Klebsiella pneumoniae,* de los cuales 940 se clasificaron como centrales. De la misma manera, se construyó el pangenoma de la especie *Pseudomonas aeruginosa* con 15150 genes de los cuales 917 se clasificaron como centrales.

Análisis filogenético del pan-genoma de las especies *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* con datos de la herramienta PIRATE.

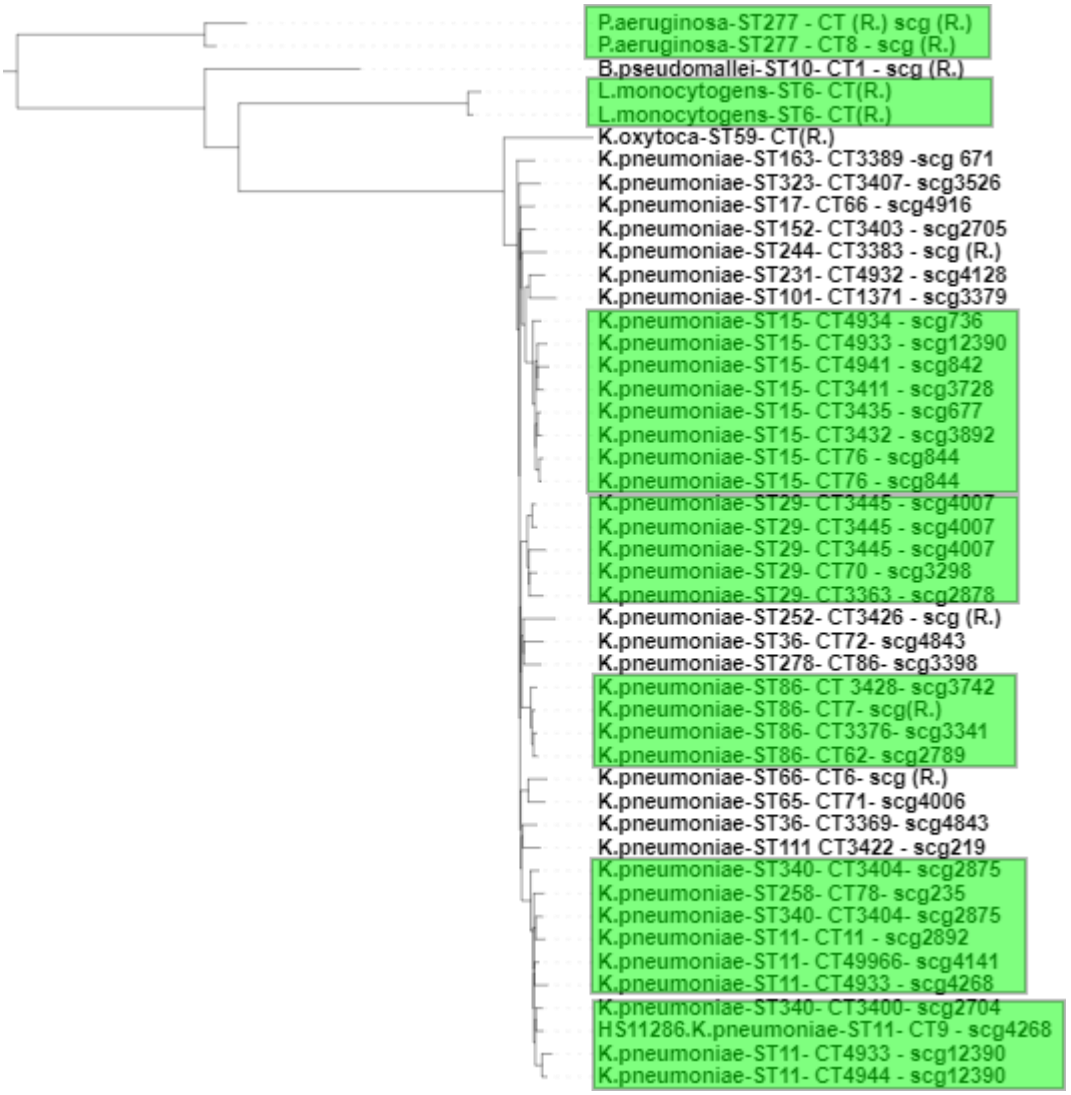

Figura 18 Árbol de máxima verosimilitud (ML) con escala 0.1 basado en enfoque pangenoma *Klebsiella Pneumoniae* PIRATE.

El árbol filogenético (Figura 18), se construyó con la matriz de presencia/ausencia (formato *binary_presence_absence.nwk)* de genes del pangenoma (Anexo 5). Al comparar con el árbol estándar (Figura 6), se observa claramente las 6 cepas filogenéticamente distantes conformado por tres especies divididas en *Pseudomonas aeruginosa* (ST 27), *Listeria monocytogenes* (ST 6),*Burkholderia pseudomallei* (ST 10) y *Klebsiella oxytoca* (ST 59). Adicionalmente, la herramienta agrupa incorrectamente el ST 258 en un grupo el cual no tiene cercanía genómica. Por tal motivo se clasifica como falso positivo. Por otra parte, el ST 17 no está agrupado en el *clústers* en el cual tiene cercanía genómica por lo que se clasifica como falso negativo.

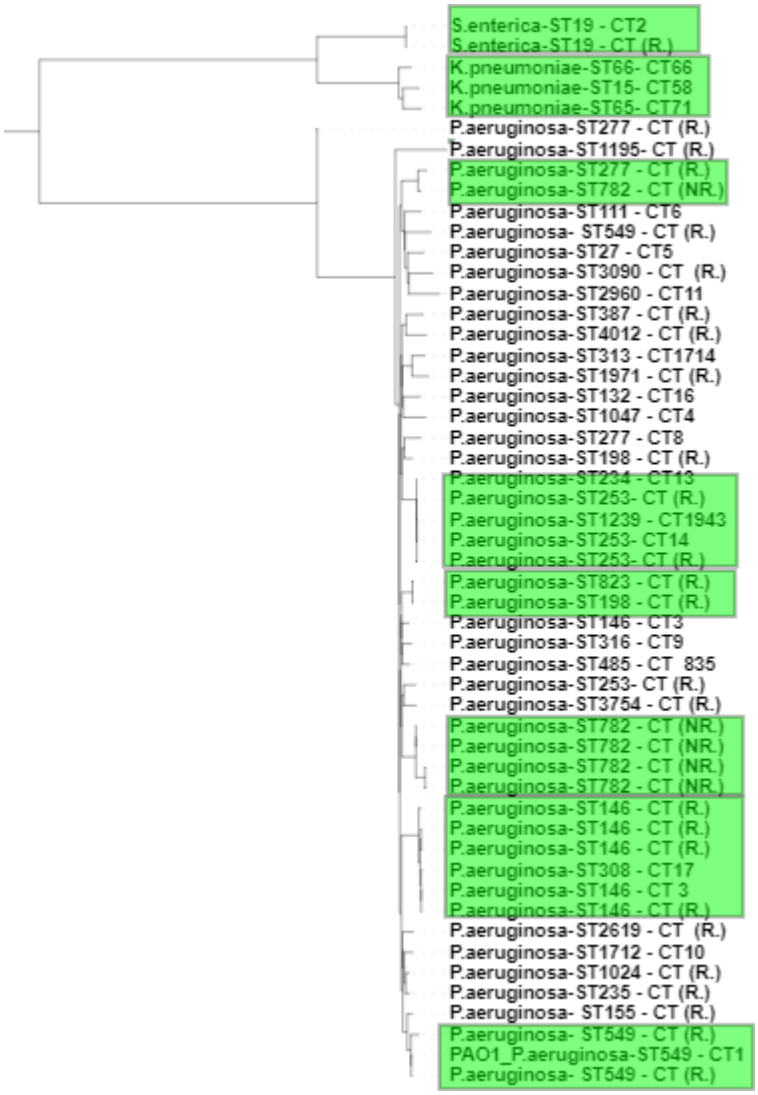

Figura 19 Árbol de máxima verosimilitud (ML) con escala 0.1 basado en enfoque pangenoma *Pseudomonas aeruginosa* PIRATE.

