Pontificia Universidad Católica del Ecuador
Facultad de ingeniería
Maestría en Biología Computacional

Ensamblaje y anotación del genoma de *Vibrio spp.* a partir de datos de secuenciación Nanopore (ONT) e Illumina.

Autor: Javier E. Gualdrón Niño

Tutora: MSc. Laura González

Ecuador, Quito 2022

ÍNDICE

1 RESUMEN

Antecedentes: Las especies bacterianas pertenecientes al género Vibrio son bacterias halófilas de tipo gramnegativas que habitan en entornos marinos, manglares, estuarios y por ende en el sistema digestivo de los crustáceos. Este proyecto tiene como objetivo generar un pipeline que permita la limpieza, ensamblaje, pulido y anotación funcional e identificación a nivel de especie de lecturas crudas para el género *Vibrio spp*. obtenidas a través de secuenciación ONT e Illumina, usando diversas herramientas bioinformáticas de código abierto disponible para GNU/Linux.

Métodos: Se realizó una investigación de tipo experimental con datos secundarios extraídos del Centro Nacional para la Información Biotecnológica (NCBI, por sus siglas en inglés) con número de accesión SRR21422415 (*Vibrio spp* CCB-PB317 Nanopore) y SRR21422416 *(Vibrio spp* CCB-PB317 Illumina). La evaluación de calidad se realizó mediante Nanoplot y FastQC respectivamente, a la muestra SRR21422415 se le aplicó una limpieza de los datos con Porechop. Posteriormente se realizó el ensamblaje de los datos limpios de Nanopore con los ensambladores Canu y Flye, los cuales fueron evaluados usando QUAST y BUSCO, seleccionando así el mejor ensamblaje. Se propuso mejorar el ensamblaje obtenido de Nanopore haciendo un pulido con los datos de Illumina, empezando con un mapeo mediante Bowtie2 e indexando con Samtools, los cuales sirvieron de entrada para Pilon. Una vez completado el proceso de pulido se procedió a evaluar el ensamblaje con QUAST y BUSCO. Todos los procesos mencionados fueron procesados en un entorno GNU/Linux a través de consola y usando diversos ambientes de Anaconda. Finalmente, el ensamblaje final fue ingresado en el Centro de Recursos de Bioinformática Bacteriana y Viral (BV-BRC, por sus siglas en inglés) para realizar el Análisis Integral del Genoma (CGA, por sus siglas en inglés) el cual comprende detalles del genoma ensamblado, anotación funcional, anotación proteica, análisis de subsistemas, genes especiales, genes anotados relacionados con mecanismos de resistencia antimicrobiana (AMR) y un árbol filogenético que permite ubicar el genoma con el más cercano.

Resultados: El genoma de Vibrio fue ensamblado a partir de las lecturas obtenidas de bases de datos públicas utilizando los ensambladores de *novo* Canu y Flye. Evaluamos la calidad de los ensamblajes mediante QUAST obteniendo las siguientes estadísticas para Canu (contigs=8, contig más largo=3,075,685 pb, total longitud=5,196,194 pb, N50=3,075,685, %GC=44.92) y Flye (contigs=4, contig más largo=3,118,351 pb, total longitud=5,175,891 pb, N50=3,118,351, %GC=44.87). Adicionalmente, evaluamos la integridad de los genes ensamblados mediante BUSCO y se obtuvo para Canu 81.5% y Flye 91.4% de genes completos (C) identificados en el dataset vibrionales_odb10 de BUSCO. Observamos mejores estadísticas en las dos herramientas utilizando el ensamblaje realizado con Flye, el cual se procedió a realizar el proceso de pulido (*polishing*) mediante lecturas de Illumina. Se evaluó y comparó la calidad del ensamblaje de Flye pulido nuevamente con QUAST (contigs=4, contig más largo=3,118,981 pb, total longitud=5,176,947 pb, N50=3,118,981, %GC=44.87, #Ns=0) y al evaluar la integridad de los genes ensamblados mediante BUSCO se obtuvo un 100.0% de genes altamente conservados de categoría completos (C) identificados en el dataset vibrionales_odb10 de BUSCO, reflejando una mejora significativa al comparar el ensamblaje de Flye sin pulir con el mismo ensamblaje posterior al proceso de pulido. Al ingresar el ensamblaje de Flye pulido al BV-BRC y realizar el Análisis Integral del Genoma (CGA), se obtuvo que el genoma analizado está constituido por 4 contigs el primero con una longitud de 3,118,981 pb (3,1 Mb) y %GC=44.72 el cual fue identificado como el cromosoma I grande, el segundo con una longitud de 1,831,277 pb (1,8 Mb) y %GC=44.95 el cual fue identificado como el cromosoma II pequeño, estos cromosomas son característicos del género *Vibrio spp.* Además, se identificaron dos contigs pequeños de 48,036 pb (48 Kb), %GC=46.45 y 178,653 pb (178 Kb), %GC=46.21 respectivamente los cuales corresponden a plásmidos. En la anotación del genoma se obtuvieron 4791 CDS, 125 ARNt, 100 regiones repetitivas, 33 ARNr, 1112 proteínas hipotéticas, 1086 proteínas según la asignación Enzyme Commission Number (EC number), 905 proteínas con asignación Gene Ontology (GO) y 4648 proteínas con asignación de familia de género cruzado por PATRIC (PGfam). En el caso de genes especiales el CGA anotó un total

de 44 genes relacionados con mecanismos de resistencia antimicrobiana, 40 genes relacionados con posibles objetivos farmacológicos, 64 genes transportadores y 114 genes relacionados con factores de virulencia, finalmente el Análisis Integral de Genoma determinó que por medio de un análisis filogenético que la especie más cercana a nuestro genoma corresponde a *Vibrio alginolyticus.*

Conclusiones: El desarrollo exponencial de los últimos años de las tecnologías de secuenciación masiva permite en la actualidad obtener millones de secuencias de ADN a una gran velocidad y un costo cada vez menor, lo que genera una elevada cantidad de datos crudos en diversas bases de datos, entre ellas el Sequence Read Archive (SRA, por sus siglas en inglés) del NCBI. Por tal motivo el diseño de pipelines (tuberías) para el análisis de dichos datos que permita extraer la mayor cantidad de información útil de dichas secuencias prácticamente sin costo alguno se convierte en una herramienta poderosa en la biología computacional. Este pipeline diseñado para bacterias permite realizar ensamblajes de datos crudos obtenidos con Nanopore con dos de los mejores ensambladores de *novo* actualmente disponible Canu y Flye, en este caso con mejor desempeño el ensamblaje generado por Flye, el cual posterior al proceso de pulido con datos crudos de illumina mejoró sustancialmente, lo cual es crucial si se desea realizar un análisis de anotación funcional ya que la misma depende de la integridad y calidad del ensamblaje. Logramos identificar que nuestra muestra corresponde a *Vibrio alginolyticus* con una estructura genómica constituida por un cromosoma I grande y un cromosoma II pequeño, siendo común para el género, junto con dos regiones pequeñas que corresponden a plásmidos. Adicionalmente se evidencia la capacidad patogénica de la especie analizada por la cantidad de genes anotados relacionados con factores de virulencia y mecanismos de resistencia antimicrobiana.

