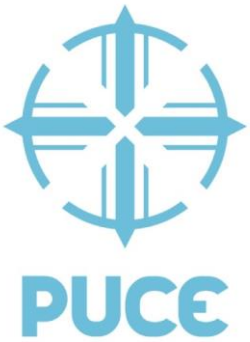
PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERIA



PROYECTO DE GRADO:

PANGENOMA DE CEPAS DE *Pseudomonas aeruginosa* PRESENTES EN HOSPITALES DE QUITO, ECUADOR.

PROYECTO DE GRADO PREVIO A LA OBTENCIÓN DEL TITULO DE

MASTER EN BIOLOGÍA COMPUTACIONAL

AUTORA: MORALES CAIZA GABRIELA NATHALY

TUTOR: DR. SERGIO ALAN CERVANTES

QUITO, 2024

Derechos de autor

La responsabilidad, derecho de la investigación, resultados, conclusiones y recomendaciones que aparecen en el presente Proyecto de grado, corresponden exclusivamente al Autor y los derechos académicos otorgados a la Pontificia Universidad Católica del Ecuador.

Ing. Gabriela Nathaly Morales Caiza

CI: 1722708391

Aprobación del director del Trabajo de Titulación

Hoja de evidencia anti plagio (Informe Turnitin)

DEDICATORIA

A mis hijos, Martin y Emilio, que son

quienes me inspiran a cumplir mis sueños.

AGRADECIMIENTOS

A mi esposo Darwin por su amor y paciencia en este camino, también por brindarme sus conocimientos para verme crecer, a mis hijos por su comprensión en el tiempo que no pude estar con ellos mientras estudiaba, a mi madre por su apoyo incondicional, siendo la segunda madre de mis hijos mientras yo cumplo mis sueños, a mi abuelita que los fines de semana de estudio nos tenía la comida lista demostrando así ese amor inmenso que siempre me ha tenido, a los Hidalgo y Flores, que siempre estuvieron prestos apoyarnos con los niños.

A mi tutor, Alan Cervantes, con sus conocimientos, amplia experiencia y amabilidad supo guiarme en este proyecto de grado. Al grupo de amigos de la maestría que fueron excelentes tanto en lo académico como en la parte humana.

Índice general

Derechos de autor 2
Aprobación del director del Trabajo de Titulación 3
Hoja de evidencia anti plagio (Informe Turnitin) 4
Dedicatoria 5
Agradecimientos 6
Índice general 7
Índice de Figuras 9
Índice de Tablas 10
Índice de Anexos 11
RESUMEN Y ABSTRACT 12
INTRODUCCIÓN 14
Justificación 14
Objetivos 16
MARCO TEÓRICO 17
Genoma de referencia 17
Pangenoma 18
Pankmers 20
Anvi'o 22
Pseudomonas aeruginosa 22
METODOLOGÍA 25
Obtención de secuencias 25
Evaluación del pangenoma 25
Analisis de similitudes entre las secuencias 25

Analisis de variabilidad genética 26
RESULTADOS 27
Calidad del pangenoma 27
Similitudes entre las secuencias 29
Variabilidad genética 31
DISCUSIÓN Y CONCLUSIONES 37
BIBLIOGRAFIA 39
ANEXOS 42

Índice de Figuras

Figura 1. Resistencia a antibioticos de *P. aeruginosa* en Ecuador 2014 a 2017. 15

Figura 2. Análisis de frecuencia alélica 19

Figura 3. Formación de biofilm de *P. aeruginosa* 24

Figura 4. Integridad del pangenoma 27

Figura 5. Estadísticas del pangenoma de *P. aeruginosa* 28

Figura 6. Comparación entre genomas 29

Figura 7. Presencia/ausencia de grupos de genes en el pangenoma *P. aeruginosa* 32

Figura 8. Grupos de genes únicos 33

Figura 9. Pangenoma P. aeruginosa - Grupo 11. 34

Figura 10. Alineamiento de la región GC_00012810 - Grupo 11 35

Figura 11. Alineamiento de la región GC_00001960 - Grupo 11 35

Figura 12. Pangenoma de *P. aeruginosa* – Grupo 8 36

Índice de Tablas

Tabla 1. Métodos de construcción de pangenoma 20

Tabla 2. Equivalencia de nucleótidos en enteros de 2-bits 21

Tabla 3. Factores de patogenicidad de *Pseudomonas aeruginosa* 23

Tabla 4. Grupos de genomas asignados en el análisis de similitud 30

Índice de Anexos

Anexo 1. Metadatos de las secuencias del pangenoma de *P. aeruginosa*..... 42

Anexo 2. Código usado para la obtención del pangenoma de *P. aeruginosa* 49

RESUMEN

Durante las últimas décadas en los estudios genómicos se ha usado el genoma de referencia como base de comparación, principalmente por el alto costo de secuenciación. Sin embargo, este genoma tiene ciertas limitaciones, al no ser el más común entre la especie y provenir de individuos de una ubicación geográfica especifica, no representa la diversidad genética real y esto conlleva a una interpretación imprecisa de las variantes genéticas. Con el fin de solventar estas limitaciones se han desarrollado nuevas estrategias y una de ellas es el desarrollo de los pangenomas. El pangenoma es el conjunto de múltiples genomas donde se presenta un genoma central constituido por los genes consenso entre todas las secuencias y un genoma no esencial constituido por las secuencias únicas de cada individuo. Especies bacterianas como *Pseudomonas aeruginosa* poseen mecanismos de resistencia mediados por modificaciones genéticas, haciendo que estos patógenos sean altamente infecciosos y por lo tanto constituyen un problema de salud pública. Conocer a fondo estas modificaciones ayuda al desarrollo de nuevas dianas terapéuticas. En este proyecto se obtuvo el pangenoma de 42 secuencias de *P. aeruginosa*, extraídas de NCBI (Acceso: PRJNA946810), se usaron dos programas bioinformaticos, Pankmers y Anvi'o, ambos basados en Python.

ABSTRACT

During the last decades, genetic and genomic studies have used the reference genome as a basis for comparison mainly because of the cost of sequencing. However, the reference genome has certain limitations, since it is not the most common among the species and comes from individuals from a specific geographic location, it does not represent the genetic diversity, and this leads to an imprecise interpretation of the genetic variants. To solve these limitations, new strategies have been developed, and the pangenome is one of them. The pangenome is the integration of multiple genomes where there is a core genome consisting of the consensus genes between all sequences and a non-essential genome consisting of the unique sequences for all genomes. Bacterial species such as *Pseudomonas aeruginosa* have resistance mechanisms mediated by genetic modifications, making these pathogens highly infectious and therefore a public health problem. Further understanding of these modifications will help the development of new therapeutic targets. In this project, we obtained the pangenome from 42 sequences of *P. aeruginosa*, extracted from NCBI (Access: PRJNA946810), using two bioinformatics programs, Pankmers and Anvi'o, both based on Python.

INTRODUCCIÓN

Justificación

A nivel general, los estándares se usan para tener un marco de referencia común. Este marco de referencia, que en este caso es el genoma, se ha utilizado durante las últimas décadas en los análisis genéticos y genómicos como una base para comparación y/o punto de partida, y de este modo se han convertido en un recurso para la obtención de grandes cantidades de datos genómicos. Estos genomas de referencia suelen ser obtenidos de un muestreo empírico o de un solo individuo de la especie a estudiar. (Ballouz, 2019)

Por otro lado, el pangenoma es la unión de todos los genomas de una especie o de una población, abarcando mucha más información que un genoma de referencia, donde se tienen secuencias centrales que son aquellas compartidas por todos los individuos, y secuencias accesorias que son las que se encuentran específicamente solo en algunos individuos, estas últimas no suelen estar presentes en un genoma de referencia. Analizar el pangenoma nos da ventajas como el mejoramiento en el llamado de variantes e identificación de genes asociados a fenotipos claves. Además, nos permite ver diferencias y similitudes derivadas de la historia evolutiva de las especies vista desde el punto de vista de genética de poblaciones (Golicz, 2020)

Las infecciones bacterianas afectan a toda la población mundial, por lo que constituyen un problema de salud pública. En el caso de *P. aeruginosa*, esta produce infecciones crónicas a nivel hospitalario debido a la asociación que hace con otros microorganismos y por su alta resistencia a antibióticos (Cendra, 2021)

En América latina la mortalidad por causa de infecciones intrahospitalarias con *P. aeruginosa* es de 33 a 61%, donde todas las cepas son resistentes a antimicrobianos (Morales & Andrade, 2006). La resistencia a antibióticos se da por la afectación de mecanismos como: producción de carbapenemasas, alteración de porinas y bombas de eflujo (Hernández, 2018). Estos mecanismos estan regulados genéticamente, por lo tanto conocer la diversidad genetica de estas bacterias es de mucha importancia para estudio de la resistencia a antibióticos.

En estudios realizados por el Instituto de Investigación en Salud Pública (INSPI) entre 2014 y 2017 se evidencia la resistencia a antibióticos de P. aeruginosa en ambientes hospitalarios, con porcentajes entre 25 a 50% (INSPI, 2019)

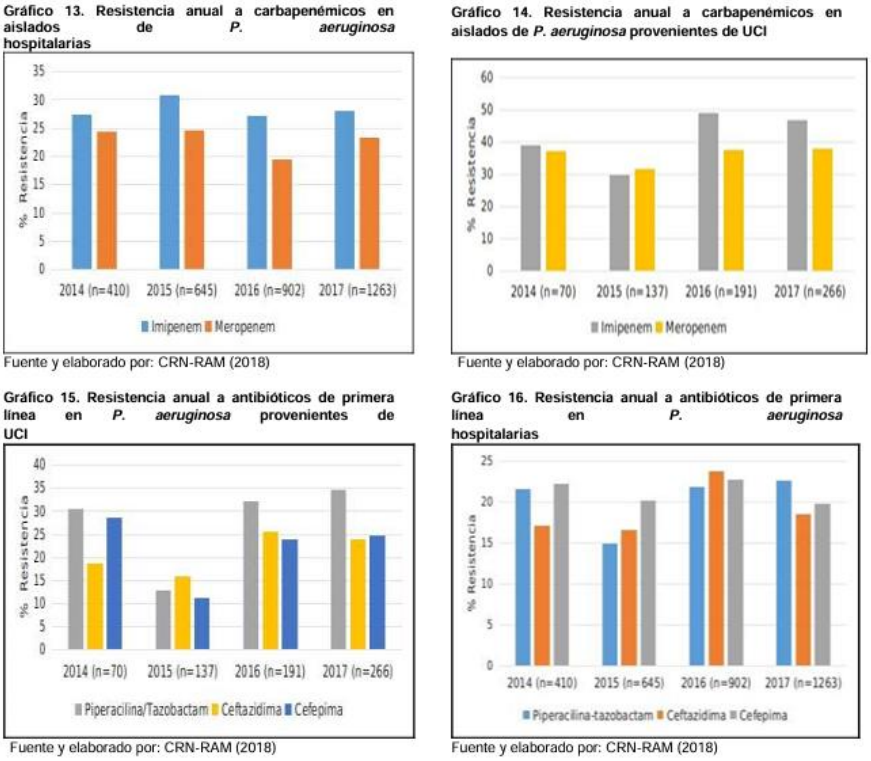


Figura 1. Resistencia a antibioticos de *P. aeruginosa* en Ecuador 2014 a 2017 reportado por el Instituto Nacional de Investigación en Salud Pública del Ecuador. Resistencias anuales a carbapenémicos y antibióticos de primera linea tanto en áreas hospitalarías generales y en UCI. (INSPI, 2019)

Dada la problemática que *P. aeruginosa* causa a nivel de salud pública, es importante obtener la mayor cantidad posible de información sobre la genética de esta baceteria con el fin de avanzar en la obtención de tratamientos específicos y eficaces. Por ello, en este proyecto buscó ampliar el conocimiento genético de las cepas de *Pseudomona aeruginosa* presentes en hospitales del Ecuador, para ellos se realizó un pangenoma de *Pseudomonas aeruginosa* con el paquete Pankmers de Python que usa un método de descomposición de k-mers de 31 pares de bases y el programa Anvi'o que usa descomposición en contigs de 20.000 pares de bases y aplica modelos ocultos de Markov.