Al igual que con la especie *Klebsiella pneumoniae*, el árbol filogenético de la especie *Pseudomonas aeruginosa* (Figura 19) se construyó con la matriz de presencia/ausencia. Al comparar con el árbol estándar (Figura 5), se evidencia 5 cepas filogenéticamente

distantes a *Pseudomonas aeruginosa* conformado por *Klebsiella pneumoniae* (ST 66, ST 15, ST65) y *Salmonella enterica* (ST 19). Sin embargo, la herramienta no agrupa 2 adecuadamente un genoma (ST 277) dentro del grupo que tiene cercanía evolutiva, lo que se un falso positivo.

5.1.3.3 Desempeño de las Herramientas Bioinformáticas.

Se evaluó el rendimiento de las herramientas útiles para tipificar a partir de árboles filogenéticos derivados de cada herramienta, con el fin de determinar si las herramientas clasificaban y agrupaban adecuadamente las cepas que estaban estrechamente relacionas de aquellas que presentan mayor divergencia genómica mediante análisis filogenéticos.

La herramienta chewBBACA por tener enfoque cgMLST, se utilizó en su evaluación esquemas de tipificación adoptados correspondientes a SeqSphere, BIGSdb y ChewBBACA. Por tal motivo, aparece duplicada en el análisis.

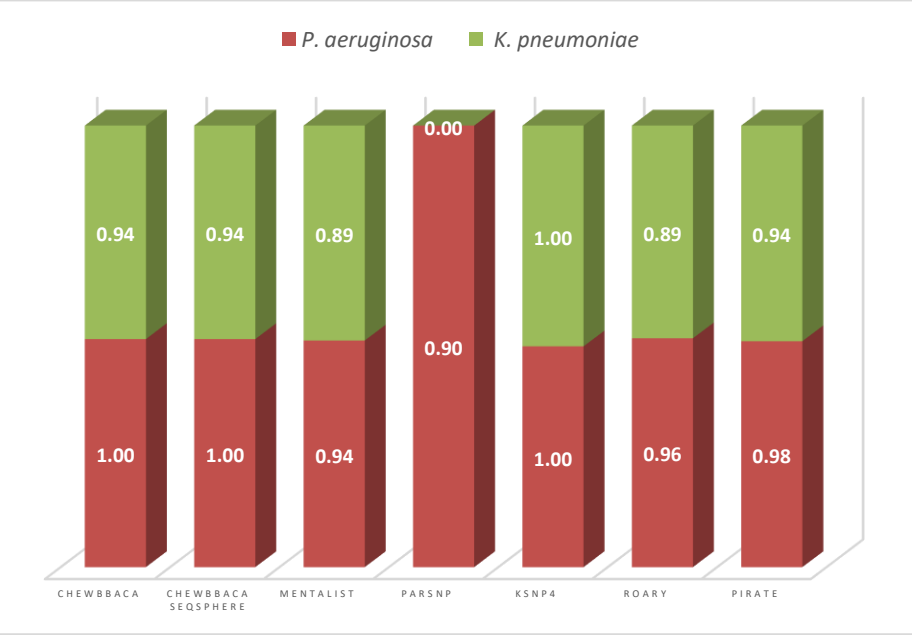

Figura 20 Comparación herramientas en la exactitud.

Se obtuvo la máxima precisión con la herramienta KSNP4 la cual utiliza un enfoque de SNP en comparación con las demás herramientas utilizadas, seguida de las herramientas chewBBACA cuyo enfoque está basado en cgMLST y PRIVATE que utiliza pangenoma las cuales obtuvieron en promedio 97%y 96% de concordancia entre el valor verdadero y el medido, indicando que clasifican mejor los datos (Figura 20). La demás herramienta está por debajo del 95% en promedio.

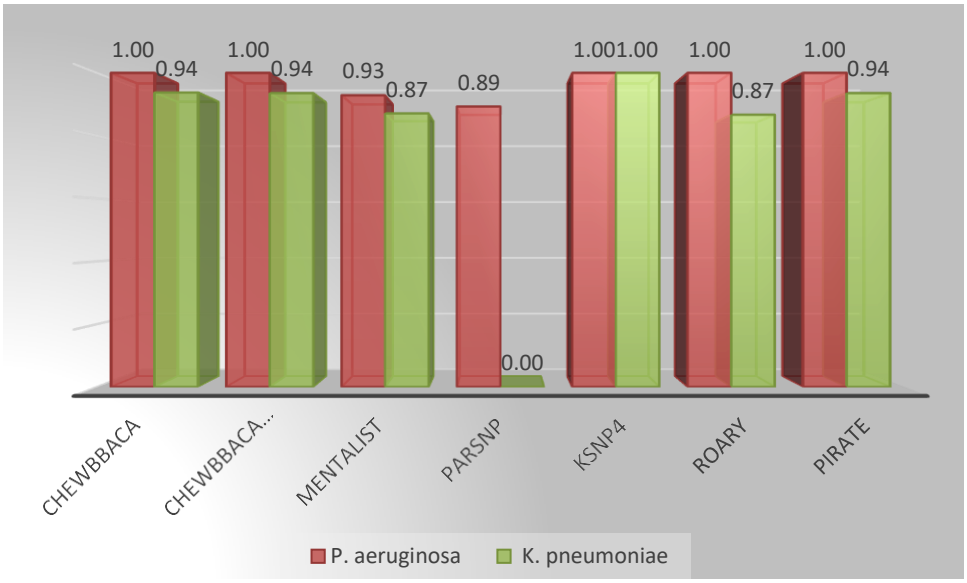

Figura 21 Comparación herramientas en la precisión.

Al igual que en la exactitud, la herramienta KSNP4 fue la más precisa con el 100%, seguida de chewBBACA y PRIVATE con el 97% en promedio, indicando que realizan mejor predicción de los cepas estrechamente relacionadas y menor probabilidad de falsos positivos (Figura 21). Las demás herramientas tienen una exactitud por debajo del 95%.

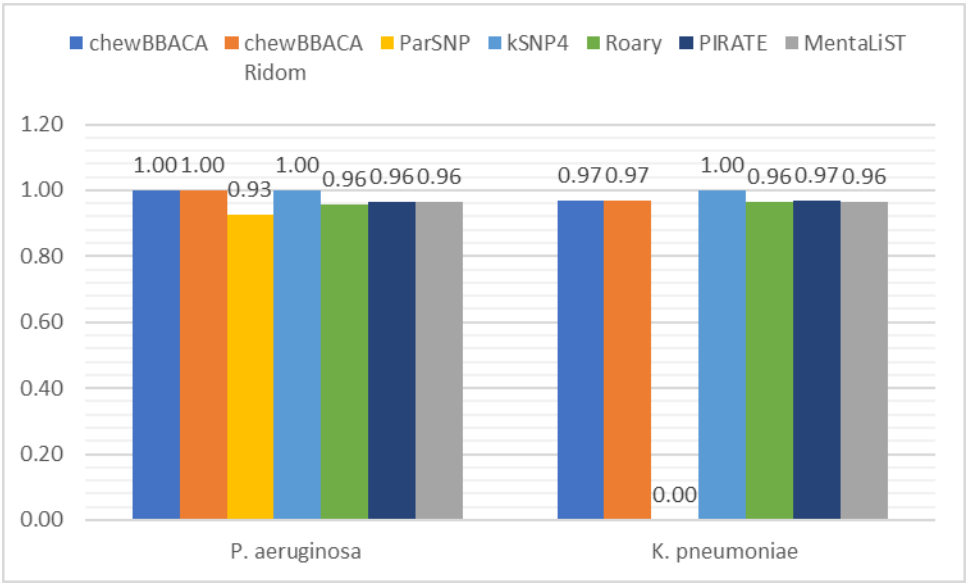

Figura 22 Comparación herramientas en la sensibilidad.