Palabras claves: *Vibrio spp*., Ensamblaje del genoma, Pulido del genoma, Anotación funcional.

2 INTRODUCCIÓN

El género *Vibrio* comprende especies bacterianas halófilas caracterizadas por ser bacilos curvos gram-negativos presentes como organismos de vida libre en hábitats acuáticos y marinos, o asociadas con peces e invertebrados marinos. Estas especies presentan mecanismos de resistenciametabolica a diferentes concentraciones de cloruro de sodio (NaCl), así como también otras características fisicoquímicas de igual importancia como crecimiento en un amplio rango de temperaturas, la aerobiosis facultativa, catalasa positiva, oxidasa positiva y antígeno somático O permitiendo su desarrollo en diversos ambientes (Cumbicos & Ruiz, 2018). Varias especies de este género bacteriano causan enfermedades transmitidas por los alimentos e infecciones de heridas que resultan del contacto con agua o mariscos crudos. El número de casos de infección humana por especies de *Vibrio spp.* ha aumentado a nivel mundial. Las enfermedades causadas por *Vibrio spp*. se dividen en dos grandes grupos: el cólera, causado por *Vibrio cholerae* y la vibriosis, causada por otras especies patógenas de *Vibrio* (Baker-Austin et al., 2018). El cólera es una enfermedad diarreica de rápida propagación que provoca una rápida deshidratación y puede ser mortal si no se trata a tiempo (WHO, 2017) Los pacientes con vibriosis presentan una variedad de síntomas según la especie de *Vibrio spp* infectante, la ruta de infección y la susceptibilidad del huésped. Los síntomas van desde enfermedades gastrointestinales, como gastroenteritis, hasta patologías extraintestinales, incluidas infecciones de la piel y septicemia (Baker-Austin et al., 2018; Pérez-Duque et al., 2021).

Según el informe de la Organización Mundial de la Salud (OMS), alrededor de 3,5 millones de personas se infectan y entre 100.000 y 120.000 mueren a causa del cólera cada año en los países en desarrollo. En 2014, la letalidad por cólera fue del 1,17 % en 42 países, lo que representó un aumento del 47 % en comparación con 2013 (WHO, 2017). Las pandemias de cólera ocupan el sexto lugar entre las diez peores pandemias del mundo, la enorme tasa de mortalidad es particularmente reportada por países que

sufren de malas instalaciones de saneamiento y falta de sistemas adecuados de distribución de agua (Lekshmi et al., 2018).

Los estudios de secuenciación han dado a conocer que las especies de *Vibrio* tienen genomas altamente plásticos ya que existe una alta probabilidad de transferencia horizontal de genes de virulencia y resistencia a antibióticos de cepas virulentas a no virulentas debido a su naturaleza de vida libre (Pérez-Duque et al., 2021). La secuenciación del genoma completo de varios aislados de *Vibrio* y los análisis genómicos comparativos han revelado una amplia gama de mutaciones, reordenamientos cromosómicos y eventos de ganancia y pérdida de genes resultantes de la duplicación o la transferencia genética horizontal (Thompson et al., 2004). Los genes de virulencia normalmente se identifican dentro de elementos genéticos móviles; por lo tanto, estos elementos pueden favorecer la aparición de nuevas cepas virulentas (Castillo et al., 2018). El genoma de *Vibrio spp.* comprende un genoma central compuesto por genes conservados necesarios para funciones esenciales y un acervo genético flexible, que incluye elementos móviles, factores de virulencia y determinantes de resistencia a los antibióticos, entre otros, que están involucrados en los procesos de adaptación (Pérez-Duque et al., 2021), sin embargo, el proceso de secuenciación por sí solo no brinda información relevante sobre ningún patógeno incluido *Vibrio spp.* Por tal motivo esta investigación tiene como objetivo generar un pipeline que permita realizar la anotación funcional e identificación a nivel de especie para *Vibrio spp*. con especial enfoque en el proceso de optimización del ensamblado, usando herramientas bioinformáticas de código abierto específicas para cada tipo de muestra y proceso pudiendo ser reproducible por otros investigadores.

3 MARCO TEÓRICO

3.1 Familia Vibrionaceae

Verón en 1965 propuso que la familia Vibrionacea estaba constituida por 7 géneros: *Vibrio, Photobacterium, Allomonas, Listonella, Enhydrobacter, Salinivibrio y Enterovibrio*, de los cuales el género *Vibrio* ha sido descrito con el mayor número de especies (Faruque et al., 2006). Las especies de Vibrionaceae también se han utilizado ampliamente en estudios fisiológicos, bioquímicos, de biología molecular y de patogenicidad (Baumann et al., 1983).

3.1.1 Género Vibrio

El género Vibrio está constituido por bacilos con forma curva de tipo gramnegativos, donde la mayoría tiene como hábitat predilecto los ambientes acuáticos de manera libre o en asociación a fitoplancton y zooplancton. Para lograr su desarrollo en diversos ambientes, es decir, se manera ubicua se vale de diversos mecanismos metabólicos como son la resistencia a diferentes concentraciones de sal, aerobiosis de tipo facultativo, catalasa/oxidasa positiva, antígeno somático O y crecimiento a diversos grados de temperatura. Permitiéndole crecer tanto en agua dulce como salada, en el tracto digestivo de los seres humano y de animales. Además, puede infectar animales y humanos a través de la ingesta de alimentos contaminados, así como también por medio de heridas (Cumbicos & Ruiz, 2018; Vadillo et al., 2002).

Si bien el género Vibrio se estima que está compuesto por al menos 80 especies, es importante considerar que 12 son patógenas para los seres humanos, entre las cuales estan; *V. cholerae, V. parahaemolyticus, V. vulnificus, V. alginolyticus, V. fluviales, V. furnisii, V. hollisae, V. carchariae*, *V. damsela, V. mimicus, V. metschnikovii y V. cincinnatiensis.* Las cinco primeras son de gran importancia clínica pues se han relacionado a nivel mundial con enfermedades como el cólera, causada por *V. cholerae*, así como un grupo de infecciones denominadas Vibriosis provocando cuadros que van desde

gastroenteritis agudas, infecciones en heridas y tejidos blandos hasta septicemia, esta última generalmente producido por el consumo de alimentos contaminados (Morris, 2003).

En el caso de *Vibrio cholerae* también se desarrolla en condiciones similares a las mencionadas como son ambientes marinos con niveles moderados de NaCl, y es la especie de Vibrio con mayor relevancia como patógeno humano. Se han caracterizado alrededor de 200 serogrupos que incluyen cepas tanto toxigénicas y no toxigénicas, dicha clasificación se basa en la presencia del polisacárido O (antígeno A), en el caso de la clasificación de serotipos depende de las combinaciones de los antígenos somáticos B y C, mientras que la clasificación a nivel biotipos se realiza de acuerdo con características fenotípicas específicas (Percival & Williams, 2014).