Objetivos

Objetivo general

- Obtener el pangenoma de diferentes cepas de *Pseudomonas aeruginosa*, presentes en hospitales de Quito, Ecuador

Objetivos específicos

- Evaluar el pangenoma de cepas de *Pseudomonas aeruginosa*, presentes en hospitales de Quito, Ecuador
- Definir las similitudes entre las secuencias de *Pseudomonas aeruginosa* presentes en hospitales de Quito, Ecuador
- Analizar la variabilidad genética de las cepas de *Pseudomona aeruginosa,* aisladas de pacientes hospitalizados.

MARCO TEÓRICO

Genoma de referencia

Desde la aparición de la secuenciación, la cantidad de datos obtenidos ha ido aumentando exponencialmente y con esto la necesidad de procesar estos datos. Según la base de datos del NCBI (National Center for Biotechnology Information), para 2019 existían $2x10^{16}$ pb en su repositorio, todos estos datos proveniente de proyectos de secuenciación. La base europea Ensembl reporta en su página oficial, que para 2020 soportaba ya más 500.000 genomas y la base de datos genómica de china CNGB para inicios del 2024 reporta que contiene $1,3x10^{16}$ pb. El procesamiento de toda esta data sería dirigido para varios campos de estudio, desde analizar las diferencias entre especies, hasta identificar genes de interés médico. Para estos análisis era necesario tener un modelo del cual partir para hacer la comparación, así nace lo que se conoce como genoma de referencia. Este genoma de referencia presenta ciertas limitaciones, al ser obtenido de uno o unos cuantos individuos de la especie a estudiar no es justamente el más común entre la especie, ni el más sano, por lo que no representa la diversidad, lo que puede llevar una incorrecta interpretación de las variantes genéticas. Los sesgos geográficos son otra limitación ya que los genomas de referencia usualmente provienen de poblaciones presentes en países desarrollados con altos recursos económicos para investigación, como son norte américa o Europa (Popejoy & Fullerton, 2016), por lo que podrían no ser de utilidad para poblaciones de especies estudiadas en países menos desarrollados. Las secuencias repetitivas o las áreas con contenido elevado de GC a menudo no se representan de manera adecuada en los genomas de referencia, lo que podría impactar la interpretación de dichas regiones en investigaciones genómicas (Li,

2010). Los genomas de referencia son estáticos, a diferencia de los genomas individuales que pueden experimentar cambios genómicos dinámicos, como mutaciones, recombinación y modificaciones epigenéticas. Esto puede resultar en una representación insuficiente de la variabilidad genómica a lo largo del tiempo (Eichler & Nickerson, 2019).

A medida que avanza el conocimiento genómico, se va haciendo obsoletos los genomas de referencia y sumado a las demás limitaciones que presenta, los investigadores buscan constantemente estrategias alternativas y una de estas es el pangenoma. (Telenti, 2016) Si bien el genoma de referencia ha servido para obtener mucha información, también ha mostrado varias limitaciones, por los sesgos que este presenta.

Pangenoma

El concepto de pangenoma aparece en 2005 cuando Teteelin y su equipo hicieron un análisis comparativo de los genomas de ocho cepas de bacterias que estaban estudiando con el fin de poder describir la especie bacteriana por completo. El pangenoma es el conjunto de múltiples genomas de la misma especie o población, donde encontramos un genoma central que contiene secuencias consenso entre todos los genomas analizados y un genoma no esencial donde está contenida la secuencia única de cada individuo. Dentro del genoma central se encuentran la mayoría de los genes constitutivos, mientras que en el genoma no esencial se puede observar genes no esenciales, que son genes únicos o faltantes en cada cepa. Son justamente estos genes no esenciales los que le confieren la diversidad genética a las bacterias porque muchos de estos participan en funciones y vías bioquímicas que les confieren ventajas selectivas como adaptación a

nuevos ecosistemas, resistencia a antibióticos o colonización de un nuevo huésped (Gong, Li, Liu, Ma, & Jiang, 2023).

En la figura 2 se muestra una comparación entre estudios con secuencia de referencia y con pangenoma (consenso), donde se observa que se obtiene mayor eficiencia con la aplicación de pangenoma (Ballouz, 2019).

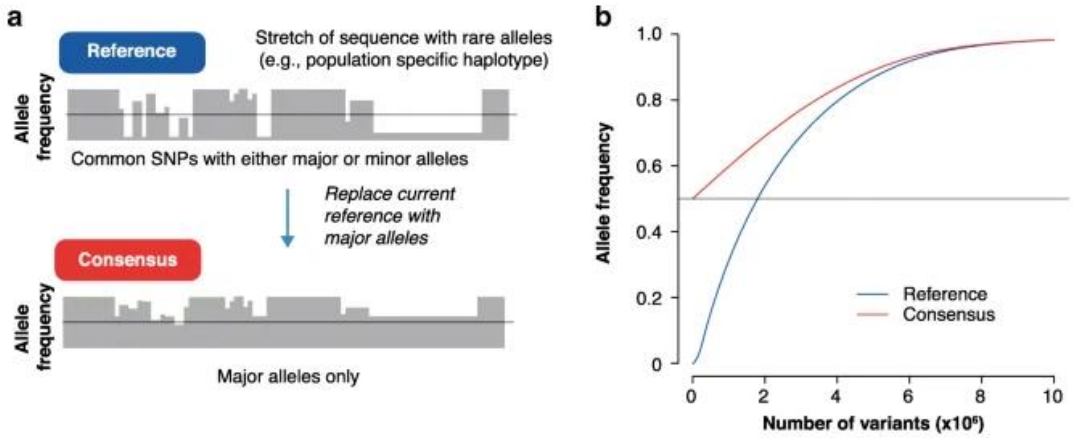


Figura 2. Análisis de frecuencia alélica. *a)* genoma de referencia (azul) y genoma consenso o pangenoma (rojo). b) frecuencia alélica en función del número de variantes obtenidas con el genoma de referencia (azul) y con el genoma consenso o pangenoma (rojo) **(Ballouz, 2019)**

Los métodos mayormente conocidos para realizar un pangenoma son: mapeo y ensamblaje iterativo, mapeo panorámico, ensamblaje de novo, pangenoma gráfico. Sus características se observan en la tabla 1, que como se observa aun presentan ciertas limitaciones, por lo que continuamente se desarrollan nuevos métodos, con el fin de seguir obteniendo la mayor cantidad de información genética de las especies o poblaciones estudiadas.

Tabla 1. Métodos de construcción de pangenoma

Característica	Mapeo y ensamblaje iterativo.	Mapeo panorámico	Ensamble de novo	Pangenoma gráfico
Referencia	Basado en referencias	Basado en referencias	Sin referencia	Sin referencia
Nivel de montaje	Lecturas cortas	contigs	Genomas	Genomas
Enfoque de alineación	Alineación iterativa	Alineación de todo el genoma		
Requisitos de profundidad de secuenciación	Baja profundidad	Baja profundidad	Alta profundidad	Alta profundidad
Ámbito de aplicación	Grupo a gran escala	Grupo a gran escala	Pocos individuos	Pocos individuos
Almacenamiento de secuencia	Transatlántico	Transatlántico	Transatlántico	Entidad gráfica
Requisitos de recursos informáticos	Bajo	Los genomas grandes tienen una gran demanda	Más alto	Más alto
Análisis de SV	Sólo se pueden identificar SNPS e indels	CNV interindividuales	SV grandes	Algunos SV altamente divergentes
Costo	Bajo	Medio	Más alto	Más alto
Limitación	El fragmento de secuencia es demasiado corto para analizar CNV	No apto para análisis a escala de genomas complejos.	Es necesario anotar el genoma, lo que daría lugar a sesgos.	Visualizar el almacenamiento es difícil y puede surgir ambigüedad durante la alineación.

Fuente: (Gong, Li, Liu, Ma, & Jiang, 2023)

Así el pangenoma se ha convertido también en una herramienta para explorar componentes genéticos faltantes y las grandes variantes estructurales, como son variaciones de número de copias, reordenamientos y variantes de presencia/ausencia. Estas variantes estructurales son importantes porque pueden afectar la dosis del gen o interferir con su función (Gong, Li, Liu, Ma, & Jiang, 2023).

Pankmers

Pankmer es un programa para análisis de variación genética de múltiples datos pangenómicos. Usa un método de descomposición de kmers por lo que no hace uso de

métodos gráficos, no requiere genoma de referencia y no asume anotaciones. La representación del pangenoma se lo hace por medio de una indexación de k-mers, donde todos los genomas de entrada conforman un set G, la descomposición de cada genoma en k-mers canónicos únicos es el set K, así para cada genoma de entrada se genera un vector de bits donde se muestra que k-mer está presente o ausente. Para la indexación se considera la secuencia complementaria de cada k-mer. Los k-mer tanto de la secuencia original como de la complementaria son representados como enteros de 64 bits sin signo. Para el cálculo de los enteros se usa los valores binarios que se muestran en la Tabla 2. Se selecciona como el k-mer canónico para la indexación al de mayor valor calculado en la transformación de binario a entero base 10.

Tabla 2. Equivalencia de nucleótidos en enteros de 2-bits

Base	Entero de 2-bit
A	11
C	10
G	01
T	00

Fuente: Aylward, et al, 2023

El valor de k se ha definido en 31 por tres razones:

1. El tamaño permite su codificación en enteros de 64 bits
2. El largo permite tener una baja tasa de k-mers que ocurren por casualidad
3. En estudios previos este número de k ha dado buenos resultados para análisis de variabilidad.

Por medio de Pankmer se obtiene matrices de adyacencia con la cual se puede genera un mapa de calor para observar la agrupación jerárquica. Gracias a que no utiliza genoma de referencia, también se puede analizar el anclaje del genoma, que consiste ubicar cada

k-mer en una posición de cada genoma del set G, para cuantificar la variación en los locus, obteniendo así valores altos que corresponden a los sitios conservados, que son los loci que están presentes en todas las secuencias del set G; y valores bajos que indican los sitios variables, es decir los loci que solo están presentes en solo uno o muy pocos de los genomas del grupo G. (Aylward, Petrus, Mamerto, Hartwick, & Michael, 2023)

Anvi'o

Es una plataforma para el análisis de datos ómicos. Está basado en Python y JavaScript, por lo que la programación se hace por medio de código fuente y la visualización por medio de una interfaz gráfica la cual está conectada a un servidor local. Para el análisis pangenómico se realizan 3 procesos: el primero, generar una base de datos donde se encuentran secuencias de ADN y aminoácidos; segundo, calcular el pangenoma a partir de la base de datos creada anteriormente, para este cálculo se identifican grupos de genes, y tercero, visualización del genoma en la interfaz gráfica interactiva (Delmont & Eren , 2018)

Pseudomonas aeruginosa

Pseudomonas aeruginosa es una bacteria Gram negativa, es un patógeno oportunista, presente alrededor del mundo, ya sea en medios terrestres o acuáticos, crece a temperaturas entre los 20 a 43 grados centígrados. Esta alta resistencia ambiental lo ha convertido en uno de los principales patógenos causantes de infecciones generales, sobre todo de neumonía nosocomial, infecciones urinarias, quirúrgicas y sepsis. Los

factores de patogenicidad que la hacen tan infecciosa son genes o proteínas que confieren características particulares a estructuras celulares como: flagelo, pili y proteínas de membrana, también los asociados a formación de biopelícula, como son: sistema de secreción de tipo 2, 3 y 5, pigmentos y efectores quorum sensing. (Paz, y otros, 2019)

Tabla 3. Factores de patogenicidad de *Pseudomonas aeruginosa*

Factores		Genes o proteínas involucradas
Asociados a célula bacteriana		
	Flagelo	FliC, FliD
	Pili (Tipo IV)	PilA, PilB, PilT, PilU
	Proteínas de membrana	LecA, LecB, LPS
	Secretados	
Factores asociados a la formación de biopelícula		PelA-PelG, Psl, alginato
	Efectores SST2	LasA, LasB
	Efectores SST3	ExoA, ExoS, ExoT, ExoU, ExoY
	Efectores SST5	EstA
	Pigmentos	Piocianina
	Efectores *quorum sensing*	LasIR, RhlIR, rhlAB, hcnABC, chic