La mayoría de las herramientas tiene una sensibilidad por encima del 96% en promedio para identificar cepas estrechamente relacionadas (Figura 22). Excepto Parsnp, esto dado que la herramienta no ejecuto el *set* de datos de *Klebsiella pneumoniae.*

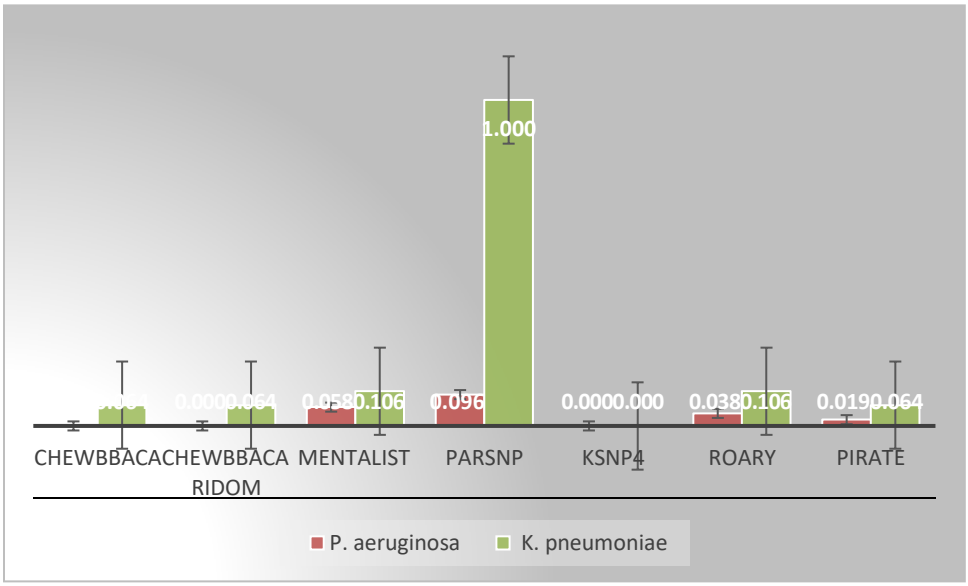

Figura 1 Comparación herramientas en la tasa de error

Adicionalmente, se verificó la tasa de error evidenciando que probabilidades de que las herramientas no agrupen las cepas altamente relacionadas están por debajo del 0.1% (Figura 23). En cuanto a la facilidad de uso (Anexo 3), todas las herramientas son fáciles de usar ya, que utilizan un lenguaje sencillo y fácil de interpretar en el momento de procesar los datos. Adicionalmente, todas las herramientas cuentan con un buzón de consulta en donde los usuarios pueden realizar consultas a sus desarrolladores con el fin de resolver problemas de uso que se puedan presentar.

De igual manera, se realizó la medición del tiempo de ejecución y la cantidad de memoria utilizada, evidenciando que las herramientas que más recursos consumieron en memoria RAM fueron kSNP4 con 7 GB y Parsnp con 6 GB en promedio. Sin embargo, la cantidad de memoria utilizada no fue proporcional al tiempo de ejecución de cada herramienta, evidenciando que las herramientas que usan enfoque pangenoma (kSNP4 ,PIRATE) tienen un tiempo promedio de ejecución de 4.5 horas, seguidas de kSNP4 con un tiempo de ejecución de 3 horas en promedio y chewBBACA con 1.5 horas de ejecución. Cabe mencionar que el proceso que más tiempo lleva es la alineación y búsqueda de homología cuando se utiliza BLAST.

6. Capítulo 6

6.1 Conclusiones y recomendaciones.

Conclusiones.

El análisis e interpretación de conjuntos de datos genómicos es un desafío importante, esto dado a la gran cantidad de herramientas con diferentes aproximaciones. Es por ello, por lo que se recomienda realizar evaluaciones comparativas para tener confianza en los resultados obtenidos

De acuerdo con los resultados obtenidos de la evaluación comparativa del presente trabajo, se puede observar que las herramientas KSNP4 con un enfoque de llamada de SNP, chewBBACA enfoque wgMLST/cgMLST y PRIVATE enfoque pangenoma presentaron mejor desempeño y poder discriminatorio entre sepas estrechamente relacionadas en comparación con las demás herramientas, lo que indica que se puede utilizar cualquiera de las tres herramientas para realizar la tipificación a partir de datos WGS.

Cabe mencionar, que ninguna de las herramientas evaluadas genera un tipo o nombre a las cepas, ya que su función final está relacionada con agrupar los genomas que estén mayormente relacionadas de acuerdo con unas características definidas e independiente en cada enfoque. Por tal motivo, se sugiere antes de ejecutar estas herramientas utilizar los servidores de nomenclatura alélica gratuitos como Enterobase, Instituto Pasteur y PubMLST, los cuales son gratuitos en la web y permiten asignarle un secuenciotipo al genoma de interés. Sin embargo, estos sitios presentan una gran desventaja relacionados con la nomenclatura wg/cgMLST, MLST o rMLST la cual no es universal para todas las especies lo que podría llevar a resultados diferentes entre los laboratorios.

Por último, el análisis filogenético de los datos de WGS de *Klebsiella pneumoniae* y *Pseudomonas aeruginosa* basada en wgMLST, cgMLST, SNP y pangenoma en relación con el árbol filogenético de referencia (Figura 5) (Figura6), arrojó resultados congruentes en cuanto al adecuado agrupamiento de los *out group.* Sin embargo, no se evidencia la misma filogenia de los genomas estrechamente relacionados en todos los árboles, debido a que algunas herramientas realizan la reconstrucción filogenética a través de algoritmos de distancias como Mínima Expansión y Neighbor-joining o de optimización como máxima verosimilitud y Parsimonia entre otros; los cuales difieren uno de otro así se utilice la misma muestra, lo que podría llevar a la interpretación errónea del resultado.

Recomendaciones.

La necesidad de la población científica y la epidemiologia para tipificar aislados bacterianos, ha desencadenado el desarrollo de un número importante de métodos, modelos, enfoque y esquemas, que han sido implementados en un sinfín de herramientas bioinformática las cuales presentan desafíos en su reproducibilidad, instalación, estabilidad, mantenimiento, actualización a corto, mediano y largo plazo. lo que puede limitar el uso de las herramientas o que afectar los resultados. Por ello, se hace necesario, implementar un consenso sobre el tipo de evaluación comparativa que se deba realizar a cada uno de los parámetros anteriormente nombrado al igual que las métricas que llevaría la evaluación.

De igual manera, se espera que las herramientas evaluadas en este trabajo sean útiles para el proceso de tipificación de otros estudios. Se recomienda realizar un estudio de las herramientas anteriormente nombradas solo con secuencias que contengan los *reads* de las cepas en formato FASTQ, esto dado se realizó la evaluación con secuencias ensambladas en formato FASTA, lo que podría cambiar los resultados generados por las herramientas.