3.2 Epidemiología

Desde 1817, ha habido siete pandemias de cólera hasta la fecha, la alta tasa de mortalidad y su propagación en los principales continentes, como Asia, América, Europa y África, lo hacen importante para la investigación. Se considera que el subcontinente indio es la patria del cólera y se ha extendido rápidamente a otros países del mundo. Los síntomas que distinguen el cólera de otras enfermedades diarreicas son las típicas heces acuosas de arroz y vómitos intensos que pueden provocar deshidratación y muerte en 48 horas si no se tratan. Según el informe de la Organización Mundial de la Salud (OMS) 3.5 millones de personas se infectan y entre 100 000 y 120 000 mueren a causa del cólera cada año en los países en desarrollo. En 2014, la letalidad por cólera fue del 1,17 % en 42 países, lo que representó un aumento del 47 % en comparación con 2013 (WHO, 2015). Las pandemias de cólera ocupan el sexto lugar entre las diez peores pandemias del mundo, la enorme tasa de mortalidad es particularmente reportada por países que tienen deficientes sistemas de salubridad y distribución de agua. Los casos transfronterizos de cólera han sido un problema grave en los países del África Subsahariana desde hace muchos años, elevando la tasa de letalidad al 5 por ciento. Los niños en edad escolar y los niños menores de cinco años

son los más afectados en los países donde el cólera es endémico. Los casos de cólera y las muertes en los países de África y el sur de Asia han representado el 99 por ciento del total de casos de cólera en todo el mundo (Lekshmi et al., 2018).

En América se han levantado alertas en todo el continente por los brotes de cólera en Haití y República Dominicana en 2010 y 2011, así como reportes de casos en Cuba, Venezuela, México y Colombia. Aunque *V. cholerae, V. parahaemolyticus* y *V. vulnificus* son los principales patógenos notificados, también existe preocupación por el aumento de los casos de infección por otras especies de Vibrio (Pérez-Duque et al., 2021).

3.3 Genómica de *Vibrio spp.*

El Vibrio proviene de una variedad de nichos ecológicos y a lo largo de la cadena alimenticia se ha investigado previamente. Estos estudios generalmente se han centrado en especies específicas de *Vibrio spp.* y usan técnicas de detección fenotípicas y moleculares; sin embargo, la mayoría no incluyen secuenciación del genoma (WGS), que proporciona la resolución molecular más alta para investigar la evolución bacteriana y la estructura de la población. Una gran investigación genómica transcontinental de *V. parahaemolyticus* identificó la actividad humana como responsable de cambiar los patrones de distribución global de este organismo, pero el alcance se limitó a una sola especie (Yang et al., 2019). Hasta la fecha se sabe poco sobre los tipos y la diversidad genómica de *Vibrio spp.* encontrados en el comercio minorista, las implicaciones para la seguridad alimentaria y, en consecuencia, para la salud humana (Janecko et al., 2021).

Así mismo a través de estudios de secuenciación se ha dado a conocer que las especies de Vibrio tienen genomas altamente plásticos ya que existe una alta probabilidad de transferencia horizontal de genes de virulencia y resistencia a antibióticos de cepas virulentas a no virulentas debido a su naturaleza de vida libre. La secuenciación del genoma completo de varios aislados de Vibrio y los análisis genómicos

comparativos han revelado una amplia gama de mutaciones, reordenamientos cromosómicos y eventos de ganancia y pérdida de genes resultantes de la duplicación o la transferencia horizontal de genes (Pérez-Duque et al., 2021). Los genes de virulencia normalmente se identifican dentro de elementos genéticos móviles; por lo tanto, estos elementos pueden favorecer la aparición de nuevas cepas virulentas. Todos los Vibrios conocidos comprenden un genoma central constituido por dos cromosomas; la presencia de dos cromosomas en *V. cholerae* se documentó por primera vez en 1998. El cromosoma I suele ser más grande, con un tamaño relativamente constante de alrededor de 3 millones de pares de bases, que codifica en promedio de 2700 proteínas necesarias para el metabolismo funcional esencial. Por el contrario, el cromosoma II es de menor tamaño alrededor de 1 millón de pares de bases las cuales codifican en promedio unas mil proteínas y contiene un "superintegrón" que varía comúnmente. Los genomas de Vibrio contienen muchas islas genómicas, que pueden contener funciones que permiten la adaptación a entornos específicos e incluso pueden representar eventos de diversidad a nivel de especie (Lukjancenko & Ussery, 2014), así mismo *Vibrio spp* comprende genes conservados necesarios para funciones esenciales y un acervo genético flexible, que incluye elementos móviles, factores de virulencia y determinantes de resistencia a los antibióticos, entre otros, que están involucrados en los procesos de adaptación (Pérez-Duque et al., 2021; Watve et al., 2016).

3.4 Herramientas utilizadas para analizar datos de secuenciación

NOTA: Todos los códigos de ejecución de herramientas utilizadas en esta investigación, se encuentran disponibles como material suplementario.

La primera etapa de análisis de datos de secuenciación incluye la revisión de calidad de los datos, sean lecturas largas o lecturas cortas. Para el caso de lecturas cortas se usa frecuentemente FastQC y para lecturas largas se utiliza Nanoplot:

3.4.1 Nanoplot

(https://github.com/wdecoster/NanoPlot) que permite evaluar la calidad de secuenciación para lo cual se siguió las indicaciones de los desarrolladores para su instalación y ejecución, obteniendo como salida un resumen estadístico, una serie de gráficos y un archivo de resumen html (de Coster et al., 2018).

3.4.2 FastQC

(https://github.com/s-andrews/FastQC) es un software ampliamente usado en bioinformática ya que permite encontrar errores o problemas en los datos de secuenciación de alto rendimiento, realizando un conjunto de análisis en los archivos de secuenciación sin procesar, es decir, en formato fastq o bam; emitiendo un informe HTML que resume los resultados obtenidos de dicho análisis (Andrews, 2010).

Seguido a la revisión de calidad, se hace una limpieza de calidad, utilizando Porechop:

3.4.3 Porechop

(https://github.com/rrwick/Porechop) es una herramienta que permite encontrar y quitar adaptadores de lecturas de Oxford Nanopore. Los adaptadores en los extremos de las lecturas se recortan, y cuando una lectura tiene un adaptador en el medio, se trata como quimera y se corta en lecturas separadas (Bonenfant et al., 2022).