Fuente: Paz, et al, 2019

Algunos de estos genes o proteínas involucrados en la patogenicidad, se pueden modificar (mutar) con el fin de proteger a la bacteria. Dentro de estos mecanismos de resistencia por modificaciones genéticas está la transferencia de plásmidos, transposones, integrones y profagos, que pueden ser de la misma especie o de una distinta. Estas modificaciones contribuyen a la reducción en la captación de medicamentos, modificaciones a las dianas de reconocimiento de antibióticos. También se pueden alterar los niveles de expresión como en el caso de las bombas de eflujo y producción de enzimas que inactivan de antibióticos. (Díaz , Mora, Del Rio, & Vidal, 2022). Esta bacteria también tiene la capacidad de generar resistencia adaptativa, para lo cual genera un biofilm que le protege de estímulos ambientales. Este biofilm ha sido

considerado como la principal causa de resistencia a antibióticos. (Bolivar , Torres , & Sanchez, 2021)

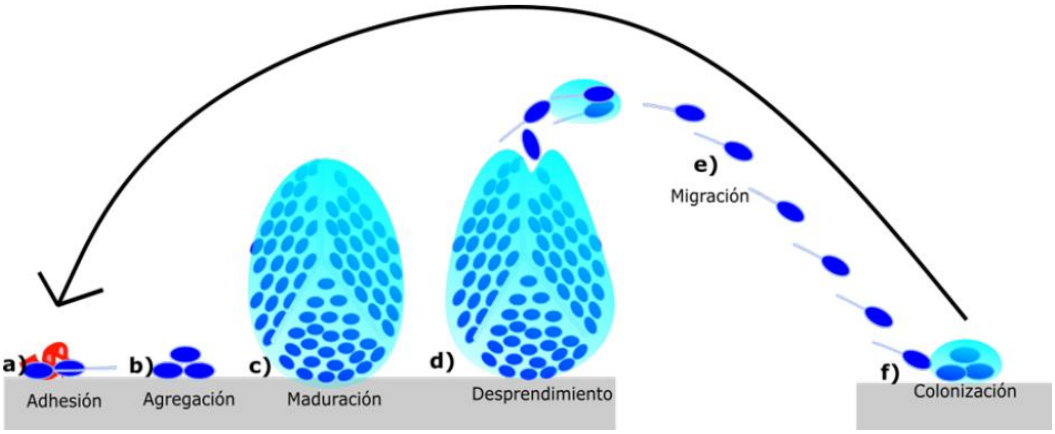


Figura 3. Formación de biofilm de *P. aeruginosa.* La formación del biofilm de *P. aeruginosa* dada por 6 etapas. **(Bolivar , Torres , & Sanchez, 2021)**

En la actualidad se conoce que *P. aeruginosa* tiene la capacidad de generar resistencia a cualquier antibiótico, además de la ya conocida resistencia intrínseca a varias familias de antibióticos. Por estos motivos ha sido catalogada como prioridad tipo 3 para investigación por la OMS en la reunión de expertos del 2017 en Ginebra (Díaz , Mora, Del Rio, & Vidal, 2022)

METODOLOGÍA

Obtención de secuencias

Las secuencias de *P. aeruginosa* con las que se trabajó en este proyecto se encuentran disponibles en NCBI (Acceso: PRJNA946810) donde se encontró 42 ensambles, con sus respectivos datos demográficos de los pacientes de quienes se obtuvieron las muestras.

Evaluación del pangenoma

Utilizando el paquete de PanKmer de Python, se evaluó la integridad del pangenoma generando la curva de acumulación, donde se comparó el número de genomas versus el número de k-mers tanto para el genoma central como para el genoma variable.

Con el programa Anvi'o en su aplicación para pangenómica, se determinó: contenido de GC, "Completion" que es el nivel de terminación el cual se evalúa por el conjunto de contigs que representan un genoma completo en función de la presencia o ausencia de genes centrales de copia única (SCG) y la redundancia que son la cantidad de copias SCG que se encuentra dentro de un genoma.

Analisis de similitudes entre las secuencias

Se generó una matriz de indexación donde se determinó la presencia/ausencia de cada k-mer en cada una de las 42 secuencias en estudio. Esta determinación fue calculada por el algoritmo determinado por el programa Pankmer, el cual divide los genomas en k-mers de 32 nucleotidos, tomando en cuenta la secuencia forware y reverse. Cada k-mer

es representado por su equivalente en enteros de 64-bit y posteriormente transformado a base 10, este último valor es el que determinará cual de las secuencia, forware o reverse, es tomada en cuenta para el análisis. Siendo el valor mas alto el que se toma en cuenta y se mostró en la matriz de indexación. A partir de la matriz se obtuvo un mapa de calor para observar las agrupaciones generadas por la compración entre genomas. Para este gráfico se utilizó la métrica de similitud de Jaccard. El codigo usado se encuentra en el Anexo 2.

Analisis de variabilidad genética

El análisis de variabilidad se realizó con el programa Anvi'o para pangenómica, utilizando el método de Ward con la distancia Euclidiana y ordenando por presencia/ausencia de genes en cada secuencia, visualizado en un círculo filogenético.

Este análisis se lo realizó para todo el conjunto de genomas que forman el pangenoma como para los grupos que se determinaron en el análisis de similitudes, para observar las diferencias entre grupos.

RESULTADOS

Calidad del pangenoma

Para el análisis de la integridad del genoma se observó que el número de k-mers compartidos del genoma central (core) va disminuyendo a medida que se aumenta el número de genomas analizados, mientras que para el pangenoma (pan) sucede lo contrario (Figura 4). Teniendo así que para nuestros 42 genomas de *Pseudomona aeruginosa*, se comparten menos de 0,5x10^7 k-mers y su diversidad es de más de 2x10^7 k-mers.

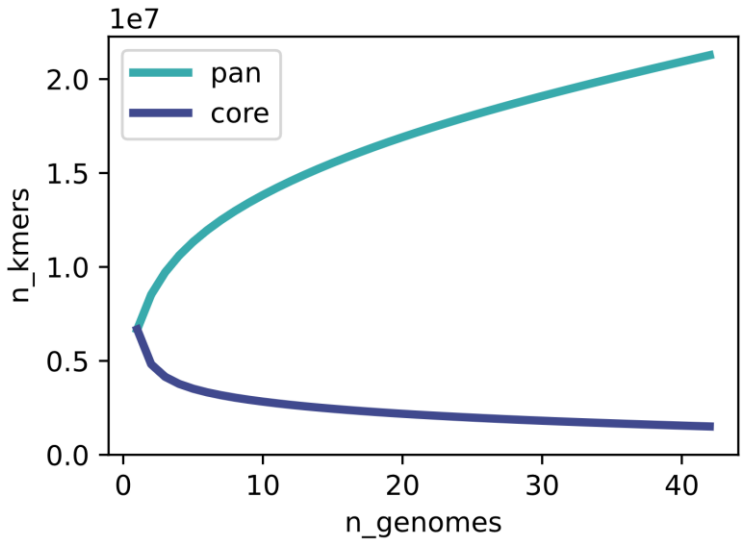


Figura 4. Integridad del pangenoma. La línea celeste corresponde al pangenoma y la línea azul, al genoma central.

En la figura 5 se observa que el contenido de GC de todas las secuencias es aproximadamente del 66%. Para el nivel de terminación “Completion” se observa que en su mayoría las secuencias tienen más del 90% de terminación, excepto dos secuencias (GCA_029382835 y GCA_029382365), que su nivel es 85,9% y al observar la longitud

de esas secuencias, también está por debajo del promedio. Sin embargo, las secuencias contiguas de la derecha tienen aproximadamente la misma longitud, pero un nivel más alto de terminación, lo que se explica en la figura 12. Todas las secuencias presentan redundancia, excepto la GCA_029382725, la mayor redundancia la presenta la secuencia GCA_029382365 con 16,9%. Los valores de cada secuencia se encuentran en el Anexo 1.

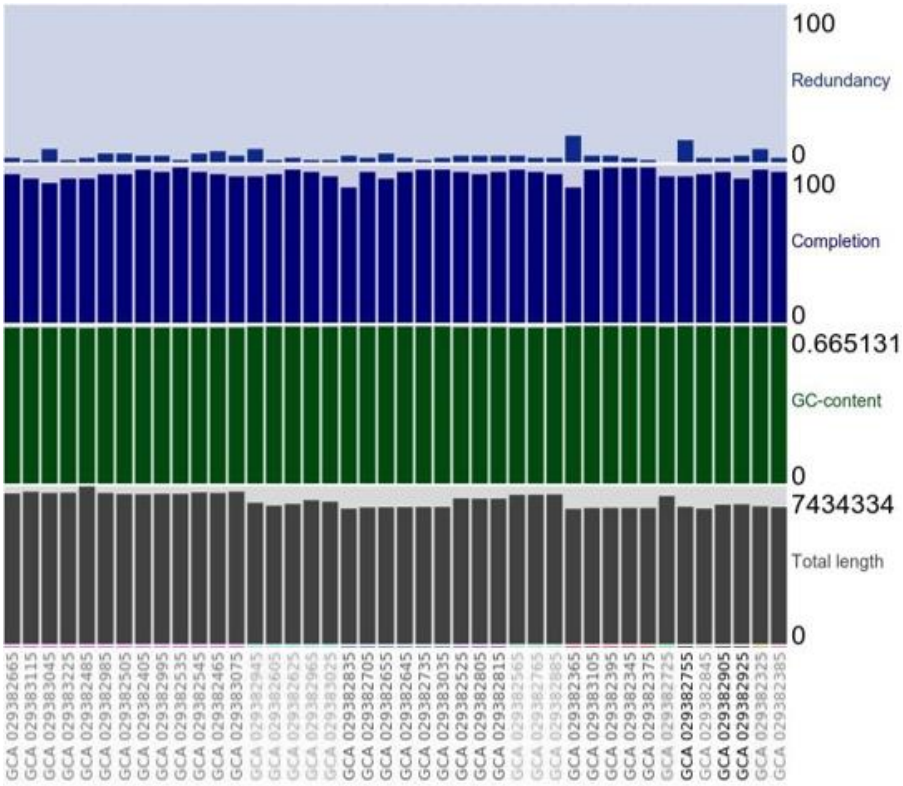


Figura 5. Estadísticas del pangenoma de *P. aeruginosa*. Los parámetros están representados por barras. Total lenght: toma como valor máximo de referencia el de la secuencia más grande. GC-content: toma como valor máximo de referencia el de la secuencia con mayor contenido de GC. Completion: es el nivel de terminación en porcentaje, evaluado en función de la presencia o ausencia de genes centrales de copia única (SCG). Redundancy: cantidad de copias SCG que se encuentra dentro de cada genoma que forma parte del pangenoma.

Similitudes entre las secuencias

En la evaluación de similitudes se aplicó un mapa de calor con el cual se visualizó 11 agrupaciones (Figura 6). En la Tabla 4 se muestra las secuencias según su agrupación del mapa de calor junto con los datos de número de paciente, fuente de aislamiento y genotipificación de la bacteria. Donde se observa que las secuencias de cada grupo guardan relación con respecto al paciente del que provienen, excepto del grupo 11 donde las secuencias son de distintos pacientes, pero su relación es con respecto a la genotipificación ST-253.