Por último, se recomienda establecer una base de datos a nivel global en donde se almacenen los secuenciotipos y característica los esquemas cgMLST/wgMLST, SNP y pangenoma, con el fin de poder comparar los aislados bacterianos a nivel mundial independiente de las aproximaciones que utilicen de las herramientas.

A. Anexo 1, Set de Datos

https://github.com/ocarabali/Anexos.git

B. Anexo 2, Arboles filogenéticos

https://github.com/ocarabali/Anexos.git

C. Anexo 3, Valores de rendimiento de las herramientas

https://github.com/ocarabali/Anexos.git

D. Anexo 4, Bibliografía del estado del arte

https://github.com/ocarabali/Anexos.git

E. Anexo 5, Formatos adicionales.

Terminal168.176.54.15Ruta/vault2/homehpc/ocarabali

Bibliografía

Alikhan, N. F., Zhou, Z., Sergeant, M. J., & Achtman, M. (2018). A genomic overview of the population structure of Salmonella. In *PLoS Genetics* (Vol. 14, Issue 4, p. e1007261). Public Library of Science. https://doi.org/10.1371/journal.pgen.1007261

Allard, M. W. (2016). The future of whole-genome sequencing for public health and the clinic. In *Journal of Clinical Microbiology* (Vol. 54, Issue 8, pp. 1946–1948). American Society for Microbiology. https://doi.org/10.1128/JCM.01082-16

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. B., & Müller-Myhsok, B. (2012). A beginners guide to SNP calling from high-Throughput DNA-sequencing data. In *Human Genetics* (Vol. 131, Issue 10, pp. 1541–1554). Springer. https://doi.org/10.1007/s00439-012-1213-z

Anani, H., Zgheib, R., Hasni, I., Raoult, D., & Fournier, P. E. (2020). Interest of bacterial pangenome analyses in clinical microbiology. *Microbial Pathogenesis*, *149*. https://doi.org/10.1016/j.micpath.2020.104275

Basset, P., & Blanc, D. S. (2014). Fast and simple epidemiological typing of Pseudomonas aeruginosa using the double-locus sequence typing (DLST) method. *European Journal of Clinical Microbiology and Infectious Diseases*, *33*(6), 927–932. https://doi.org/10.1007/s10096-013-2028-0

Bathke, J., & Lühken, G. (2021). OVarFlow: a resource optimized GATK 4 based Open source Variant calling workFlow. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/S12859-021-04317-Y

Bayliss, S. C., Thorpe, H. A., Coyle, N. M., Sheppard, S. K., & Feil, E. J. (2019). PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *GigaScience*, *8*(10), 1–9. https://doi.org/10.1093/GIGASCIENCE/GIZ119

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, *45*(D1), D37–D42. https://doi.org/10.1093/nar/gkw1070

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., & Sayers, E. W. (2018). GenBank. *Nucleic Acids Research*, *46*(D1), D41–D47. https://doi.org/10.1093/nar/gkx1094

Blanc, D. S., Magalhães, B., Koenig, I., Senn, L., & Grandbastien, B. (2020). Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumericsTM) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa. *Frontiers in Microbiology*, *11*, 1729. https://doi.org/10.3389/fmicb.2020.01729

Bogaerts, B., Winand, R., Fu, Q., Van Braekel, J., Ceyssens, P.-J., Mattheus, W., Bertrand, S., De Keersmaecker, S. C. J., Roosens, N. H. C., & Vanneste, K. (2019). Validation of a Bioinformatics Workflow for Routine Analysis of Whole-Genome Sequencing Data and Related Challenges for Pathogen Typing in a European National Reference Center: Neisseria meningitidis as a Proof-of-Concept. *Frontiers in Microbiology*, *10*(MAR), 362. https://doi.org/10.3389/fmicb.2019.00362

Botes, J., Williamson, G., Sinickas, V., & Gürtler, V. (2003). Genomic typing of Pseudomonas aeruginosa isolates by comparison of Riboprinting and PFGE: correlation of experimental results with those predicted from the complete genome sequence of isolate PAO1. *Journal of Microbiological Methods*, *55*(1), 231–240. https://doi.org/10.1016/s0167-7012(03)00156-8

Bou, G., Fernández-Olmos, A., García, C., Sáez-Nieto, J. A., & Valdezate, S. (2011).

Métodos de identificación bacteriana en el laboratorio de microbiología. *Enfermedades Infecciosas y Microbiología Clínica*, *29*(8), 601–608. https://doi.org/10.1016/J.EIMC.2011.03.012

Brilhante, M., Gobeli Brawand, S., Endimiani, A., Rohrbach, H., Kittl, S., Willi, B., Schuller, S., & Perreten, V. (2021). Two high-risk clones of carbapenemase-producing Klebsiella pneumoniae that cause infections in pets and are present in the environment of a veterinary referral hospital. *Journal of Antimicrobial Chemotherapy*, *76*(5), 1140–1149. https://doi.org/10.1093/JAC/DKAB028

Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R., & Boulesteix, A. L. (2021). On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biology*, *22*(1). https://doi.org/10.1186/S13059-021-02365-4

Bush, S. J. (2021). Generalizable characteristics of false-positive bacterial variant calls. *Microbial Genomics*, *7*(8). https://doi.org/10.1099/MGEN.0.000615

Bush, S. J., Foster, D., Eyre, D. W., Clark, E. L., de Maio, N., Shaw, L. P., Stoesser, N., Peto, T. E. A., Crook, D. W., & Walker, A. S. (2020). Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. *GigaScience*, *9*(2), 1–21. https://doi.org/10.1093/GIGASCIENCE/GIAA007

Carattoli, A., Zankari, E., Garciá-Fernández, A., Larsen, M. V., Lund, O., Villa, L., Aarestrup, F. M., & 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

Chang, C. H., Chang, Y. C., Underwood, A., Chiou, C. S., & Kao, C. Y. (2007). VNTRDB: A bacterial variable number tandem repeat locus database. *Nucleic Acids Research*, *35*(SUPPL. 1). https://doi.org/10.1093/nar/gkl872

Chen, Y., Gonzalez-Escalona, N., Hammack, T. S., Allard, M. W., Strain, E. A., & Brown, E. W. (2016). Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of Listeria monocytogenes. *Applied and Environmental Microbiology*, *82*(20), 6258–6272. https://doi.org/10.1128/AEM.01532-16

Clarke, T. H., Brinkac, L. M., Inman, J. M., Sutton, G., & Fouts, D. E. (2018). PanACEA: A bioinformatics tool for the exploration and visualization of bacterial pan-chromosomes. *BMC Bioinformatics*, *19*(1), 246. https://doi.org/10.1186/s12859-018-2250-y

Coll, F., Gouliouris, T., Bruchmann, S., Phelan, J., Raven, K. E., Clark, T. G., Parkhill, J., & Peacock, S. J. (2022). PowerBacGWAS: a computational pipeline to perform power calculations for bacterial genome-wide association studies. *Communications Biology*, *5*(1). https://doi.org/10.1038/S42003-022-03194-2

Coll, F., Raven, K. E., Knight, G. M., Blane, B., Harrison, E. M., Leek, D., Enoch, D. A., Brown, N. M., Parkhill, J., & Peacock, S. J. (2020). Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant Staphylococcus aureus: a genomic epidemiology analysis. *The Lancet Microbe*, *1*(8), e328–e335. https://doi.org/10.1016/S2666-5247(20)30149-X

Dalsass, M., Bodini, M., Lambert, C., Mortier, M. C., Romanelli, M., Medini, D., Muzzi, A., & Brozzi, A. (2019). STRAIN: An R package for multi-locus sequence typing from whole genome sequencing data. *BMC Bioinformatics*, *20*(Suppl 9). https://doi.org/10.1186/s12859-019-2887-1