Una vez tenemos los datos limpios, podemos realizar el ensamblaje del genoma. Para este proceso existen dos aproximaciones. La primera realiza el ensamblaje de lecturas cortas y utiliza las lecturas largas para unir los contigs generados; y la segunda hace un ensamblaje de lecturas largas seguido de corrección de errores con lecturas cortas. Esta última será la aproximación utilizada en este estudio. Algunas de las herramientas para ensamblar lecturas largas de Nanopore son Canu y Flye:

3.4.4 Canu

(https://github.com/marbl/canu) es una bifurcación de Celera Assembler (ensamblador WGS de novo), diseñado para la secuenciación de una sola molécula de alto ruido (como PacBio RS II / Sequel u Oxford Nanopore MinION). Es una tubería de ensamblaje jerárquica que se ejecuta en cuatro pasos: detecta superposiciones en secuencias de alto ruido usando MHAP, genera consenso de secuencia corregida, recorta secuencias corregidas, ensambla secuencias corregidas recortadas (Koren et al., 2017).

3.4.5 Flye

(https://github.com/fenderglass/Flye) es un ensamblador de novo para lecturas de secuenciación de una sola molécula, como las producidas por PacBio y Oxford Nanopore Technologies. El paquete representa una canalización completa: toma lecturas sin procesar de PacBio/ONT como entrada y salidas de contigs pulidos (Kolmogorov et al., 2019).

Los genomas resultantes de este ensamblaje de lecturas largas requieren ser pulidos con lecturas cortas que presentan una tasa de error muy baja, proceso que requiere el mapeo de lecturas cortas al genoma con herramientas como Bowtie2 e indexado con Samtools, y una sustitución basada en la identificación de errores con el software Pilon:

3.4.6 Bowtie2

(https://github.com/BenLangmead/bowtie2) es una herramienta usada para la alineación de lecturas de secuenciación en base a secuencias de referencia de gran longitud, con características como su gran rapidez y eficiencia en memoria computacional. Logrando alinear entre 50-1000 caracteres, lo que le hace particularmente eficiente para alinear genomas relativamente largos como por ejemplo, de mamíferos. Esto lo logra indexando el genoma con un índice FM, lo cual se traduce en menor espacio de

memoria necesaria, por ejemplo, para el genoma del humano, la memoria necesaria suele ser de alrededor de 3,2 GB (Langmead & Salzberg, 2012b).

3.4.7 Pilon

(https://github.com/broadinstitute/pilon/wiki) es una herramienta de software que se puede utilizar para mejorar automáticamente los borradores de ensamblajes y encuentra variaciones entre cepas, Pilon requiere como entrada un archivo FASTA del genoma junto con uno o más archivos BAM de lecturas alineadas con el archivo FASTA de entrada para analizando la alineación de lectura para identificar inconsistencias entre el genoma de entrada y la evidencia en las lecturas. Luego intenta realizar mejoras en el genoma de entrada, que incluyen: diferencias de base única, pequeños indeles, eventos de sustitución de bloque o indel más grandes, relleno de huecos, identificación de desmontajes locales, incluida la apertura opcional de nuevos espacios (Walker et al., 2014).Finalmente, podemos realizar la revisión de la calidad del genoma utilizando herramientas de código abierto. Las más utilizadas son QUAST y BUSCO, estas dos herramientas nos aportan información acerca de la completitud del genoma y nos permiten tomar decisiones de mejora en la secuenciación, proceso de ensamblaje o pulido, e incluso para comparar varios ensamblajes y elegir el mejor.

3.4.8 QUAST

(https://github.com/ablab/quast) es una herramienta de evaluación del ensamblaje del genoma, el cual evalúa ensamblajes de genoma/metagenoma. El paquete QUAST funciona con y sin genomas de referencia (Gurevich et al., 2013).

3.4.9 BUSCO

(https://github.com/WenchaoLin/BUSCO-Mod) la evaluación de integridad de BUSCO emplea conjuntos de ortólogos universales de copia única de evaluación comparativa de OrthoDB

(www.orthodb.org) para proporcionar medidas cuantitativas de la integridad de los ensamblajes del genoma, conjuntos de genes anotados y transcriptomas en términos del contenido genético esperado (Simão et al., 2015).

Una vez completado el proceso de evaluación de la calidad de nuestro ensamblaje pulido (final), se puede ingresar el mismo en cualquier base de datos según la necesidad del investigador, en este caso se eligió:

3.4.10 El Centro de Investigación de Bioinformática Bacteriana y Viral (BV-BRC)

(https://www.bv-brc.org/) el cual se formó en 2019 a través de la fusión de tres recursos de BRC derivados del Instituto Nacional de Alergias y Enfermedades Infecciosas (NIAID) (Greene et al., 2007): el Centro de Integración de Recursos de PAThosystems (PATRIC) (Davis et al., 2019), la Base de Datos de Investigación de Influenza (IRD) (Zhang et al., 2017) y la Base de datos y recurso de análisis para virus patógenos (ViPR) (Pickett et al., 2012). PATRIC fue uno de los BRC originales y fue diseñado para analizar a través de la bioinformática a patógenos bacterianos. En 2012, el recurso nacional de base de datos de patógenos microbianos (NMPDR) se fusionó con PATRIC-BRC, incorporando los conocidos recursos de anotación SEED y RAST (Overbeek et al., 2014). IRD y ViPR fueron diseñados para respaldar los análisis de la influenza y una variedad de otros patógenos virales humanos importantes, respectivamente. El recurso BV-BRC combinado actual cuenta con el respaldo de equipos de investigadores de la Universidad de Chicago, el Instituto J. Craig Venter, la Universidad de Virginia y la organización Fellowship for Interpretation of Genomes (FIG), así como muchos colaboradores cercanos en otras instituciones (Olson et al., 2022).

4 PLANTEAMIENTO DEL PROBLEMA

Vibrio spp. se encuentra comúnmente en los Estados Unidos de América, donde el número de casos de infección ha aumentado desde el año 2000, alcanzando aproximadamente 80.000 casos y 100 muertes al año. De manera similar, los países de Asia informaron sobre los impactos de las infecciones por *Vibrio spp*. en humanos y la muerte masiva de animales marinos, lo que afectó a la industria pesquera. En Europa, la presencia de *Vibrio vulnificus*, *Vibrio parahaemolyticus* y otras especies de Vibrio ha suscitado preocupación con respecto a su potencial para generar problemas clínicos significativos y brotes asociados con el cambio climático. En América se han levantado alertas en todo el continente por los brotes de cólera en Haití y República Dominicana en 2010 y 2011, así como reportes de casos en Cuba, Venezuela y México (Janecko et al., 2021). Igualmente, Colombia por su ubicación geográfica se han reportados casos de colera y vibriosis, siendo *V. cholerae, V. parahaemolyticus* y *V. vulnificus* son los principales patógenos notificados, también existe preocupación por el aumento de los casos de infección por otras especies de *Vibrio spp.* y la aparición de linajes epidémicos (Pérez-Duque et al., 2021).