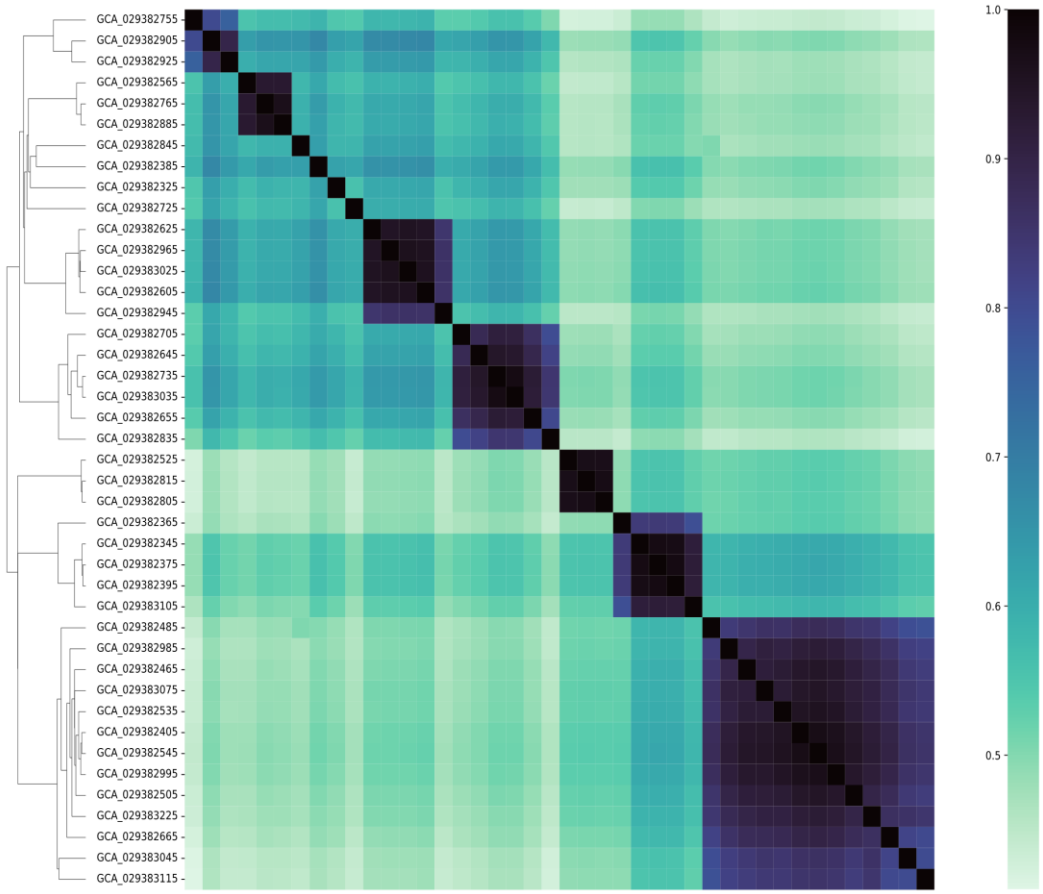


Figura 6. Comparación entre genomas. Los ejes “x” y el eje “y” se usan los genomas en estudio generando una comparación de cada genoma versus cada uno de los genomas del estudio. La correlación entre los genomas está dada de 0 a 1, siendo cero el celeste claro y 1 el negro.

Tabla 4. Grupos de genomas asignados en el análisis de similitud

Grupo	# Assembly	Paciente	Fuente de aislamiento	Genotipo (*P. aeruginosa*)
1	GCA_029382925	6	Hisopado rectal	
	GCA_029382905	6	Secreción traqueal	
	GCA_029382755	6	Hisopado rectal	
2	GCA_029382765	8	Herida	ST-389
	GCA_029382565	8	Hisopado rectal 1	ST-389
	GCA_029382885	8	Hisopado rectal 2	
3	GCA_029382845	2	Hisopado rectal	ST-274
4	GCA_029382385	16	Hisopado rectal	ST-2433
5	GCA_029382325	15	Hisopado rectal	ST-3142
6	GCA_029382725	7	Hisopado rectal	
7	GCA_029382625	5	Hisopado rectal 4	
	GCA_029382965	5	Hisopado rectal 2	
	GCA_029383025	5	Hisopado rectal 3	
	GCA_029382605	5	Esputo	
	GCA_029382945	5	Hisopado rectal 1	
8	GCA_029383035	11	Hisopado rectal 1	
	GCA_029382835	11	Hisopado rectal 3	
	GCA_029382735	11	Hisopado rectal 4	
	GCA_029382705	11	Hisopado rectal 5	
	GCA_029382655	11	Hisopado rectal 2	
	GCA_029382645	11	Orina	
9	GCA_029382815	9	Hisopado rectal 2	
	GCA_029382805	9	Hisopado rectal 1	
	GCA_029382525	9	Secreción traqueal	
10	GCA_029383105	14	Hisopado rectal 2	
	GCA_029382375	14	Fémur	
	GCA_029382365	14	Hisopado rectal 1	
	GCA_029382395	14	Acetábulo	
	GCA_029382345	14	Herida	
11	GCA_029383115	1	Hisopado rectal 1	ST-253
	GCA_029383045	1	Hisopado rectal 2	ST-253
	GCA_029383225	1	Catéter	
	GCA_029382485	3	Secreción traqueal	ST-253
	GCA_029383075	4	Hisopado rectal	ST-253
	GCA_029382985	4	Esputo	ST-253
	GCA_029382505	10	Hisopado rectal	ST-253
	GCA_029382405	10	Secreción traqueal	ST-253
	GCA_029382665	12	Sangre	ST-253
	GCA_029382995	13	Hisopado rectal 2	ST-253
	GCA_029382545	13	Hisopado rectal 1	ST-253
	GCA_029382535	13	Herida	ST-253
	GCA_029382465	13	Orina	ST-253

Variabilidad genética

Para el análisis de la variabilidad con el programa Avi'o, bajo el método de Ward y filtrando por presencia/ausencia de grupos de genes (gene cluster), se agruparon de manera similar al mapa de calor (Figura 6), excepto los grupos 1, 3, 4, 5 y 6. (Figura 7). Los genes que son parte del genoma central (core) están agrupados al inicio del círculo (sentido horario), mientras que los genes únicos de cada secuencia van presentándose a medida que avanza el recorrido del círculo.

La cantidad de grupos de genes por secuencia varia de 5405 a 6204, mientras que los grupos de genes únicos de cada secuencia varían de 6 a 497 grupos. En la figura 8 se observa los bloques de genes únicos de cada secuencia que por su tamaño son los que determinan el tamaño de la barra de "Singleton gene cluster". Se trata de genes únicos ya que al analizar su alineamiento están presentes en solo una secuencia de las del grupo como se observa en la figura 10, mientras que si seleccionamos una región de la parte de los genes del genoma central, tenemos alineamientos para todas las secuencias y para unas más de un alineamiento (figura 11), que son los que contribuyen a la redundancia.

Las secuencias del grupo 11 pertenecen a diferentes pacientes como se parecía en la Tabla 1, sin embargo, están identificadas 12 de las 13 secuencias como genotipo ST-253, por lo que haciendo una comparación solo de este grupo se observa que a pesar de ser de un mismo genotipo identificado tienen variaciones notorias (Figura 9).

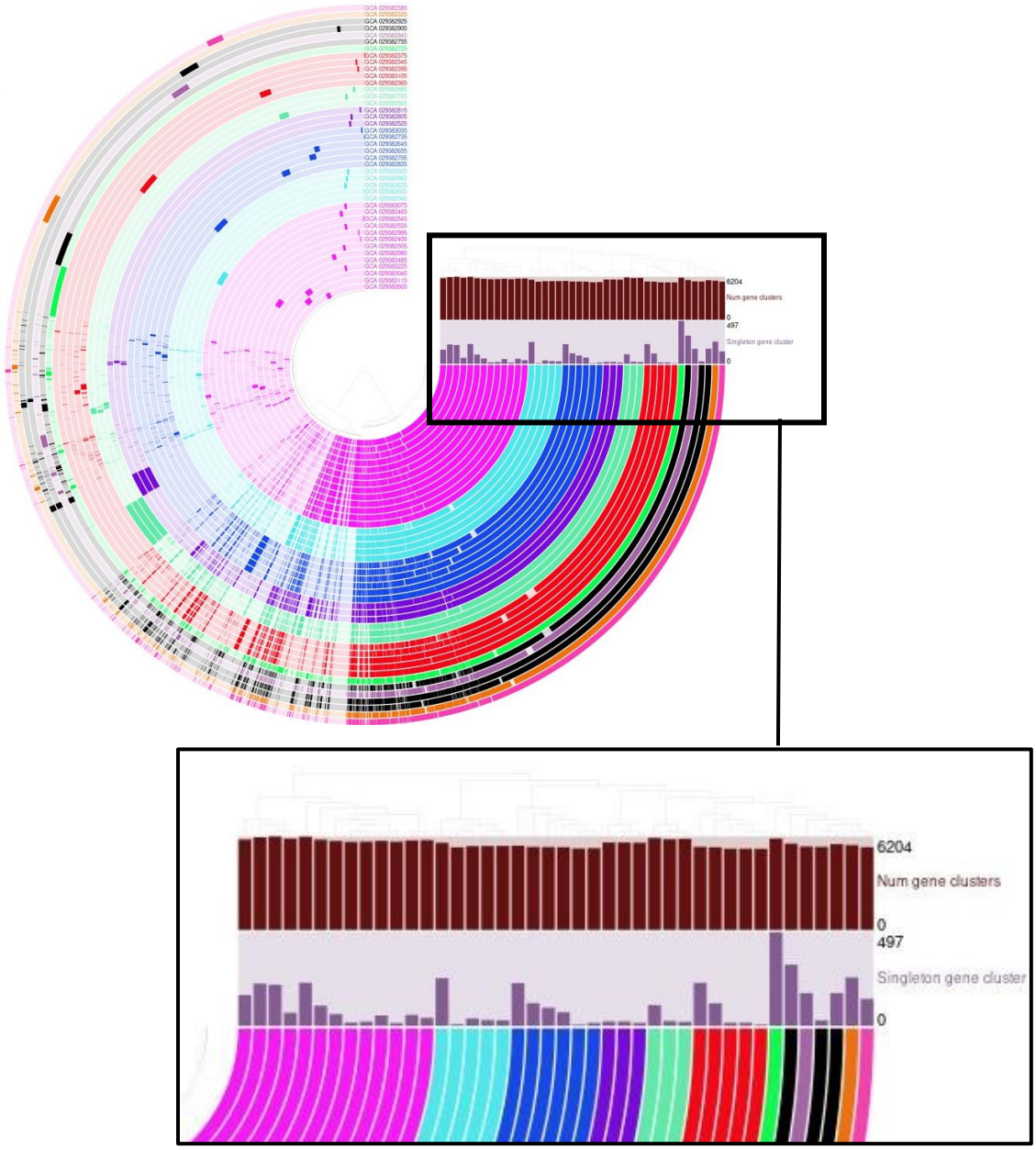


Figura 7. Presencia/ausencia de grupos de genes en el pangenoma *P. aeruginosa*. En los gráficos de barras se observa en marrón “Num gene clusters” que es la cantidad de grupos de genes de cada secuencia, en morado los grupos de genes únicos. Los colores de las secuencias representan los grupos obtenidos en el análisis de similitud. De izquierda a derecha: rosa = grupo 11, celeste = grupo 7, azul = grupo 8, morado oscuro = grupo 9, verde agua = grupo 2, rojo = grupo 10, verde claro = grupo 6, negro = grupo 1, morado claro = grupo 3, naranja = grupo 5 y fuccia = grupo 4.

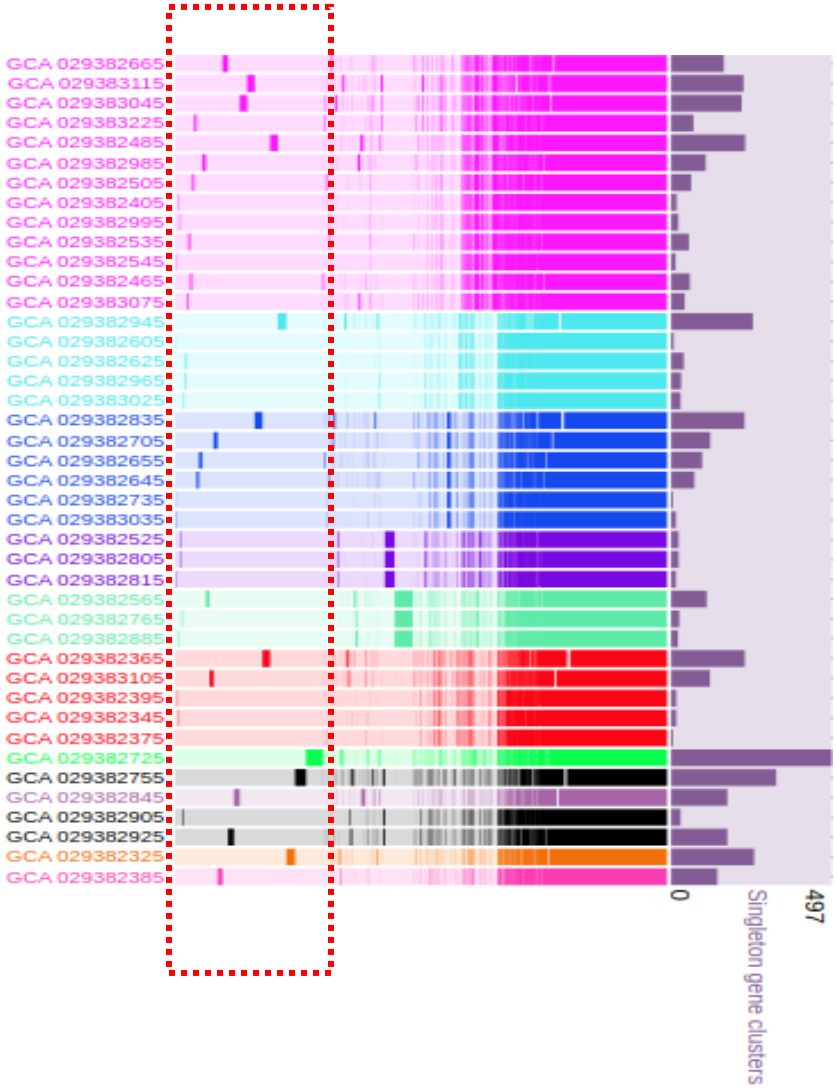


Figura 8. Grupos de genes únicos. En la línea punteada se marcan los grupos más representativos, por tamaño, de los grupos de genes únicos.