Das, D., Baruah, R., Sarma Roy, A., Singh, A. K., Deka Boruah, H. P., Kalita, J., & Bora,

T. C. (2015). Complete genome sequence analysis of Pseudomonas aeruginosa N002 reveals its genetic adaptation for crude oil degradation. *Genomics*, *105*(3), 182–190. https://doi.org/10.1016/j.ygeno.2014.12.006

De Rosa, F. G., Corcione, S., Pagani, N., & Di Perri, G. (2015). From ESKAPE to ESCAPE, From KPC to CCC. *Clinical Infectious Diseases*, *60*(8), 1289–1290. https://doi.org/10.1093/CID/CIU1170

Deneke, C., Uelze, L., Brendebach, H., Tausch, S. H., & Malorny, B. (2021). Decentralized Investigation of Bacterial Outbreaks Based on Hashed cgMLST. *Frontiers in Microbiology*, *12*, 874. https://doi.org/10.3389/fmicb.2021.649517

Escalona, M., Rocha, S., & Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. In *Nature Reviews Genetics* (Vol. 17, Issue 8, pp. 459–469). Nature Publishing Group. https://doi.org/10.1038/nrg.2016.57

Ettorchi -Tardy, A., Levif, M., & Michel, P. (2012). Benchmarking: A method for continuous quality improvement in health. *Healthcare Policy*, *7*(4). https://doi.org/10.12927/hcpol.2012.22872

Feijao, P., Yao, H. T., Fornika, D., Gardy, J., Hsiao, W., Chauve, C., & Chindelevitch, L. (2018). MentaLiST - A fast MLST caller for large MLST schemes. *Microbial Genomics*, *4*(2), e000146. https://doi.org/10.1099/mgen.0.000146

Fouts, D. E., Brinkac, L., Beck, E., Inman, J., & Sutton, G. (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Research*, *40*(22). https://doi.org/10.1093/NAR/GKS757

Francisco, A. P., Bugalho, M., Ramirez, M., & Carriço, J. A. (2009). Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics*, *10*(1), 1–15. https://doi.org/10.1186/1471-2105-10-152/FIGURES/5

Francisco, A. P., Vaz, C., Monteiro, P. T., Melo-Cristino, J., Ramirez, M., & Carriço, J. A. (2012). PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*, *13*(1), 87. https://doi.org/10.1186/1471-2105-13-87

Friedman, S., Gauthier, L., Farjoun, Y., & Banks, E. (2020). Lean and deep models for more accurate filtering of SNP and INDEL variant calls. *Bioinformatics (Oxford, England)*, *36*(7), 2060–2067. https://doi.org/10.1093/BIOINFORMATICS/BTZ901

Gardner, S. N., & Hall, B. G. (2013). When whole-genome alignments just won't work: KSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLoS ONE*, *8*(12). https://doi.org/10.1371/journal.pone.0081760

Gardner, S. N., Slezak, T., & Hall, B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics (Oxford, England)*, *31*(17), 2877–2878. https://doi.org/10.1093/BIOINFORMATICS/BTV271

Guigon, G., Cheval, J., Cahuzac, R., & Brisse, S. (2008). MLVA-NET--a standardised web database for bacterial genotyping and surveillance. *Euro Surveillance : Bulletin Européen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, *13*(19), 18863. https://doi.org/10.2807/ese.13.19.18863-en

Guimarães, L. C., Florczak-Wyspianska, J., Jesus, L. B. de, Viana, M. V. C., Silva, A., Ramos, R. T. J., Soares, S. de C., & Soares, S. de C. (2015). Inside the Pan-genome - Methods and Software Overview. *Current Genomics*, *16*(4), 245. https://doi.org/10.2174/1389202916666150423002311

Gupta, A., Jordan, I. K., & Rishishwar, L. (2017). stringMLST: a fast k-mer based tool for

multilocus sequence typing. *Bioinformatics*, *33*(1), 119–121. https://doi.org/10.1093/BIOINFORMATICS/BTW586

Gupta, A. K. (1996). Classification. *Springer Geology*, 69–87. https://doi.org/10.1007/978-81-322-2083-1_3

Hall, B. G. (2014). SNP-associations and phenotype predictions from hundreds of microbial genomes without genome alignments. *PLoS ONE*, *9*(2), 90490. https://doi.org/10.1371/journal.pone.0090490

Hallgren, M. B., Overballe-Petersen, S., Lund, O., Hasman, H., & Clausen, P. T. L. C. (2021). MINTyper: an outbreak-detection method for accurate and rapid SNP typing of clonal clusters with noisy long reads. *Biology Methods and Protocols*, *6*(1). https://doi.org/10.1093/BIOMETHODS/BPAB008

Henry, V. J., Bandrowski, A. E., Pepin, A. S., Gonzalez, B. J., & Desfeux, A. (2014). OMICtools: an informative directory for multi-omic data analysis. *Database: The Journal of Biological Databases and Curation*, *2014*. https://doi.org/10.1093/DATABASE/BAU069

Inouye, M., Dashnow, H., Raven, L. A., Schultz, M. B., Pope, B. J., Tomita, T., Zobel, J., & Holt, K. E. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine*, *6*(11), 90. https://doi.org/10.1186/s13073-014-0090-6

INS. (2018). *INFORME DE RESULTADOS DE LA VIGILANCIA POR LABORATORIO DE RESISTENCIA ANTIMICROBIANA EN INFECCIONES ASOCIADAS A LA ATENCIÓN EN SALUD*.

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., Wimalarathna, H., Harrison, O. B., Sheppard, S. K., Cody, A. J., & Maiden, M. C. J. (2012). Ribosomal multilocus sequence typing: Universal characterization of bacteria from domain to strain. *Microbiology*, *158*(4), 1005–1015. https://doi.org/10.1099/mic.0.055459-0

Jolley, K. A., Chan, M. S., & Maiden, M. C. J. (2004). mlstdbNet - Distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics*, *5*(1), 86. https://doi.org/10.1186/1471-2105-5-86

Jolley, K. A., & Maiden, M. C. J. (2014). Using MLST to study bacterial variation: Prospects in the genomic era. *Future Microbiology*, *9*(5), 623–630. https://doi.org/10.2217/fmb.14.24

Jonas, D., Spitzmüller, B., Daschner, F. D., Verhoef, J., & Brisse, S. (2004). Discrimination of Klebsiella pneumoniae and Klebsiella oxytoca phylogenetic groups and other Klebsiella species by use of amplified fragment length polymorphism. *Research in Microbiology*, *155*(1), 17–23. https://doi.org/10.1016/j.resmic.2003.09.011

Kimura, B. (2018). Will the emergence of core genome MLST end the role of in silico MLST? In *Food Microbiology* (Vol. 75, pp. 28–36). Academic Press. https://doi.org/10.1016/j.fm.2017.09.003

Kingry, L. C., Rowe, L. A., Respicio-Kingry, L. B., Beard, C. B., Schriefer, M. E., & Petersen, J. M. (2016). Whole genome multilocus sequence typing as an epidemiologic tool for Yersinia pestis. *Diagnostic Microbiology and Infectious Disease*, *84*(4), 275–280. https://doi.org/10.1016/j.diagmicrobio.2015.12.003

Kozyreva, V. K., Truong, C. L., Greninger, A. L., Crandall, J., Mukhopadhyay, R., & Chaturvedi, V. (2017). Validation and implementation of clinical laboratory

improvements act-compliant whole-genome sequencing in the public health microbiology laboratory. *Journal of Clinical Microbiology*, *55*(8), 2502–2520. https://doi.org/10.1128/JCM.00361-17

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, *35*(6), 1547. https://doi.org/10.1093/MOLBEV/MSY096