Las especies de Vibrio son bacterias halófilas gramnegativas que habitan de forma natural en diversos entornos marinos o acuáticos, en la cadena alimentaria, *Vibrio spp*. Se encuentran predominantemente en productos del mar, cuyo consumo se ha asociado con enfermedades humanas (Watve et al., 2016). En particular, *Vibrio parahaemolyticus, V. vulnificus, V. alginolyticus* y *V. cholerae* son patógenos microbianos importantes con factores de virulencia característicos. No se conocen bien todos los factores de virulencia, sin embargo, la presencia de genes de hemolisina termoestable tdh y trh en *V. parahaemolyticus*, el gen de hemolisina vvhA en *V. vulnificus*, así como los genes que codifican las toxinas ctxA y ctxB en el caso de *V. cholerae* están vinculados con factores de virulencia y son utilizados por laboratorios de salud pública para evaluar la patogenicidad humana. Para la caracterización adicional de los ecotipos de *V. vulnificus* se ha utilizado genes vcgC (clínico), vcgE (ambiental) y mutaciones dentro de la estructura del gen pilF puede informar el linaje epidemiológico de cepas presumiblemente más

virulentas. Las enfermedades asociadas a Vibrio se manifiestan como gastroenteritis de leve a grave, infecciones de heridas y, en algunos casos, septicemia. En el caso de *Vibrio cholerae*, los serotipos O1 y O139, que son los agentes causantes de la enfermedad del cólera pandémico, están vinculados a fuentes de agua contaminada (Janecko et al., 2021; Watve et al., 2016).

Una vez comprendido el problema y tomando en cuenta que los aislamientos recuperados pertenecientes a vibrios no relacionados con el cólera no han sido genotipados ni caracterizados a nivel del genoma, se desconoce si estos aislamientos portan genes de toxina/virulencia y rasgos de resistencia antimicrobiana (Pérez-Duque et al., 2021). Por lo tanto, los estudios de caracterización genética son importantes para comprender los riesgos para la salud pública asociados con la presencia de estas bacterias en el medio ambiente; las muestras clínicas a menudo se aíslan de regiones geográficas limitadas y se derivan clonalmente con pocas o ninguna diferencia genética (Watve et al., 2016).

Una caracterización ampliada de los genomas de aislamientos ambientales de *Vibrio spp*., que tienden a ser mucho más diversos desde el punto de vista genético y fenotípico, debería aumentar sustancialmente la diversidad de secuencias disponibles de este importante patógeno humano, sin embargo, las secuencias por sí solas no ofrecen ninguna clave sobre los patógenos secuenciados, para extraer información genómica es necesario la aplicación de diversas herramientas bioinformáticas que permitan determinar su estructura cromosómica a nivel de subsistemas y determinación de genes relacionados con factores de virulencia y resistencia antimicrobiana a fin de establecer su grado de patogenicidad y fuentes potenciales de brotes epidémicos para adoptar medidas oportunas ante una contingencia, que permitan reducir el impacto de este patógeno.

5 DISEÑO Y ALCANCE

El diseño de la investigación corresponde al tipo experimental, tomando en cuenta que es un proceso basado en la búsqueda, recuperación, análisis e interpretación de datos obtenidos a partir de datos secundarios, es decir, datos de secuenciación obtenido por otros investigadores depositados en cualquier base de datos que pueda alojar datos SRA, como es el caso del NCBI (Arias, 2006; Caro et al., 2005).

Tomando en cuenta el diseño de la investigación, se define la misma con un propósito de crear ciencia aplicada la cual tiene como objetivo aplicar diversos algoritmos bioinformáticos de código abierto de manera combinada para crear un pipeline que permita un ensamblaje de alta calidad y un Análisis Integral del Genoma (CGA) para *Vibrio spp.*, el cual está dirigido a incrementar los postulados teóricos y prácticos sobre este género bacteriano perteneciente a la familia Vibrionaceae (Arias, 2006; Noreña et al., 2012).

Se propone un alcance de tipo descriptivo donde se busca especificar las propiedades, las características y el perfil genómico de cualquier muestra secuenciada de *Vibrio spp* a partir de los archivos crudos (FASTQ) de ONT e Illumina, con el fin de determinar la anotación funcional, estructural y su árbol filogenético (Arias, 2006; Hernández-Sampieri et al., 2014), pudiendo ser reproducible por otros investigadores.

6 OBJETIVOS

6.1 OBJETIVO GENERAL

Generar un pipeline que permita la limpieza, ensamblaje y pulido, sirviendo como entrada para realizar el Análisis Integral del Genoma (CGA) del Centro de Recursos de Bioinformática Bacteriana y Viral (BV-BRC) a partir de lecturas crudas provenientes del SRA del NCBI para el género *Vibrio spp*. obtenidas a través de secuenciación ONT e Illumina.

6.2 OBJETIVOS ESPECÍFICOS

- Evaluar la calidad de lecturas crudas obtenidas a través de secuenciación ONT e Ilumina y limpiar las lecturas.
- Ensamblar de lecturas crudas obtenidas a través de secuenciación ONT usando dos ensambladores.
- Realizar el proceso de pulido (Polishing) con datos Illumina.
- Evaluar la calidad de los ensamblajes antes y después del pulido usando QUAST y BUSCO.
- Realizar un Análisis Integral del Genoma (CGA) por medio del Centro de Recursos de Bioinformática Bacteriana y Viral (BV-BRC).

7 DATOS

Se ha planteado un método de recolección de datos fundamentado en el tipo secundario, es decir, aquellos datos recolectados por otros investigadores y depositados en el Sequence Read Archive (SRA) del NCBI (Hernández-Sampieri et al., 2014).

Este método de recolección de datos del tipo secundario tiene como propósito general, recabar datos que son difíciles de recolectar para esta investigación, considerando las notables ventajas de poder utilizar datos de secuenciación sin incurrir en gastos operativos y tiempo de análisis *in vitro*, esto va acorde al alcance propuesto de la investigación actual (D'Ancona & Angeles, 2012; Hernández-Sampieri et al., 2014), las características de las muestras usadas para este proyecto se pueden apreciar en la Tabla 1

Tabla 1

Características de secuenciación de las muestras usadas.

Características	***SRR21422415 Vibrio sp CCB-PB317 Nanopore***	***SRR21422416 Vibrio sp CCB-PB317 Illumina***
Experiment Accession	SRX17426752	SRX17426751
Experiment Title	Vibrio sp CCB-PB317 Nanopore sequencing	Vibrio sp CCB-PB317 Illumina sequencing
Organism Name	Vibrio sp. CCB-PB317	Vibrio sp. CCB-PB317
Instrument	MinION	Illumina NovaSeq 6000
Submitter	Universiti Sains Malaysia	Universiti Sains Malaysia
Study Accession	SRP395637	
Study Title	Complete Whole Genome Sequence of Vibrio sp. CCB-PB317 Isolated From Mangrove	Complete Whole Genome Sequence of Vibrio sp. CCB-PB317 Isolated From Mangrove
Sample Accession	SRS14982227	SRS14982227
Total Size, Mb	136.61	357.48

Total RUNs	1	1
Total Spots	46174	4094156
Total Bases	154325573	1228246800
Library Name	CCB-PB317 Nanopore	CCB-PB317 Illumina
Library Strategy	WGS	WGS
Library Source	GENOMIC	GENOMIC
Library Selection	RANDOM	RANDOM