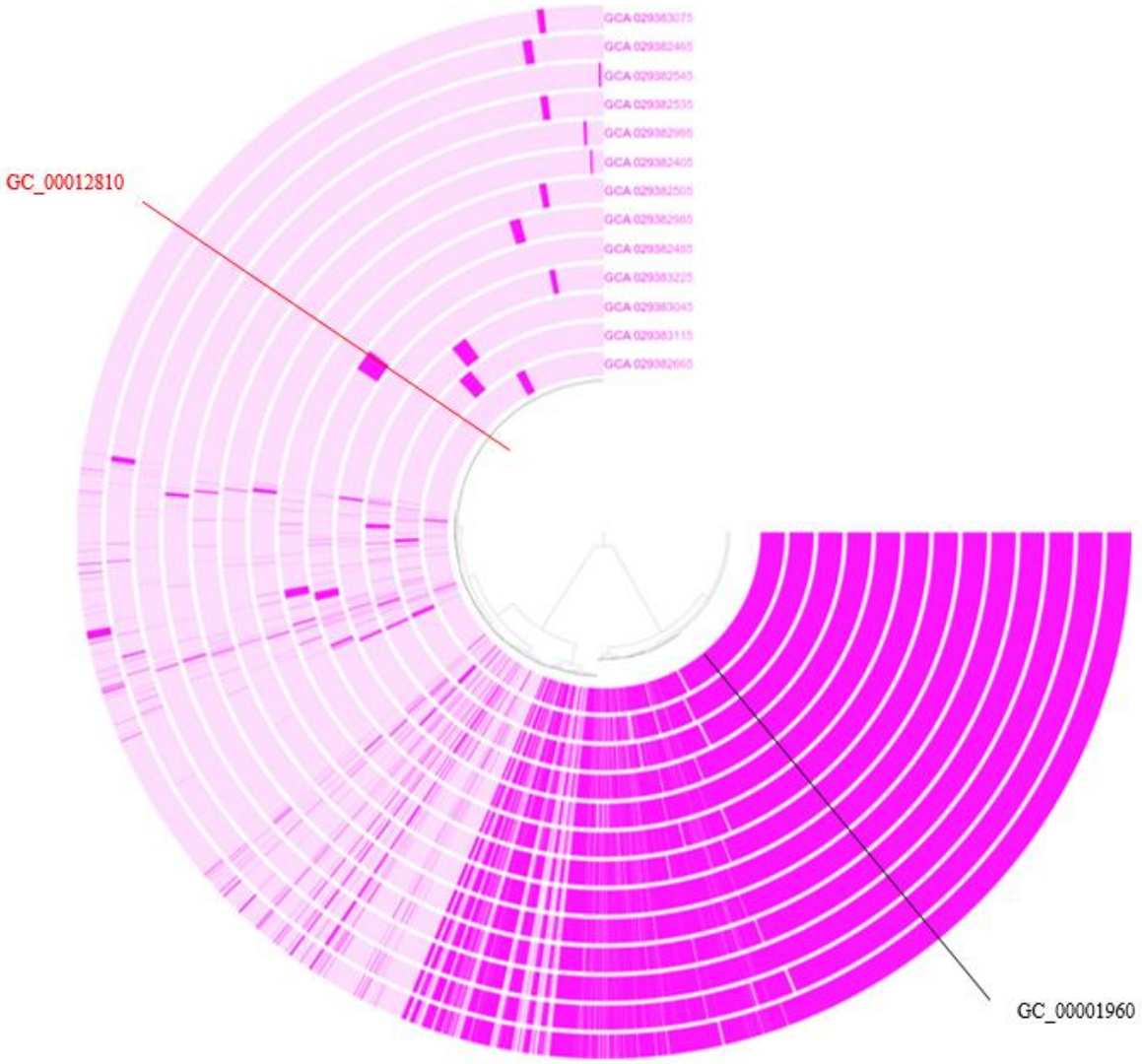


Figura 9. Pangenoma *P. aeruginosa* - Grupo 11. El gráfico presenta las presencias/ausencias de grupos de genes. Al inicio del círculo (sentido horario) se agrupan los genes presentes en todas las secuencias y a medida que avanza el círculo se presentan los genes únicos de cada secuencia. Las regiones GC_00001960 y GC_00012810 marcadas son las mencionadas en las figuras 10 y 11 respectivamente.

Figura 10. Alineamiento de la región GC_00012810 - Grupo 11. En el alineamiento de esta región solo se presenta un gen correspondiente a la secuencia GCA_029382485.

Figura 11. Alineamiento de la región GC_00001960 – Grupo En el alineamiento de esta región se detallan 48 genes, ya que toma en cuenta las alineaciones de genes presentes en todas las secuencias del estudio, sin embargo, en el gráfico están filtradas solo las secuencias del grupo 11. El alineamiento muestra una sección de la región analizada.

En el análisis de las estadísticas del pangenoma se observó un bajo nivel de terminación en dos secuencias en particular, GCA_029382835 y GCA_029382365, que pertenecen a los grupos 8 y 10 respectivamente. Tomando como ejemplo el grupo 8, en la figura 12 se puede observar que el bajo nivel de terminación está relacionado a la ausencia de un grupo de genes centrales.

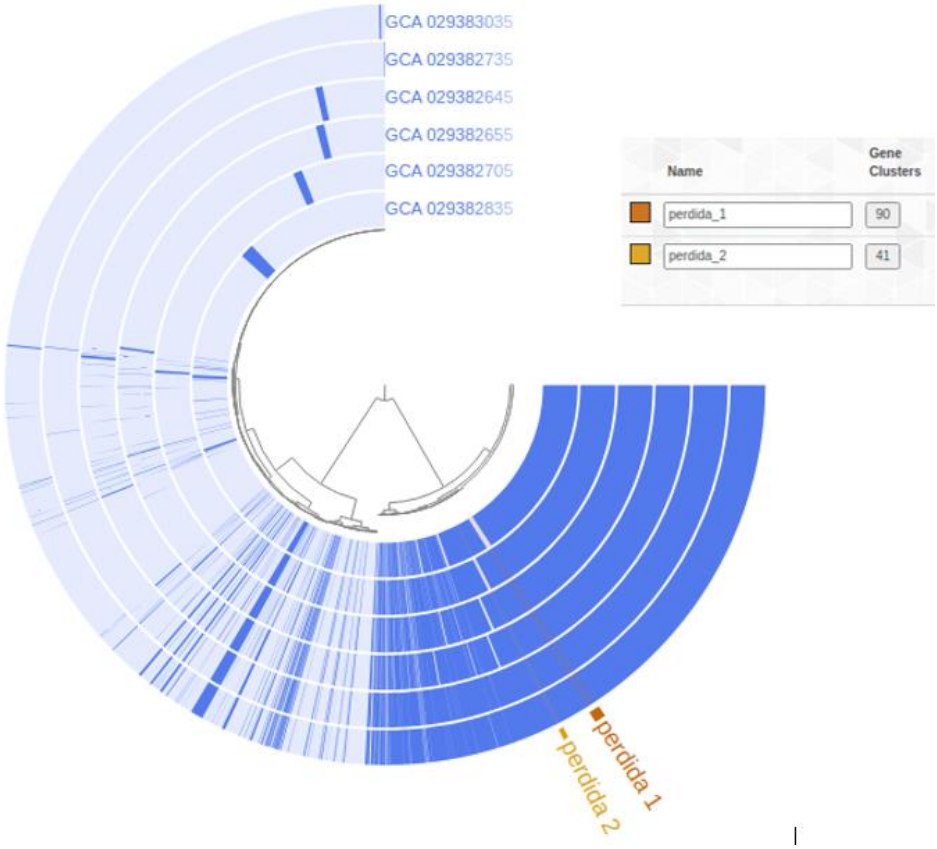


Figura 12. Pangenoma de *P. aeruginosa* – Grupo 8. Las regiones de perdida 1 y 2 corresponden a las secuencias GCA_029382835 y GCA_029382705 respectivamente.

En GCA_029382835 la región con ausencia denominada perdida 1, contiene 90 grupos de genes en las otras secuencias. Se analizó la secuencia contigua con el fin de evidenciar que, si las regiones en el gráfico se ven casi del mismo tamaño, no lo son, tanto así que en dicha secuencia su región de perdida 2 solo contiene 41 grupos de genes.

DISCUSIÓN Y CONCLUSIONES

En 2000 se secuenció por completo la cepa tipo PAO1 de *Pseudomonas aeruginosa*, obteniendo 6,3 millones de pares de bases con un contenido de GC del 66,6% (Stover, 2000). Las secuencias de *P. aeruginosa* usadas en este estudio tienen longitudes de 6,3 a 7,4 millones de pares de bases y contenido de CG del 66, 5%, por lo que consideramos son secuencias representativas para el estudio de un pangenoma.

La redundancia genética es que dos o más genes realizan la misma función, por lo que si uno de estos genes se inactiva puede tener un efecto bajo o nulo en el fenotipo (Nowak, 1997), así la redundancia que presentan las cepas de *P. aeruginosa* estudiadas pueden representar una ventaja para su supervivencia y capacidad infecciosa.

En el pangenoma obtenido, si bien la gran parte de su material genético es bastante homogéneo (genoma central), también tenemos la parte variable con importante número de grupos de genes únicos. Tomando en cuenta que la conservación genómica extendida entre diferentes variantes de *P. aeruginosa*, combinada con la presencia de genes específicos, podría ser la clave detrás de su notable habilidad para colonizar una amplia gama de hábitats y utilizar una diversidad de compuestos como fuente de nutrientes (Soberón, 2024). Se podría tratar de este caso ya que las muestras de donde se aislaron las cepas de *P. aeruginosa* provenían de diferentes ambientes como recto, tráquea, piel (herida), vías respiratorias inferiores (esputo), uretra (orina), fémur, acetábulo, catéter y sangre.

Markussen T. et al, 2014, en su estudio sobre la diversificación y evolución dentro del huésped de P. aeruginosa, observaron la variedad de entornos del huésped da lugar a la evolución y la estabilidad de la población de patógenos durante la infección, por lo que seria recomendable realizar un estudio de este tipo con las secuencias de este proyecto, con el fin de establecer el nivel de diversificación, para generar nuevas dianas terapéuticas.

BIBLIOGRAFIA

Aylward, A., Petrus, S., Mamerto, A., Hartwick, N., & Michael, T. (2023). PanKmer: k-mer based and reference-free pangenome analysis. *Bioinformatics, 39*(10), 1-5.

Ballouz, S. D. (2019). Is it time to change the reference genome? *Genome biology*, 159.