Kwong, J. C., Mccallum, N., Sintchenko, V., & Howden, B. P. (2015). Whole genome sequencing in clinical and public health microbiology. *Pathology*, *47*(3), 199–210. https://doi.org/10.1097/PAT.0000000000000235

Labbé, G., Kruczkiewicz, P., Robertson, J., Mabon, P., Schonfeld, J., Kein, D., Rankin, M. A., Gopez, M., Hole, D., Son, D., Knox, N., Laing, C. R., Bessonov, K., Taboada, E. N., Yoshida, C., Ziebell, K., Nichani, A., Johnson, R. P., Van Domselaar, G., & Nash, J. H. E. (2021). Rapid and accurate SNP genotyping of clonal bacterial pathogens with BioHansel. *Microbial Genomics*, *7*(9), 651. https://doi.org/10.1099/MGEN.0.000651

Larsen, M. V., Cosentino, S., Lukjancenko, O., Saputra, D., Rasmussen, S., Hasman, H., Sicheritz-Pontén, T., Aarestrup, F. M., Ussery, D. W., & Lund, O. (2014). Benchmarking of methods for genomic taxonomy. *Journal of Clinical Microbiology*, *52*(5), 1529–1539. https://doi.org/10.1128/JCM.02981-13

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256. https://doi.org/10.1093/NAR/GKZ239

Li, F., Wang, Y., Li, C., Marquez-Lago, T. T., Leier, A., Rawlings, N. D., Haffari, G., Revote, J., Akutsu, T., Chou, K. C., Purcell, A. W., Pike, R. N., Webb, G. I., Ian Smith, A., Lithgow, T., Daly, R. J., Whisstock, J. C., & Song, J. (2019). Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. *Briefings in Bioinformatics*, *20*(6), 2150. https://doi.org/10.1093/BIB/BBY077

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/BIOINFORMATICS/BTP352

Li, W., Raoult, D., & Fournier, P. E. (2009). Bacterial strain typing in the genomic era. In *FEMS Microbiology Reviews* (Vol. 33, Issue 5, pp. 892–916). Oxford Academic. https://doi.org/10.1111/j.1574-6976.2009.00182.x

Lindgreen, S., Adair, K. L., & Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, *6*. https://doi.org/10.1038/SREP19233

Liu, J., Li, L., Peters, B. M., Li, B., Chen, D., Xu, Z., & Shirtliff, M. E. (2018). Complete genomic analysis of multidrug-resistance Pseudomonas aeruginosa Guangzhou-Pae617, the host of megaplasmid pBM413. *Microbial Pathogenesis*, *117*, 265–269. https://doi.org/10.1016/j.micpath.2018.02.049

Liu, Y. Y., Chiou, C. S., & Chen, C. C. (2016). PGAdb-builder: A web service tool for creating pan-genome allele database for molecular fine typing. *Scientific Reports*, *6*. https://doi.org/10.1038/srep36213

Liu, Y. Y., Lin, J. W., & Chen, C. C. (2019). Cano-wgMLST_BacCompare: A bacterial

genome analysis platform for epidemiological investigation and comparative genomic analysis. In *Frontiers in Microbiology* (Vol. 10, Issue JULY). Frontiers Media S.A. https://doi.org/10.3389/fmicb.2019.01687

Luis, J. (2012). Hipótesis, Método & Diseño de Investigación. In *Daena: International Journal of Good Conscience* (Vol. 7, Issue 2).

Lüth, S., Deneke, C., Kleta, S., & Dahouk, S. Al. (2021). Translatability of wgs typing results can simplify data exchange for surveillance and control of listeria monocytogenes. *Microbial Genomics*, *7*(1), 1–12. https://doi.org/10.1099/mgen.0.000491

Magalhães, B., Valot, B., Abdelbary, M. M. H., Prod'hom, G., Greub, G., Senn, L., & Blanc, D. S. (2020). Combining Standard Molecular Typing and Whole Genome Sequencing to Investigate Pseudomonas aeruginosa Epidemiology in Intensive Care Units. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.00003

Maiden, M. C. J., Rensburg, M. J. J. van, Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., & McCarthy, N. D. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews. Microbiology*, *11*(10), 728. https://doi.org/10.1038/NRMICRO3093

Maiden, M. C. J., Van Rensburg, M. J. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., & McCarthy, N. D. (2013). MLST revisited: The gene-by-gene approach to bacterial genomics. In *Nature Reviews Microbiology* (Vol. 11, Issue 10, pp. 728–736). https://doi.org/10.1038/nrmicro3093

Mamede, R., Vila-Cerqueira, P., Silva, M., Carriço, J. A., & Ramirez, M. (2021). Chewie Nomenclature Server (chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas. *Nucleic Acids Research*, *49*(D1), D660–D666. https://doi.org/10.1093/NAR/GKAA889

Mangul, S., Martin, L. S., Hill, B. L., Lam, A. K. M., Distler, M. G., Zelikovsky, A., Eskin, E., & Flint, J. (2019). Systematic benchmarking of omics computational tools. In *Nature Communications* (Vol. 10, Issue 1, pp. 1–11). Nature Publishing Group. https://doi.org/10.1038/s41467-019-09406-4

Marcos-Zambrano, L. J., Escribano, P., Bouza, E., & Guinea, J. (2014). Use of molecular typing tools for the study of hospital outbreaks of candidemia. *Revista Iberoamericana de Micologia*, *31*(2), 97–103. https://doi.org/10.1016/j.riam.2013.06.003

Martin, J., Phan, H. T. T., Findlay, J., Stoesser, N., Pankhurst, L., Navickaite, I., De Maio, N., Eyre, D. W., Toogood, G., Orsi, N. M., Kirby, A., Young, N., Turton, J. F., Hill, R. L. R., Hopkins, K. L., Woodford, N., Peto, T. E. A., Walker, A. S., Crook, D. W., & Wilcox, M. H. (2017). Covert dissemination of carbapenemase-producing Klebsiella pneumoniae (KPC) in a successfully controlled outbreak: long- and short-read whole-genome sequencing demonstrate multiple genetic modes of transmission. *The Journal of Antimicrobial Chemotherapy*, *72*(11), 3025–3034. https://doi.org/10.1093/jac/dkx264

Martínez-Carranza, E., García-Reyes, S., González-Valdez, A., & Soberón-Chávez, G. (2020). Tracking the genome of four Pseudomonas aeruginosa isolates that have a defective Las quorum-sensing system, but are still virulent. *Access Microbiology*, *2*(7). https://doi.org/10.1099/acmi.0.000132

Merchán, M. A., Caicedo, M. I. T., & Torres, A. K. D. (2017). Técnicas de Biología Molecular en el desarrollo de la investigación. Revisión de la literatura. *Revista*

Habanera de Ciencias Medicas, *16*(5), 796–807.