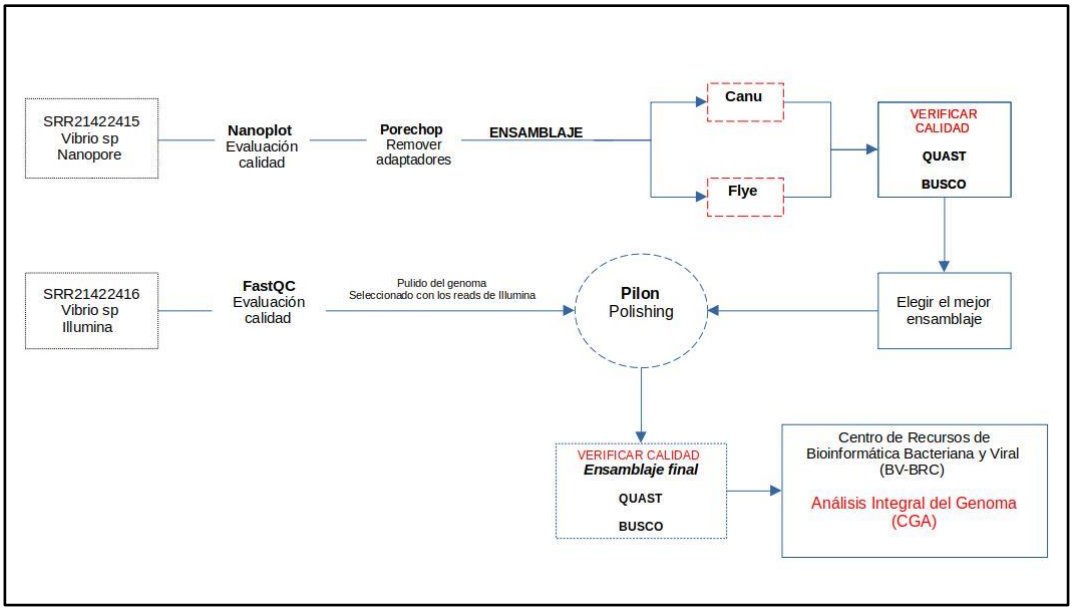
Nota: Esta tabla muestra todas las características disponibles en el Sequence Read Archive (SRA) del NCBI de las dos muestras seleccionadas para este estudio.

8 METODOLOGÍA Y RESULTADOS

El pipeline diseñado fue ejecutado en un entorno GNU/Linux a través de consola por medio de la instalación de las herramientas directamente en el ordenador o usando Bioconda. Bioconda permite instalar miles de paquetes de software relacionados con la investigación biomédica utilizando el administrador de paquetes Conda. Por último, el análisis del genoma ensamblado fue realizado a través del Centro de Recursos de Bioinformática Bacteriana y Viral (BV-BRC). La descripción general del pipeline diseñado se puede observar en la Figura 1.

Figura 1

Pipeline para el ensamblaje de lecturas ONT usando lecturas de Illumina para el pulido (polishing)



Nota: La figura muestra el pipeline o pasos a seguir de manera general que se siguió para realizar el ensamblaje pulido final de inicio a fin, el cual sirvió como entrada para el Análisis Integral del Genoma (CGA).

Se seleccionó una muestra de *Vibrio spp*. proveniente de manglar y secuenciada usando diferentes tecnologías (SRR21422415 Vibrio sp CCB-PB317 Nanopore y SRR21422416 Vibrio sp CCB-PB317

Illumina). Para esta muestra, fueron extraídas las lecturas crudas del NCBI-SRA en formato fastq y se procedió analizar la calidad de las mismas. Las lecturas de *Vibrio sp* CCB-PB317 Nanopore (SRR21422415) fueron analizadas por medio de Nanoplot, el cual permite la evaluación integral de la calidad de alineaciones y datos de secuenciación de lectura larga (de Coster et al., 2018). Los resultados pueden ser observados en la Tabla 2, Figura 2 y Figura 3. En paralelo se analizó la calidad de las lecturas de *Vibrio spp* CCB-PB317 Illumina (SRR21422416) por medio de FastQC, herramienta ampliamente utilizada para la evaluación integral de lecturas crudas de Illumina (Andrews, 2010), Los resultados generados se presentan en la Tabla 3 y Figura 4.

Tabla 2

Resumen de la evaluación de calidad de la muestra SRR21422415 usando Nanoplot

Medida	***Valor***
Mean read length	3342.3
Mean read quality	12.2
Median read length	2583
Median read quality	12.4
Number of reads	46174
Read length N50	4263
STDEV read length	2397.9
Total bases	154325573

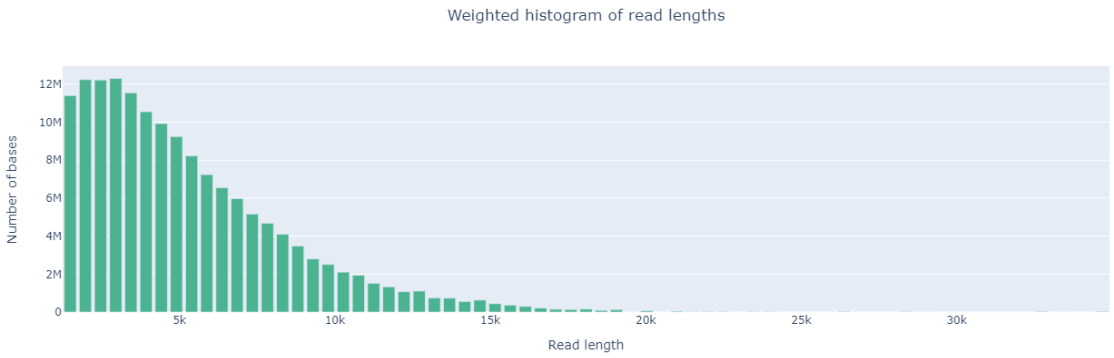
Number, percentage and megabases of reads above quality cutoffs	
>Q5	46174 (100.0%) 154.3Mb
>Q7	46174 (100.0%) 154.3Mb
>Q10	36710 (79.5%) 122.8Mb
>Q12	25725 (55.7%) 86.6Mb
>Q15	5619 (12.2%) 19.1Mb
Top 5 highest mean basecall quality scores and their read lengths	
1	18.9 (1150)
2	18.6 (1882)
3	18.4 (1019)
4	18.3 (1122)
5	18.3 (1168)
Top 5 longest reads and their mean basecall quality score	
1	34675 (14.5)

2	32441 (13.6)
3	28123 (9.5)
4	26258 (11.6)
5	23643 (7.7)

Nota: En la tabla se puede observar todas las métricas de calidad emitidas por el programa Nanoplot cuando se analizó la muestra SRR21422415 Vibrio sp CCB-PB317 secuenciada mediante tecnología Nanopore.

Figura 2

Histograma ponderado de las longitudes de lecturas de la muestra SRR21422415 generado por Nanoplot



Nota: El histograma refleja la el promedio de las longitudes o largo de las lecturas con respecto al número de bases de la muestra SRR21422415 Vibrio sp CCB-PB317 secuenciada mediante Nanopore por cada lectura, generado por Nanoplot.