Bolivar , A., Torres , M., & Sanchez, Y. (2021). Biofilms de Pseudomonas aeruginosa como mecanismos. *Revista de la Facultad de Ciencias de la Salud de la Universidad, 23*(2), 47-57. doi:https://doi.org/10.47373/rfcs.2021.v232.1780

Cendra, M. &. (2021). Pseudomonas aeruginosa biofilms and their partners in crime. *Biotechnology Advances*, 49.

Delmont , T., & Eren , A. (2018). Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* , 6:e4320. doi:https://doi.org/10.7717/peerj.4320

Díaz , E., Mora, C., Del Rio, L., & Vidal, p. (2022). Tratamiento de las infecciones graves por Pseudomonas aeruginosa multirresistente. *Medicina Intensiva, 46*(9), 508-520.

Eichler, E., & Nickerson, D. (2019). The continuing integration of human genetics and pathobiology. *Neuron*, 377-388.

Golicz, A. B. (2020). Pangenomics Comes of Age: From Bacteria to Plant And Animal Applications. *Trends in Genetics*, 132-145.

Gong, Y., Li, Y., Liu, X., Ma, Y., & Jiang, L. (2023). A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? *Animal Sci Biotechnol*, 73. doi:https://doi.org/10.1186/s40104-023-00860-1

GRC. (2022). *Genome Reference Consortium*. Obtenido de https://www.ncbi.nlm.nih.gov/grc/help/faq/#human-reference-genome-individuals

Hernández, A. (2018). Infecciones nosocomiales por Pseudomonas aeruginosa multiresistente incluido carbapenémicos: factores predictivos y pronósticos. *Esp Quimioter*, 123-130.

INSPI. (2019). *Reporte de datos de resistencia los antimicrobiano en el Ecuador 2014-2018*. Recuperado el 28 de 11 de 2023, de Ministerio de Salud Pública del Ecuador: //efaidnbmnnnibpcajpcglclefindmkaj/https://www.salud.gob.ec/wp-content/uploads/2019/08/gaceta_ram2018.pdf

Li, R. L. (2010). Building the sequence map of the human pan-genome. *Nature Biotechnology*, 57-63.

Markussen, T., Marvig , R., Gómez-Lozano, M., Aanæs, K., Burleigh, A., Høiby NJohansen, H., . . . Jelsbak, L. (2014). Environmental Heterogeneity Drives Within-Host Diversification and Evolution of Pseudomonas aeruginosa. *mBio, 5*(5). doi:https://doi.org/10.1128/mbio.01592-14

Morales, J., & Andrade, J. (2006). Factores asociados a la mortalidad y patrones de susceptibilidad antibiótica en bacteremias por Pseudomonas aeruginosa. *Bol. Med. Hosp. Infant. Mex*, 291-300.

Nowak, M. B. (1997). Evolution of genetic redundancy. *Nature 388*, 167–171. doi:https://doi.org/10.1038/40618

Paz, V., Mangwani, S., Martínez, A., Álvarez, D., Solano, S., & Vázquez, R. (2019). Pseudomonas aeruginosa: patogenicidad y resistencia antimicrobiana en la infección urinaria. *Revista chilena de infectología, 36*(2), 180-189. Obtenido de https://dx.doi.org/10.4067/S0716-10182019000200180

Popejoy, A., & Fullerton, S. (2016). Genomics is failing on diversity. *Nature*, 161-164.

Soberón, G. (2024). *MIcrobios.* Obtenido de http://www.biblioweb.tic.unam.mx/libros/microbios/Cap3/

Stover, C. P. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, 959–964. doi:https://doi.org/10.1038/35023079

Telenti, A. P. (2016). Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences*, 11901-11906.

ANEXOS

Anexo 1. Metadatos de las secuencias del pangenoma de *P. aeruginosa*

Item Name	GCA_029382665
Layer Name	percent_completion
total_length	7099465
gc_content	0.657692938834123
percent_completion	94.3661971830986
percent_redundancy	2.8169014084507
num_genes	7654
avg_gene_length	777.373530180298
num_genes_per_kb	1.07810940683559
singleton_gene_clusters	163
num_gene_clusters	5996

Item Name	GCA_029383115
Layer Name	percent_completion
total_length	7182098
gc_content	0.657583898186853
percent_completion	91.5492957746479
percent_redundancy	1.40845070422535
num_genes	8509
avg_gene_length	669.992478552121
num_genes_per_kb	1.18475130804397
singleton_gene_clusters	224
num_gene_clusters	6146

Item Name	GCA_029383045
Layer Name	percent_completion
total_length	7126159
gc_content	0.65775349665928
percent_completion	88.7323943661972
percent_redundancy	8.45070422535211
num_genes	8444
avg_gene_length	676.049147323543
num_genes_per_kb	1.18493005839471
singleton_gene_clusters	218
num_gene_clusters	6204

Item Name	GCA_029383225
Layer Name	total_length
total_length	7138327
gc_content	0.657493835740503
percent_completion	91.5492957746479
percent_redundancy	1.40845070422535
num_genes	7810
avg_gene_length	772.834955185659
num_genes_per_kb	1.09409389623087
singleton_gene_clusters	69
num_gene_clusters	6060

Item Name	GCA_029382485
Layer Name	total_length
total_length	7434334
gc_content	0.655549508537012
percent_completion	91.5492957746479
percent_redundancy	2.8169014084507
num_genes	7953
avg_gene_length	790.789513391173
num_genes_per_kb	1.06976630320887
singleton_gene_clusters	229
num_gene_clusters	6191

Item Name	GCA_029382985
Layer Name	total_length
total_length	7123642
gc_content	0.657985620276819
percent_completion	94.3661971830986
percent_redundancy	5.63380281690141
num_genes	7705
avg_gene_length	777.802855288774
num_genes_per_kb	1.08160965977796
singleton_gene_clusters	107
num_gene_clusters	5989

Item Name	GCA_029382505
Layer Name	total_length
total_length	7074056
gc_content	0.657947859050027
percent_completion	94.3661971830986
percent_redundancy	5.63380281690141
num_genes	7332
avg_gene_length	828.67062193126
num_genes_per_kb	1.0364633811211
singleton_gene_clusters	62
num_gene_clusters	5904

Item Name	GCA_029382405
Layer Name	total_length
total_length	7064450
gc_content	0.65799743787556
percent_completion	97.1830985915493
percent_redundancy	4.22535211267606
num_genes	7018
avg_gene_length	881.219008264463
num_genes_per_kb	0.993424824296301
singleton_gene_clusters	17
num_gene_clusters	5841

Item Name	GCA_029382995
Layer Name	total_length
total_length	7083703
gc_content	0.657943310158543
percent_completion	95.7746478873239
percent_redundancy	4.22535211267606
num_genes	7059
avg_gene_length	875.612834679133
num_genes_per_kb	0.99651269964311
singleton_gene_clusters	22
num_gene_clusters	5858

Item Name	GCA_029382535
Layer Name	total_length
total_length	7080980
gc_content	0.657803722083666
percent_completion	98.5915492957746
percent_redundancy	1.40845070422535
num_genes	7243
avg_gene_length	848.963551014773
num_genes_per_kb	1.022881013645
singleton_gene_clusters	55
num_gene_clusters	5909

Item Name	GCA_029382545
Layer Name	total_length
total_length	7148091
gc_content	0.657730014908876
percent_completion	95.7746478873239
percent_redundancy	5.63380281690141
num_genes	6987
avg_gene_length	900.121082009446
num_genes_per_kb	0.977463773194829
singleton_gene_clusters	14
num_gene_clusters	5851

Item Name	GCA_029382465
Layer Name	total_length
total_length	7125424
gc_content	0.65788056963347
percent_completion	94.3661971830986
percent_redundancy	7.04225352112676
num_genes	7500
avg_gene_length	812.8752
num_genes_per_kb	1.05256894186227
singleton_gene_clusters	58
num_gene_clusters	5936

Item Name	GCA_029383075
Layer Name	total_length
total_length	7183101
gc_content	0.657679461836886
percent_completion	92.9577464788732
percent_redundancy	4.22535211267606
num_genes	7692
avg_gene_length	798.382995319813
num_genes_per_kb	1.07084669977493
singleton_gene_clusters	42
num_gene_clusters	5956

Item Name	GCA_029382945
Layer Name	total_length
total_length	6662600
gc_content	0.661955993155825
percent_completion	92.9577464788732
percent_redundancy	8.45070422535211
num_genes	7886
avg_gene_length	676.468425057063
num_genes_per_kb	1.18362200942575
singleton_gene_clusters	253
num_gene_clusters	5779

Item Name	GCA_029382605
Layer Name	total_length
total_length	6525922
gc_content	0.662954598599248
percent_completion	94.3661971830986
percent_redundancy	1.40845070422535
num_genes	6522
avg_gene_length	880.010579576817
num_genes_per_kb	0.999399012124264
singleton_gene_clusters	8
num_gene_clusters	5485

Item Name	GCA_029382625
Layer Name	total_length
total_length	6606490
gc_content	0.662110893984552
percent_completion	97.1830985915493
percent_redundancy	2.8169014084507
num_genes	6814
avg_gene_length	841.603023187555
num_genes_per_kb	1.03141002256872
singleton_gene_clusters	39
num_gene_clusters	5561

Item Name	GCA_029382965
Layer Name	total_length
total_length	6783690
gc_content	0.660962396571777
percent_completion	95.7746478873239
percent_redundancy	1.40845070422535
num_genes	7028
avg_gene_length	841.858992601025
num_genes_per_kb	1.03601432258844
singleton_gene_clusters	31
num_gene_clusters	5568

Item Name	GCA_029383025
Layer Name	total_length
total_length	6713166
gc_content	0.661273533233053
percent_completion	92.9577464788732
percent_redundancy	1.40845070422535
num_genes	6894
avg_gene_length	853.402088772846
num_genes_per_kb	1.02693721561481
singleton_gene_clusters	29
num_gene_clusters	5575

Item Name	GCA_029382835
Layer Name	total_length
total_length	6386065
gc_content	0.663697754407448
percent_completion	85.9154929577465
percent_redundancy	4.22535211267606
num_genes	7291
avg_gene_length	707.287340556851
num_genes_per_kb	1.14170463344798
singleton_gene_clusters	227
num_gene_clusters	5590

Item Name	GCA_029382705
Layer Name	total_length
total_length	6442630
gc_content	0.663061358482483
percent_completion	95.7746478873239
percent_redundancy	2.8169014084507
num_genes	6956
avg_gene_length	781.622052903968
num_genes_per_kb	1.07968329703863
singleton_gene_clusters	121
num_gene_clusters	5512

Item Name	GCA_029382655
Layer Name	total_length
total_length	6447555
gc_content	0.663322918532684
percent_completion	91.5492957746479
percent_redundancy	5.63380281690141
num_genes	6956
avg_gene_length	785.282346175963
num_genes_per_kb	1.07885857507226
singleton_gene_clusters	96
num_gene_clusters	5505

Item Name	GCA_029382645
Layer Name	total_length
total_length	6456168
gc_content	0.663192314698131
percent_completion	95.7746478873239
percent_redundancy	2.8169014084507
num_genes	6833
avg_gene_length	810.991804478267
num_genes_per_kb	1.05836775003377
singleton_gene_clusters	72
num_gene_clusters	5498

Item Name	GCA_029382735
Layer Name	total_length
total_length	6463565
gc_content	0.663107279032546
percent_completion	97.1830985915493
percent_redundancy	1.40845070422535
num_genes	6230
avg_gene_length	918.35393258427
num_genes_per_kb	0.963864368966662
singleton_gene_clusters	6
num_gene_clusters	5405

Item Name	GCA_029383035
Layer Name	total_length
total_length	6463491
gc_content	0.663166545756774
percent_completion	97.1830985915493
percent_redundancy	2.8169014084507
num_genes	6331
avg_gene_length	900.163481282578
num_genes_per_kb	0.979501634642951
singleton_gene_clusters	15
num_gene_clusters	5429

Item Name	GCA_029382525
Layer Name	total_length
total_length	6870960
gc_content	0.659831959435072
percent_completion	95.7746478873239
percent_redundancy	4.22535211267606
num_genes	7087
avg_gene_length	839.087343022435
num_genes_per_kb	1.0314424767427
singleton_gene_clusters	23
num_gene_clusters	5800