Michael Dunne, W., Pouseele, H., Monecke, S., Ehricht, R., & van Belkum, A. (2018). Epidemiology of transmissible diseases: Array hybridization and next generation sequencing as universal nucleic acid-mediated typing tools. *Infection, Genetics and Evolution*, *63*, 332–345. https://doi.org/10.1016/j.meegid.2017.09.019

Mirande, C., Bizine, I., Giannetti, A., Picot, N., & van Belkum, A. (2018). Epidemiological aspects of healthcare-associated infections and microbial genomics. *European Journal of Clinical Microbiology and Infectious Diseases*, *37*(5), 823–831. https://doi.org/10.1007/S10096-017-3170-X/TABLES/4

Miro, E., Rossen, J. W. A., Chlebowicz, M. A., Harmsen, D., Brisse, S., Passet, V., Navarro, F., Friedrich, A. W., & García-Cobos, S. (2020). Core/Whole Genome Multilocus Sequence Typing and Core Genome SNP-Based Typing of OXA-48-Producing Klebsiella pneumoniae Clinical Isolates From Spain. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.02961

Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2019). Integrated omics: Tools, advances and future approaches. In *Journal of Molecular Endocrinology* (Vol. 62, Issue 1, pp. R21–R45). BioScientifica Ltd. https://doi.org/10.1530/JME-18-0055

Moradigaravand, D., Martin, V., Peacock, S. J., & Parkhill, J. (2017). Evolution and epidemiology of multidrug-resistant Klebsiella pneumoniae in the United Kingdom and Ireland. *MBio*, *8*(1). https://doi.org/10.1128/mBio.01976-16

Nadon, C. A., Trees, E., Ng, L. K., Møller Nielsen, E., Reimer, A., Maxwell, N., Kubota, K. A., & Gerner-Smidt, P. (2013). Development and application of MLVA methods as a tool for inter-laboratory surveillance. *Eurosurveillance*, *18*(35), 20565. https://doi.org/10.2807/1560-7917.ES2013.18.35.20565

Neoh, H. min, Tan, X. E., Sapri, H. F., & Tan, T. L. (2019). Pulsed-field gel electrophoresis (PFGE): A review of the “gold standard” for bacteria typing and current alternatives. In *Infection, Genetics and Evolution* (Vol. 74). Elsevier B.V. https://doi.org/10.1016/j.meegid.2019.103935

Nguyen, K. T., Bonasera, R., Benson, G., Hernandez-Morales, A. C., Gill, J. J., & Liu, M. (2019). Complete Genome Sequence of Klebsiella pneumoniae Myophage May. *Microbiology Resource Announcements*, *8*(19). https://doi.org/10.1128/MRA.00252-19

Noller, A. C., McEllistrem, M. C., Pacheco, A. G. F., Boxrud, D. J., & Harrison, L. H. (2003). Multilocus Variable-Number Tandem Repeat Analysis Distinguishes Outbreak and Sporadic Escherichia coli O157:H7 Isolates. *Journal of Clinical Microbiology*, *41*(12), 5389–5397. https://doi.org/10.1128/JCM.41.12.5389-5397.2003

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Papić, B., Diricks, M., & Kušar, D. (2021). Analysis of the Global Population Structure of Paenibacillus larvae and Outbreak Investigation of American Foulbrood Using a Stable wgMLST Scheme. *Frontiers in Veterinary Science*, *8*, 582677. https://doi.org/10.3389/fvets.2021.582677

Payne, M., Kaur, S., Wang, Q., Hennessy, D., Luo, L., Octavia, S., Tanaka, M. M., Sintchenko, V., & Lan, R. (2020). Multilevel genome typing: genomics-guided scalable resolution typing of microbial pathogens. *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease*

Bulletin, *25*(20). https://doi.org/10.2807/1560-7917.ES.2020.25.20.1900519

Peix, A., Ramírez-Bahena, M. H., & Velázquez, E. (2009). Historical evolution and current status of the taxonomy of genus Pseudomonas. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *9*(6), 1132–1147. https://doi.org/10.1016/J.MEEGID.2009.08.001

Peix, A., Ramírez-Bahena, M. H., & Velázquez, E. (2018). The current status on the taxonomy of Pseudomonas revisited: An update. *Infection, Genetics and Evolution*, *57*, 106–116. https://doi.org/10.1016/j.meegid.2017.10.026

Pérez-Losada, M., Arenas, M., & Castro-Nallar, E. (2018). Microbial sequence typing in the genomic era. *Infection, Genetics and Evolution*, *63*, 346–359. https://doi.org/10.1016/j.meegid.2017.09.022

Perrin, A., & Rocha, E. P. C. (2021). PanACoTA: a modular tool for massive microbial comparative genomics. *NAR Genomics and Bioinformatics*, *3*(1), lqaa106. https://doi.org/10.1093/nargab/lqaa106

Platt, S., Pichon, B., George, R., & Green, J. (2006). *RESEARCH NOTE A bioinformatics pipeline for high-throughput microbial multilocus sequence typing (MLST) analyses*. https://doi.org/10.1111/j.1469-0691.2006.01541.x

Riley, L. W. (2018). Laboratory Methods in Molecular Epidemiology: Bacterial Infections. *Microbiology Spectrum*, *6*(6). https://doi.org/10.1128/MICROBIOLSPEC.AME-0004-2018

Robinson, M. D., & Vitek, O. (2019). Benchmarking comes of age. *Genome Biology*, *20*(1). https://doi.org/10.1186/S13059-019-1846-5

Rouli, L., Merhej, V., Fournier, P. E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, *7*, 72–85. https://doi.org/10.1016/j.nmni.2015.06.005

Royer, G., Fourreau, F., Boulanger, B., Mercier-Darty, M., Ducellier, D., Cizeau, F., Potron, A., Podglajen, I., Mongardon, N., & Decousser, J. W. (2020). outbreak. *Journal of Hospital Infection*, *104*(1), 33–39. https://doi.org/10.1016/j.jhin.2019.07.014

Ruppé, E., Olearo, F., Pires, D., Baud, D., Renzi, G., Cherkaoui, A., Goldenberger, D., Huttner, A., François, P., Harbarth, S., & Schrenzel, J. (2017). Clonal or not clonal? Investigating hospital outbreaks of KPC-producing Klebsiella pneumoniae with whole-genome sequencing. *Clinical Microbiology and Infection*, *23*(7), 470–475. https://doi.org/10.1016/j.cmi.2017.01.015

Saharman, Y. R., Pelegrin, A. C., Karuniawati, A., Sedono, R., Aditianingsih, D., Goessens, W. H. F., Klaassen, C. H. W., van Belkum, A., Mirande, C., Verbrugh, H. A., & Severin, J. A. (2019). Epidemiology and characterisation of carbapenem-non-susceptible Pseudomonas aeruginosa in a large intensive care unit in Jakarta, Indonesia. *International Journal of Antimicrobial Agents*, *54*(5), 655–660. https://doi.org/10.1016/j.ijantimicag.2019.08.003

Sahl, J. W., Lemmer, D., Travis, J., Schupp, J. M., Gillece, J. D., Aziz, M., Driebe, E. M., Drees, K. P., Hicks, N. D., Williamson, C. H. D., Hepp, C. M., Smith, D. E., Roe, C., Engelthaler, D. M., Wagner, D. M., & Keim, P. (2016). NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. *Microbial Genomics*, *2*(8), e000074. https://doi.org/10.1099/mgen.0.000074

Salipante, S. J., SenGupta, D. J., Cummings, L. A., Land, T. A., Hoogestraat, D. R., &

Cookson, B. T. (2015). Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *Journal of Clinical Microbiology*, *53*(4), 1072–1079. https://doi.org/10.1128/JCM.03385-14

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., … McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nature Methods*, *14*(11), 1063–1071. https://doi.org/10.1038/nmeth.4458

Seth-Smith, H. M. B., Bonfiglio, F., Cuénod, A., Reist, J., Egli, A., & Wüthrich, D. (2019). Evaluation of Rapid Library Preparation Protocols for Whole Genome Sequencing Based Outbreak Investigation. *Frontiers in Public Health*, *7*(AUG), 241. https://doi.org/10.3389/fpubh.2019.00241

Singh, M., Malik, M. A., Singh, D. K., Doimari, S., Bhavna, & Sharma, R. (2020). Multilocus variable number tandem repeat analysis (MLVA)-typing of Brucella abortus isolates of India reveals limited genetic diversity. *Tropical Animal Health and Production*, *52*(3), 1187–1194. https://doi.org/10.1007/s11250-019-02110-x