Figura 3

Gráfico de puntos: Longitud de las lecturas vs promedio de calidad de las lecturas de la muestra SRR21422415 generado por Nanoplot

Nota: Este grafico de puntos demuestra la relación entre la longitud de las lecturas (eje X) y la calidad de dichas lecturas (eje Y), demostrando una distribución uniforme desde las lecturas más cortas hasta las lecturas más largas (alrededor de 18 mil pares de bases) de la muestra SRR21422415 Vibrio sp CCB-PB317 secuenciada mediante tecnología Nanopore, generado Nanoplot.

Tabla 3

Resumen de la evaluación de calidad de la muestra SRR21422416 mediante FastQC

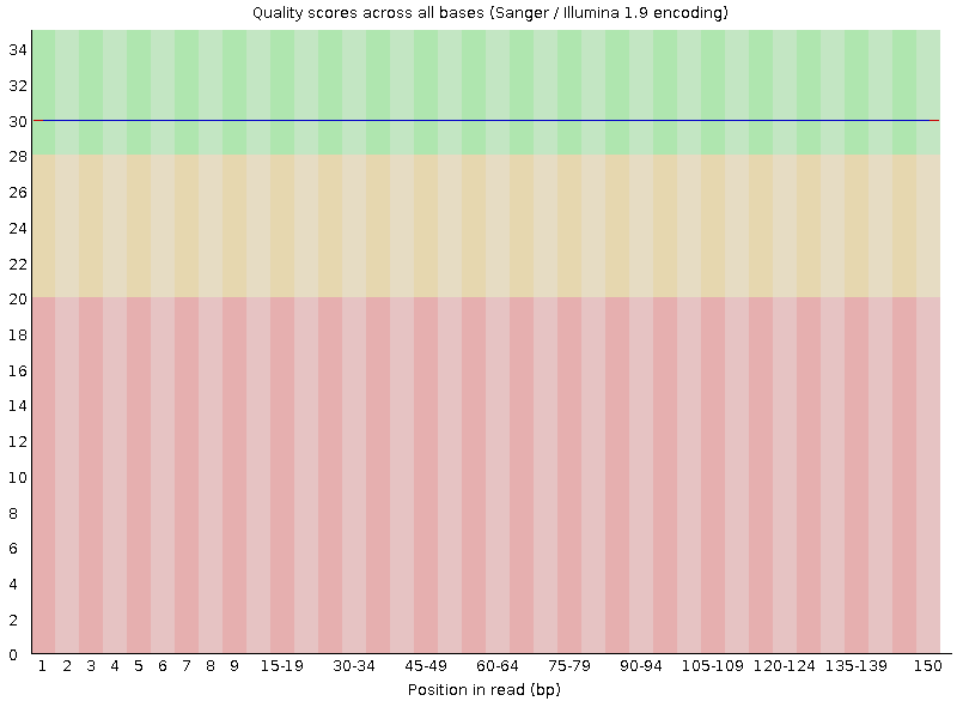
Medida	***Valor***
Filename	SRR21422416.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8188312

Sequences flagged as poor quality	0
Sequence length	150
%GC	45

Nota: Resumen de la evaluación de calidad de la muestra SRR21422416 Vibrio sp CCB-PB317 secuenciada mediante tecnología Illumina generado a través de FastQC con un promedio de longitud de lectura de 150 pb y un %GC del 45.

Figura 4

Promedio de la calidad de secuenciación (Phred score) por lectura de la muestra SRR21422416 mediante FastQC

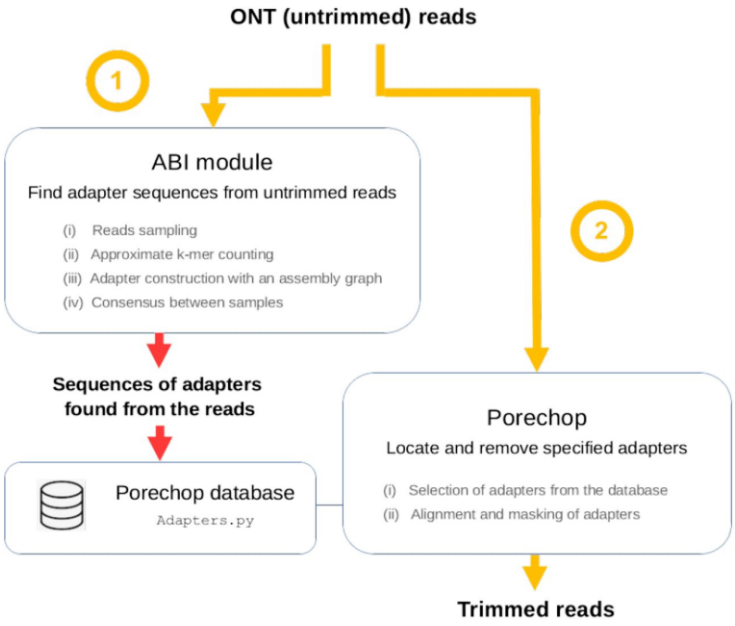


Nota: Se puede observar en este gráfico cuál es el promedio de calidad (Phred Score) en el eje Y por lecturas de toda la secuenciación del eje X, de la muestra RR21422416 Vibrio sp CCB-PB317 secuenciada mediante tecnología Illumina generado a través de la herramienta FastQC.

Analizando la calidad de las lecturas de Illumina se determinó que no era necesaria la limpieza de la misma. Así mismo, tomando en cuenta que la secuenciación de lecturas largas presenta una calidad inferior a la obtenida por Illumina con un promedio de calidad de lectura de 12.2, se procedió a realizar la limpieza de las lecturas Nanopore mediante Porechop permitiendo encontrar y quitar adaptadores de lecturas de Oxford Nanopore de la siguiente manera: los adaptadores en los extremos de las lecturas se recortan, y cuando una lectura tiene un adaptador en el medio, se trata como quimera y se corta en lecturas separadas (Figura 5) (Bonenfant et al., 2022), en nuestro caso se puede observar las estadísticas proporcionadas por Porechop en la Tabla 4, una vez finalizado la limpieza o trimado Porechop genera como archivo de salida la secuencia trimada en formato fasta.

Figura 5

Algoritmo de Porechop para la limpieza de lecturas ONT.



Nota: En la figura se aprecia el algoritmo utilizado por la herramienta Porechop (Bonenfant et al., 2022) para realizar el proceso de limpieza/trimado de todas las lecturas de la muestra SRR21422415 *Vibrio sp* CCB-PB317 secuenciada mediante tecnología Nanopore.