Item Name	GCA_029382805
Layer Name	total_length
total_length	6855163
gc_content	0.659833763252602
percent_completion	94.3661971830986
percent_redundancy	4.22535211267606
num_genes	7029
avg_gene_length	845.093043107128
num_genes_per_kb	1.02535855091994
singleton_gene_clusters	23
num_gene_clusters	5799

Item Name	GCA_029382815
Layer Name	total_length
total_length	6855258
gc_content	0.659829433115428
percent_completion	95.7746478873239
percent_redundancy	4.22535211267606
num_genes	6995
avg_gene_length	850.153252323088
num_genes_per_kb	1.02038464489593
singleton_gene_clusters	15
num_gene_clusters	5783

Item Name	GCA_029382565
Layer Name	total_length
total_length	7037322
gc_content	0.657643916251097
percent_completion	97.1830985915493
percent_redundancy	4.22535211267606
num_genes	7799
avg_gene_length	759.342736248237
num_genes_per_kb	1.10823406972141
singleton_gene_clusters	110
num_gene_clusters	6103

Item Name	GCA_029382765
Layer Name	total_length
total_length	7041890
gc_content	0.657518507105337
percent_completion	95.7746478873239
percent_redundancy	2.8169014084507
num_genes	6994
avg_gene_length	889.77752359165
num_genes_per_kb	0.993199268946263
singleton_gene_clusters	25
num_gene_clusters	6015

Item Name	GCA_029382885
Layer Name	total_length
total_length	7055341
gc_content	0.657353060610394
percent_completion	94.3661971830986
percent_redundancy	2.8169014084507
num_genes	7394
avg_gene_length	832.491885312415
num_genes_per_kb	1.04800037305071
singleton_gene_clusters	21
num_gene_clusters	6043

Item Name	GCA_029382365
Layer Name	total_length
total_length	6374110
gc_content	0.664811087351803
percent_completion	85.9154929577465
percent_redundancy	16.9014084507042
num_genes	7309
avg_gene_length	702.059105212751
num_genes_per_kb	1.14666988803143
singleton_gene_clusters	228
num_gene_clusters	5517

Item Name	GCA_029383105
Layer Name	total_length
total_length	6417746
gc_content	0.664913039562488
percent_completion	97.1830985915493
percent_redundancy	4.22535211267606
num_genes	6891
avg_gene_length	784.817588158468
num_genes_per_kb	1.07374146624064
singleton_gene_clusters	120
num_gene_clusters	5492

Item Name	GCA_029382395
Layer Name	total_length
total_length	6422496
gc_content	0.664827039207187
percent_completion	98.5915492957746
percent_redundancy	4.22535211267606
num_genes	6263
avg_gene_length	907.084464314226
num_genes_per_kb	0.975166041364603
singleton_gene_clusters	16
num_gene_clusters	5390

Item Name	GCA_029382345
Layer Name	total_length
total_length	6420783
gc_content	0.66483713902183
percent_completion	98.5915492957746
percent_redundancy	2.8169014084507
num_genes	6260
avg_gene_length	907.372044728435
num_genes_per_kb	0.974958973072287
singleton_gene_clusters	16
num_gene_clusters	5391

Item Name	GCA_029382375
Layer Name	total_length
total_length	6421072
gc_content	0.664841166708612
percent_completion	98.5915492957746
percent_redundancy	1.40845070422535
num_genes	6152
avg_gene_length	928.072821846554
num_genes_per_kb	0.958095470662843
singleton_gene_clusters	6
num_gene_clusters	5386

Item Name	GCA_029382725
Layer Name	total_length
total_length	6978603
gc_content	0.660865648898497
percent_completion	92.9577464788732
percent_redundancy	-1
num_genes	7830
avg_gene_length	741.852873563218
num_genes_per_kb	1.12200106525618
singleton_gene_clusters	497
num_gene_clusters	6064

Item Name	GCA_029382755
Layer Name	total_length
total_length	6472127
gc_content	0.665131571120282
percent_completion	92.9577464788732
percent_redundancy	14.0845070422535
num_genes	8112
avg_gene_length	614.677144970414
num_genes_per_kb	1.25337466338346
singleton_gene_clusters	325
num_gene_clusters	5727

Item Name	GCA_029382845
Layer Name	total_length
total_length	6386085
gc_content	0.664530146404252
percent_completion	94.3661971830986
percent_redundancy	2.8169014084507
num_genes	7290
avg_gene_length	722.058436213992
num_genes_per_kb	1.14154446738495
singleton_gene_clusters	174
num_gene_clusters	5535

Item Name	GCA_029382905
Layer Name	total_length
total_length	6572209
gc_content	0.663617362138057
percent_completion	95.7746478873239
percent_redundancy	2.8169014084507
num_genes	6796
avg_gene_length	844.034726309594
num_genes_per_kb	1.03405110823469
singleton_gene_clusters	29
num_gene_clusters	5516

Item Name	GCA_029382925
Layer Name	total_length
total_length	6590279
gc_content	0.663995712472871
percent_completion	91.5492957746479
percent_redundancy	4.22535211267606
num_genes	7629
avg_gene_length	715.823436885568
num_genes_per_kb	1.15761411618537
singleton_gene_clusters	174
num_gene_clusters	5692

Item Name	GCA_029382325
Layer Name	total_length
total_length	6492286
gc_content	0.663741400178612
percent_completion	97.1830985915493
percent_redundancy	8.45070422535211
num_genes	7175
avg_gene_length	760.403205574913
num_genes_per_kb	1.10515772102461
singleton_gene_clusters	257
num_gene_clusters	5628

Item Name	GCA_029382385
Layer Name	total_length
total_length	6453935
gc_content	0.66426776842345
percent_completion	95.7746478873239
percent_redundancy	2.8169014084507
num_genes	6518
avg_gene_length	865.433261736729
num_genes_per_kb	1.00992650220369
singleton_gene_clusters	143
num_gene_clusters	5496