Sitto, F., & Battistuzzi, F. U. (2020). Estimating Pangenomes with Roary. *Molecular Biology and Evolution*, *37*(3), 933–939. https://doi.org/10.1093/MOLBEV/MSZ284

Tadee, P., Tadee, P., Hitchings, M. D., Pascoe, B., Sheppard, S. K., & Patchanee, P. (2018). High Resolution Whole Genome Multilocus Sequence Typing (wgMLST) Schemes for Salmonella enterica Weltevreden Epidemiologic Investigations. *Biotechnol. Lett*, *46*(2), 162–170. https://doi.org/10.4014/mbl.1802.02008

Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelsen, P. A., Murray, B. E., Persing, D. H., & Swaminathan, B. (1995). Interpreting chromosomal DNA restriction patterns produced by pulsed- field gel electrophoresis: Criteria for bacterial strain typing. In *Journal of Clinical Microbiology* (Vol. 33, Issue 9, pp. 2233–2239). American Society for Microbiology. https://doi.org/10.1128/jcm.33.9.2233-2239.1995

Timme, R. E., Rand, H., Shumway, M., Trees, E. K., Simmons, M., Agarwala, R., Davis, S., Tillman, G. E., Defibaugh-Chavez, S., Carleton, H. A., Klimke, W. A., & Katz, L. S. (2017). Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ*, *2017*(10). https://doi.org/10.7717/peerj.3893

Tissot, F., Blanc, D. S., Basset, P., Zanetti, G., Berger, M. M., Que, Y. A., Eggimann, P., & Senn, L. (2016). New genotyping method discovers sustained nosocomial Pseudomonas aeruginosa outbreak in an intensive care burn unit. *Journal of Hospital Infection*, *94*(1), 2–7. https://doi.org/10.1016/j.jhin.2016.05.011

Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, *15*(11), 1–15. https://doi.org/10.1186/S13059-014-0524-X/TABLES/4

Tsai, M. H., Liu, Y. Y., & Soo, V. W. (2017). PathoBacTyper: A web server for pathogenic bacteria identification and molecular genotyping. *Frontiers in Microbiology*, *8*(AUG), 1474. https://doi.org/10.3389/fmicb.2017.01474

Uelze, L., Grützke, J., Borowiak, M., Hammerl, J. A., Juraschek, K., Deneke, C., Tausch, S. H., & Malorny, B. (2020). Typing methods based on whole genome sequencing data. *One Health Outlook*, *2*(1), 1–19. https://doi.org/10.1186/s42522-020-0010-1

van Beek, J., Räisänen, K., Broas, M., Kauranen, J., Kähkölä, A., Laine, J., Mustonen, E., Nurkkala, T., Puhto, T., Sinkkonen, J., Torvinen, S., Vornanen, T., Vuento, R.,

Jalava, J., & Lyytikäinen, O. (2019). Tracing local and regional clusters of carbapenemase-producing Klebsiella pneumoniae ST512 with whole genome sequencing, Finland, 2013 to 2018. *Eurosurveillance*, *24*(38). https://doi.org/10.2807/1560-7917.ES.2019.24.38.1800522

Van Belkum, A., Struelens, M., De Visser, A., Verbrugh, H., & Tibayrenc, M. (2001). Role of Genomic Typing in Taxonomy, Evolutionary Genetics, and Microbial Epidemiology. *Clinical Microbiology Reviews*, *14*(3), 547. https://doi.org/10.1128/CMR.14.3.547-560.2001

van Dorp, L., Wang, Q., Shaw, L. P., Acman, M., Brynildsrud, O. B., Eldholm, V., Wang, R., Gao, H., Yin, Y., Chen, H., Ding, C., Farrer, R. A., Didelot, X., Balloux, F., & Wang, H. (2019). Rapid phenotypic evolution in multidrug-resistant Klebsiella pneumoniae hospital outbreak strains. *Microbial Genomics*, *5*(4), 1–11. https://doi.org/10.1099/mgen.0.000263

Vaz, C., Francisco, A. P., Silva, M., Jolley, K. A., Bray, J. E., Pouseele, H., Rothganger, J., Ramirez, M., & Carriço, J. A. (2014). TypOn: the microbial typing ontology. *Journal of Biomedical Semantics*, *5*(1), 43. https://doi.org/10.1186/2041-1480-5-43

Vernikos, G. S. (2020). A Review of Pangenome Tools and Recent Studies. *The Pangenome: Diversity, Dynamics and Evolution of Genomes*, 89–112. https://doi.org/10.1007/978-3-030-38281-0_4

Wang, G., Zhao, G., Chao, X., Xie, L., & Wang, H. (2020). The Characteristic of Virulence, Biofilm and Antibiotic Resistance of Klebsiella pneumoniae. *International Journal of Environmental Research and Public Health*, *17*(17), 1–17. https://doi.org/10.3390/IJERPH17176278

Weber, L. M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P. P., Boulesteix, A. L., Saeys, Y., & Robinson, M. D. (2019). Essential guidelines for computational method benchmarking. In *Genome Biology* (Vol. 20, Issue 1, pp. 1–12). BioMed Central Ltd. https://doi.org/10.1186/s13059-019-1738-8

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*, 1338. https://doi.org/10.12688/F1000RESEARCH.15931.2

Yoshimura, D., Kajitani, R., Gotoh, Y., Katahira, K., Okuno, M., Ogura, Y., Hayashi, T., & Itoh, T. (2019). Evaluation of SNP calling methods for closely related bacterial isolates and a novel high-accuracy pipeline: BactSNP. *Microbial Genomics*, *5*(5). https://doi.org/10.1099/mgen.0.000261

Youenou, B., Brothier, E., & Nazaret, S. (2014). Diversity among strains of Pseudomonas aeruginosa from manure and soil, evaluated by multiple locus variable number tandem repeat analysis and antibiotic resistance profiles. *Research in Microbiology*, *165*(1), 2–13. https://doi.org/10.1016/j.resmic.2013.10.004

Zheng, S. (2017). Bogaerts Contexts and details matter. In *Genome Biology* (Vol. 18, Issue 1, p. 129). BioMed Central Ltd. https://doi.org/10.1186/s13059-017-1258-3

Zhi, X. Y., Zhao, W., Li, W. J., & Zhao, G. P. (2011). Prokaryotic systematics in the genomics era. *Antonie van Leeuwenhoek 2011 101:1*, *101*(1), 21–34. https://doi.org/10.1007/S10482-011-9667-X

Zhou, G. H., Gotou, M., Kajiyama, T., & Kambara, H. (2005). Multiplex SNP typing by bioluminometric assay coupled with terminator incorporation (BATI). *Nucleic Acids Research*, *33*(15), 1–11. https://doi.org/10.1093/nar/gni132

Zhou, K., Lokate, M., Deurenberg, R. H., Tepper, M., Arends, J. P., Raangs, E. G. C., Lo-

Ten-Foe, J., Grundmann, H., Rossen, J. W. A., & Friedrich, A. W. (2016). Whole genome sequencing for the molecular characterization of carbapenem-resistant Klebsiella pneumoniae strains isolated at the Italian ASST Fatebenefratelli Sacco Hospital, 2012-2014Use of whole-genome sequencing to trace, control and characterize the r. *Scientific Reports*, *6*, 20840. https://doi.org/10.1038/srep20840

Zhou, Z., Alikhan, N. F., Sergeant, M. J., Luhmann, N., Vaz, C., Francisco, A. P., Carriço, J. A., & Achtman, M. (2018). GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Research*, *28*(9), 1395–1404. https://doi.org/10.1101/GR.232397.117