Anexo 2. Código usado para la obtención del pangenoma de *P. aeruginosa*

```
PANKMERS
## ACTIVAR PANKMER
  conda activate
  conda activate pankmer
  pankmer --version

  ## IR A CARPETA DONDE ESTAN LOS GENOMAS A ANALIZAR
  cd /home/gmorales/Documents/pseudomonas_ecu/

 ## GENERAR LAS INDEXACIONES
pankmer index -g pseudomonas_ecu_genomes/ -o pseudomonas_ecu_index.tar
  tar -tvf pseudomonas_ecu_index.tar
pankmer adj-matrix -i pseudomonas_ecu_index.tar -o
pseudomonas_ecu_adj_matrix.csv
  pankmer adj-matrix -i pseudomonas_ecu_index.tar -o
pseudomonas_ecu_adj_matrix.tsv


  ## GRAFICOS
## Newick tree
pankmer tree --newick --metric jaccard -i pseudomonas_ecu_adj_matrix.csv

## count total and diagnostic K-mers (resultado los kmers de cada secuencia)
pankmer count -i pseudomonas_ecu_index.tar

#Calculate collection curves
pankmer collect -i pseudomonas_ecu_index.tar -o pseudomonas_ecu_collect.svg

## Clustermap (similitudes)
 pankmer clustermap -i pseudomonas_ecu_adj_matrix.csv   -o
pseudomonas_ecu_adj_matrix.svg   --metric jaccard   --width 15   --height 15

### UPSET PLOT
#GRUPO1
pankmer upset -i pseudomonas_ecu_index.tar -o pseudomonas_ecu_upsetG1.svg -g
GCA_029382925 GCA_029382905 GCA_029382755

#GRUPO2
pankmer upset -i pseudomonas_ecu_index.tar -o pseudomonas_ecu_upsetG2.svg -g
GCA_029382765 GCA_029382565 GCA_029382885

## GRUPO 7
```

```
pankmer upset -i pseudomonas_ecu_index.tar -o pseudomonas_ecu_upsetG7.svg -g
GCA_029383025 GCA_029382965 GCA_029382945 GCA_029382625
GCA_029382605

#GRUPO 8
pankmer upset -i pseudomonas_ecu_index.tar -o pseudomonas_ecu_upsetG8.svg -g
GCA_029383035 GCA_029382835 GCA_029382735 GCA_029382705
GCA_029382655 GCA_029382645

#GRUPO 9
pankmer upset -i pseudomonas_ecu_index.tar -o pseudomonas_ecu_upsetG9.svg -g
GCA_029382815 GCA_029382805 GCA_029382525

 #GRUPO 10

pankmer upset -i pseudomonas_ecu_index.tar -o pseudomonas_ecu_upsetG10.svg -g
GCA_029383105 GCA_029382375 GCA_029382365 GCA_029382395.
GCA_029382345

### GRUPO 11

pankmer upset -i pseudomonas_ecu_index.tar -o pseudomonas_ecu_upsetG11.svg -g
GCA_029383115 GCA_029383045 GCA_029382485 GCA_029383075
GCA_029382985 GCA_029382505 GCA_029382405 GCA_029382665
GCA_029382995 GCA_029382545 GCA_029382535 GCA_029382465
GCA_029383225

ANVI'O
  cd ..
  pwd
 mkdir -p ~/miniconda3
cd mkdir -p ~/miniconda3
cd ~/miniconda3
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -O
~/miniconda3/miniconda.sh
 bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3
~/miniconda3/bin/conda init bash
exit
conda update conda
conda activate anvio-8
cd pseudomonas_ecu_genomes/
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382325.fna.gz -o contigs-
fasta-GCA_029382325 --simplify-names --report-file GCA_029382325-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382345.fna.gz -o contigs-
fasta-GCA_029382345 --simplify-names --report-file GCA_029382345-txt
```

```
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382365.fna.gz -o contigs-
fasta-GCA_029382365 --simplify-names --report-file GCA_029382365-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382375.fna.gz -o contigs-
fasta-GCA_029382375 --simplify-names --report-file GCA_029382375-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382385.fna.gz -o contigs-
fasta-GCA_029382385 --simplify-names --report-file GCA_029382385-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382395.fna.gz -o contigs-
fasta-GCA_029382395 --simplify-names --report-file GCA_029382395-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382405.fna.gz -o contigs-
fasta-GCA_029382405 --simplify-names --report-file GCA_029382405-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382465.fna.gz -o contigs-
fasta-GCA_029382465 --simplify-names --report-file GCA_029382465-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382485.fna.gz -o contigs-
fasta-GCA_029382485 --simplify-names --report-file GCA_029382485-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382505.fna.gz -o contigs-
fasta-GCA_029382505 --simplify-names --report-file GCA_029382505-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382525.fna.gz -o contigs-
fasta-GCA_029382525 --simplify-names --report-file GCA_029382525-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382535.fna.gz -o contigs-
fasta-GCA_029382535 --simplify-names --report-file GCA_029382535-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382545.fna.gz -o contigs-
fasta-GCA_029382545 --simplify-names --report-file GCA_029382545-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382565.fna.gz -o contigs-
fasta-GCA_029382565 --simplify-names --report-file GCA_029382565-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382605.fna.gz -o contigs-
fasta-GCA_029382605 --simplify-names --report-file GCA_029382605-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382625.fna.gz -o contigs-
fasta-GCA_029382625 --simplify-names --report-file GCA_029382625-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382645.fna.gz -o contigs-
fasta-GCA_029382645 --simplify-names --report-file GCA_029382645-txt
```

```
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382655.fna.gz -o contigs-
fasta-GCA_029382655 --simplify-names --report-file GCA_029382655-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382665.fna.gz -o contigs-
fasta-GCA_029382665 --simplify-names --report-file GCA_029382665-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382705.fna.gz -o contigs-
fasta-GCA_029382705 --simplify-names --report-file GCA_029382705-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382725.fna.gz -o contigs-
fasta-GCA_029382725 --simplify-names --report-file GCA_029382725-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382735.fna.gz -o contigs-
fasta-GCA_029382735 --simplify-names --report-file GCA_029382735-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382755.fna.gz -o contigs-
fasta-GCA_029382755 --simplify-names --report-file GCA_029382755-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382765.fna.gz -o contigs-
fasta-GCA_029382765 --simplify-names --report-file GCA_029382765-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382805.fna.gz -o contigs-
fasta-GCA_029382805 --simplify-names --report-file GCA_029382805-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382815.fna.gz -o contigs-
fasta-GCA_029382815 --simplify-names --report-file GCA_029382815-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382835.fna.gz -o contigs-
fasta-GCA_029382835 --simplify-names --report-file GCA_029382835-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382845.fna.gz -o contigs-
fasta-GCA_029382845 --simplify-names --report-file GCA_029382845-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382885.fna.gz -o contigs-
fasta-GCA_029382885 --simplify-names --report-file GCA_029382885-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382905.fna.gz -o contigs-
fasta-GCA_029382905 --simplify-names --report-file GCA_029382905-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382925.fna.gz -o contigs-
fasta-GCA_029382925 --simplify-names --report-file GCA_029382925-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382945.fna.gz -o contigs-
fasta-GCA_029382945 --simplify-names --report-file GCA_029382945-txt
```

```
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382965.fna.gz -o contigs-
fasta-GCA_029382965 --simplify-names --report-file GCA_029382965-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382985.fna.gz -o contigs-
fasta-GCA_029382985 --simplify-names --report-file GCA_029382985-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029382995.fna.gz -o contigs-
fasta-GCA_029382995 --simplify-names --report-file GCA_029382995-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029383025.fna.gz -o contigs-
fasta-GCA_029383025 --simplify-names --report-file GCA_029383025-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029383035.fna.gz -o contigs-
fasta-GCA_029383035 --simplify-names --report-file GCA_029383035-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029383045.fna.gz -o contigs-
fasta-GCA_029383045 --simplify-names --report-file GCA_029383045-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029383075.fna.gz -o contigs-
fasta-GCA_029383075 --simplify-names --report-file GCA_029383075-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029383105.fna.gz -o contigs-
fasta-GCA_029383105 --simplify-names --report-file GCA_029383105-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029383115.fna.gz -o contigs-
fasta-GCA_029383115 --simplify-names --report-file GCA_029383115-txt
anvi-script-reformat-fasta
/home/gnmorales/pseudomonas_ecu_genomes/GCA_029383225.fna.gz -o contigs-
fasta-GCA_029383225 --simplify-names --report-file GCA_029383225-txt
anvi-gen-contigs-database -n GCA_029382325 -f contigs-fasta-GCA_029382325 -o
contigs-db-GCA_029382325.db
anvi-gen-contigs-database -n GCA_029382345 -f contigs-fasta-GCA_029382345 -o
contigs-db-GCA_029382345.db
anvi-gen-contigs-database -n GCA_029382365 -f contigs-fasta-GCA_029382365 -o
contigs-db-GCA_029382365.db
anvi-gen-contigs-database -n GCA_029382375 -f contigs-fasta-GCA_029382375 -o
contigs-db-GCA_029382375.db
anvi-gen-contigs-database -n GCA_029382385 -f contigs-fasta-GCA_029382385 -o
contigs-db-GCA_029382385.db
anvi-gen-contigs-database -n GCA_029382395 -f contigs-fasta-GCA_029382395 -o
contigs-db-GCA_029382395.db
anvi-gen-contigs-database -n GCA_029382405 -f contigs-fasta-GCA_029382405 -o
contigs-db-GCA_029382405.db
anvi-gen-contigs-database -n GCA_029382465 -f contigs-fasta-GCA_029382465 -o
contigs-db-GCA_029382465.db
```

```
anvi-gen-contigs-database -n GCA_029382485 -f contigs-fasta-GCA_029382485 -o contigs-db-GCA_029382485.db
anvi-gen-contigs-database -n GCA_029382505 -f contigs-fasta-GCA_029382505 -o contigs-db-GCA_029382505.db
anvi-gen-contigs-database -n GCA_029382525 -f contigs-fasta-GCA_029382525 -o contigs-db-GCA_029382525.db
anvi-gen-contigs-database -n GCA_029382535 -f contigs-fasta-GCA_029382535 -o contigs-db-GCA_029382535.db
anvi-gen-contigs-database -n GCA_029382545 -f contigs-fasta-GCA_029382545 -o contigs-db-GCA_029382545.db
anvi-gen-contigs-database -n GCA_029382565 -f contigs-fasta-GCA_029382565 -o contigs-db-GCA_029382565.db
anvi-gen-contigs-database -n GCA_029382605 -f contigs-fasta-GCA_029382605 -o contigs-db-GCA_029382605.db
anvi-gen-contigs-database -n GCA_029382625 -f contigs-fasta-GCA_029382625 -o contigs-db-GCA_029382625.db
anvi-gen-contigs-database -n GCA_029382645 -f contigs-fasta-GCA_029382645 -o contigs-db-GCA_029382645.db
anvi-gen-contigs-database -n GCA_029382655 -f contigs-fasta-GCA_029382655 -o contigs-db-GCA_029382655.db
anvi-gen-contigs-database -n GCA_029382665 -f contigs-fasta-GCA_029382665 -o contigs-db-GCA_029382665.db
anvi-gen-contigs-database -n GCA_029382705 -f contigs-fasta-GCA_029382705 -o contigs-db-GCA_029382705.db
anvi-gen-contigs-database -n GCA_029382725 -f contigs-fasta-GCA_029382725 -o contigs-db-GCA_029382725.db
anvi-gen-contigs-database -n GCA_029382735 -f contigs-fasta-GCA_029382735 -o contigs-db-GCA_029382735.db
anvi-gen-contigs-database -n GCA_029382755 -f contigs-fasta-GCA_029382755 -o contigs-db-GCA_029382755.db
anvi-gen-contigs-database -n GCA_029382765 -f contigs-fasta-GCA_029382765 -o contigs-db-GCA_029382765.db
anvi-gen-contigs-database -n GCA_029382805 -f contigs-fasta-GCA_029382805 -o contigs-db-GCA_029382805.db
anvi-gen-contigs-database -n GCA_029382815 -f contigs-fasta-GCA_029382815 -o contigs-db-GCA_029382815.db
anvi-gen-contigs-database -n GCA_029382835 -f contigs-fasta-GCA_029382835 -o contigs-db-GCA_029382835.db
anvi-gen-contigs-database -n GCA_029382845 -f contigs-fasta-GCA_029382845 -o contigs-db-GCA_029382845.db
anvi-gen-contigs-database -n GCA_029382885 -f contigs-fasta-GCA_029382885 -o contigs-db-GCA_029382885.db
anvi-gen-contigs-database -n GCA_029382905 -f contigs-fasta-GCA_029382905 -o contigs-db-GCA_029382905.db
anvi-gen-contigs-database -n GCA_029382925 -f contigs-fasta-GCA_029382925 -o contigs-db-GCA_029382925.db
```

```
anvi-gen-contigs-database -n GCA_029382945 -f contigs-fasta-GCA_029382945 -o
contigs-db-GCA_029382945.db
anvi-gen-contigs-database -n GCA_029382965 -f contigs-fasta-GCA_029382965 -o
contigs-db-GCA_029382965.db
anvi-gen-contigs-database -n GCA_029382985 -f contigs-fasta-GCA_029382985 -o
contigs-db-GCA_029382985.db
anvi-gen-contigs-database -n GCA_029382995 -f contigs-fasta-GCA_029382995 -o
contigs-db-GCA_029382995.db
anvi-gen-contigs-database -n GCA_029383025 -f contigs-fasta-GCA_029383025 -o
contigs-db-GCA_029383025.db
anvi-gen-contigs-database -n GCA_029383035 -f contigs-fasta-GCA_029383035 -o
contigs-db-GCA_029383035.db
 anvi-gen-contigs-database -n GCA_029383045 -f contigs-fasta-GCA_029383045 -o
contigs-db-GCA_029383045.db
anvi-gen-contigs-database -n GCA_029383075 -f contigs-fasta-GCA_029383075 -o
contigs-db-GCA_029383075.db
anvi-gen-contigs-database -n GCA_029383105 -f contigs-fasta-GCA_029383105 -o
contigs-db-GCA_029383105.db
anvi-gen-contigs-database -n GCA_029383115 -f contigs-fasta-GCA_029383115 -o
contigs-db-GCA_029383115.db
anvi-gen-contigs-database -n GCA_029383225 -f contigs-fasta-GCA_029383225 -o
contigs-db-GCA_029383225.db
anvi-migrate *.db --migrate-dbs-safely
anvi-migrate *.db --migrate-safely
anvi-script-gen-genomes-file --input-dir /home/gnmorales/pseudomonas_ecu_genomes/
-o external-genomes_ecu.txt
anvi-estimate-genome-taxonomy -e external-genomes_ecu.txt
anvi-estimate-sc-taxonomy -e external-genomes_ecu.txt
anvi-estimate-scg-taxonomy -e external-genomes_ecu.txt
anvi-estimate-genome-completeness -e external-genomes_ecu.txt
anvi-estimate-genome-completeness -e external-genomes_ecu.txt --debug
anvi-run-hmms -c contigs-db-GCA_029382325.db
anvi-run-hmms -c contigs-db-GCA_029382325.db
anvi-run-hmms -c contigs-db-GCA_029382345.db
anvi-run-hmms -c contigs-db-GCA_029382365.db
anvi-run-hmms -c contigs-db-GCA_029382375.db
anvi-run-hmms -c contigs-db-GCA_029382385.db
anvi-run-hmms -c contigs-db-GCA_029382395.db
anvi-run-hmms -c contigs-db-GCA_029382405.db
anvi-run-hmms -c contigs-db-GCA_029382465.db
anvi-run-hmms -c contigs-db-GCA_029382485.db
anvi-run-hmms -c contigs-db-GCA_029382505.db
anvi-run-hmms -c contigs-db-GCA_029382525.db
anvi-run-hmms -c contigs-db-GCA_029382535.db
anvi-run-hmms -c contigs-db-GCA_029382545.db
anvi-run-hmms -c contigs-db-GCA_029382565.db
anvi-run-hmms -c contigs-db-GCA_029382605.db
```

```
anvi-run-hmms -c contigs-db-GCA_029382625.db
anvi-run-hmms -c contigs-db-GCA_029382645.db
anvi-run-hmms -c contigs-db-GCA_029382655.db
anvi-run-hmms -c contigs-db-GCA_029382665.db
anvi-run-hmms -c contigs-db-GCA_029382705.db
 anvi-run-hmms -c contigs-db-GCA_029382725.db
anvi-run-hmms -c contigs-db-GCA_029382735.db
anvi-run-hmms -c contigs-db-GCA_029382755.db
 anvi-run-hmms -c contigs-db-GCA_029382765.db
 anvi-run-hmms -c contigs-db-GCA_029382805.db
anvi-run-hmms -c contigs-db-GCA_029382815.db
anvi-run-hmms -c contigs-db-GCA_029382835.db
anvi-run-hmms -c contigs-db-GCA_029382845.db
anvi-run-hmms -c contigs-db-GCA_029382885.db
anvi-run-hmms -c contigs-db-GCA_029382905.db
anvi-run-hmms -c contigs-db-GCA_029382925.db
anvi-run-hmms -c contigs-db-GCA_029382945.db
anvi-run-hmms -c contigs-db-GCA_029382965.db
anvi-run-hmms -c contigs-db-GCA_029382985.db
anvi-run-hmms -c contigs-db-GCA_029382995.db
anvi-run-hmms -c contigs-db-GCA_029383025.db
anvi-run-hmms -c contigs-db-GCA_029383035.db
anvi-run-hmms -c contigs-db-GCA_029383045.db
 anvi-run-hmms -c contigs-db-GCA_029383075.db
anvi-run-hmms -c contigs-db-GCA_029383105.db
anvi-run-hmms -c contigs-db-GCA_029383115.db
anvi-run-hmms -c contigs-db-GCA_029383225.db
anvi-script-gen-genomes-file --input-dir /home/gnmorales/pseudomonas_ecu_genomes/
-o external-genomes_ecu.txt
anvi-estimate-genome-completeness -e external-genomes_ecu.txt
anvi-gen-genomes-storage -e external-genomes_ecu.txt -o V_pseudomonas_ecu-
GENOMES.db
anvi-pan-genome -g V_pseudomonas_ecu-GENOMES.db --project-name
V_pseudomonas_ecu --num-threads 12 --minbit 0.5 --mcl-inflation 10 --use-ncbi-blast
anvi-display-pan -p V_pseudomonas_ecu/V_pseudomonas_ecu-PAN.db -g
V_pseudomonas_ecu-GENOMES.db
```