Tabla 4

Estadísticas de limpieza/trimado de la muestra SRR21422415 mediante Porechop

Total lecturas de entrada	46,174
Lecturas que contienen adaptadores en el inicio	45,914
Lecturas que contienen adaptadores en el final	8,978
Bases removidas del inicio (pb)	5,105,988
Bases removidas del final (pb)	90,501
Lecturas divididas en función de los adaptadores intermedios	734

Nota: En la tabla se puede observar los datos estadísticos emitidos por Porechop al realizar la limpieza/trimado de la muestra SRR21422415 *Vibrio sp* CCB-PB317 secuenciada mediante tecnología Nanopore.

Una vez concluida la limpieza de la muestra de Nanopore, se procedió a realizar el ensamblaje de la misma usando el archivo fasto generado por Porechop a través de Canu y Flye.

1) Canu es un algoritmo que permite corregir, recortar y ensamblar genomas obtenidos a través de secuenciación de lectura larga. La fase inicial mejorará la calidad de las bases, la segunda fase recortará las lecturas en el segmento que parece ser una secuencia con alta calidad. La fase de ensamblaje ordenará las lecturas en contigs en formato fasta, generará secuencias de consenso y creará gráficos de caminos alternativos, es decir, se obtienen tres archivos de salida (Koren et al., 2017).

2) Flye es un algoritmo de ensamblaje de lectura larga que genera rutas arbitrarias en un gráfico repetido desconocido, llamado *disjointigs* y construye un gráfico repetido preciso a partir de estos *disjointigs* (Kolmogorov et al., 2019).

La decisión de usar dos algoritmos de ensamblaje de *novo* se realizó pensando en poder obtener el mejor ensamblaje en términos de contigüidad y completitud funcional. Luego de realizar los dos ensamblajes, los mismos fueron evaluados usando QUAST el cual permite evaluar la calidad del genoma ensamblado usando diversas métricas, además funciona con o sin genoma de referencia, este último siendo nuestro caso. Los resultados estadísticos de los ensamblajes a través de QUAST se puede observar en la Tabla 5.

Tabla 5

Tabla comparativa de la evaluación de la calidad entre los ensamblajes realizados con Canu y Flye mediante la herramienta QUAST

Referencia estadística	***canu_ensamblaje***	***flye_ensamblaje***
# contigs	8	4
# contigs (>= 0 bp)	8	4
# contigs (>= 1000 bp)	8	4
# contigs (>= 5000 bp)	7	4
# contigs (>= 10000 bp)	5	4
# contigs (>= 25000 bp)	3	4

# contigs (>= 50000 bp)	3	3
Largest contig	3075685	3118351
Total length	5196194	5175891
Total length (>= 0 bp)	5196194	5175891
Total length (>= 1000 bp)	5196194	5175891
Total length (>= 5000 bp)	5192017	5175891
Total length (>= 10000 bp)	5176037	5175891
Total length (>= 25000 bp)	5152727	5175891
Total length (>= 50000 bp)	5152727	5127870
N50	3075685	3118351
N90	1839558	1830892
auN	2482709	2532992
L50	1	1
L90	2	2

GC (%)	44.92	44.87
Mismatches		
# N's per 100 kbp	0	0
# N's	0	0

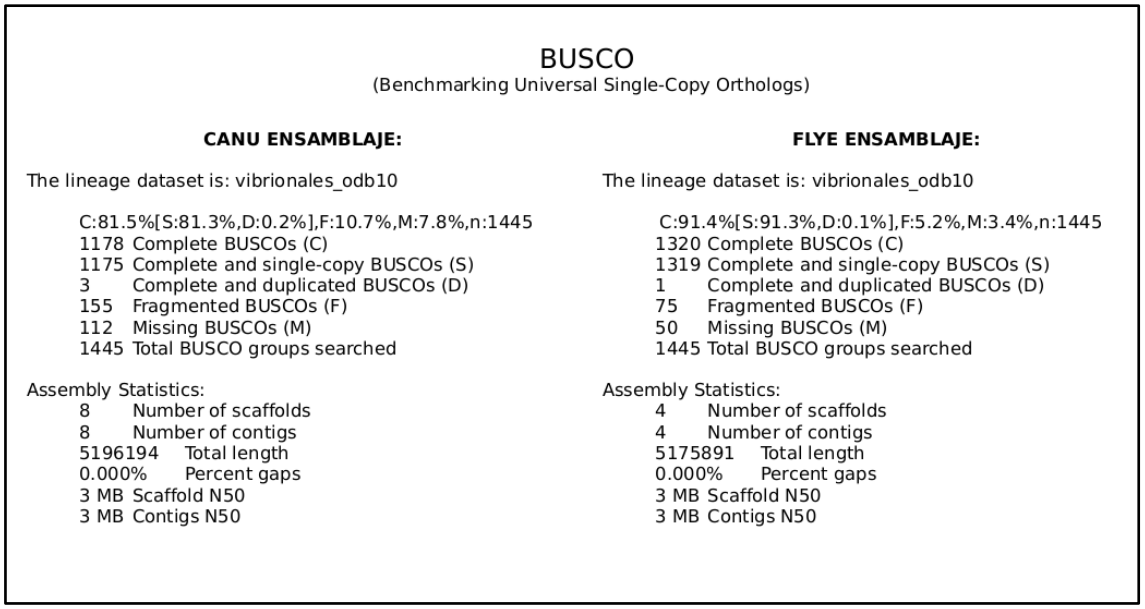
Nota: Esta tabla permite la comparación rápida entre las distintas estadísticas emitidas por la herramienta QUAST la cual permite evaluar la calidad de ensamblajes genómicos.

Los dos ensamblajes parecen tener métricas similares, sin embargo, cabe resaltar que Flye ensambló todo el genoma en solo 4 contigs a diferencia de Canu que lo realizó en 8. Tener menor cantidad de contigs determina que el genoma se encuentra estructurado en menor cantidad de partes. Así mismo el contig más largo reportado lo obtuvo Flye con 3.118.351 bp a comparación de Canu con 3.075.685 bp, el total de longitud del ensamblaje con Flye fue 5.175.891 bp a comparación de Canu con 5.196.194 bp. Lo anterior indica que Flye logró ensamblar el genoma completo de manera más contigua y eliminando zonas repetidas o ambiguas. Otra métrica resaltante fue el N50 el cual para Flye fue de 3,118,351 bp en comparación con 3,075,685 bp obtenido con Canu, siendo mejor el N50 generado por Flye el cual comprende el 60.2% del total del ensamblaje, en comparación con el 59.2% obtenido por Canu.

Una vez determinado que el mejor ensamblaje según la evaluación realizada a través de QUAST fue el generado por Flye, se procedió a evaluar los mismos a través de BUSCO. Los resultados pueden ser observados en la Figura 6.

Figura 6

Comparación de las estadísticas de los ensamblajes Canu y Flye mediante BUSCO



Nota: En esta figura se puede observar la comparación de las diversas métricas en cada ensamblaje realizado mediante la herramienta BUSCO, la cual considero el dataset vibrionales_odb10 como el más óptimo para realizar la evaluación de genes altamente conservados.

Para los dos ensamblajes BUSCO logro agruparlos en el linaje de vibrionales, sin embargo, al momento de comparar los mismos se puede evidenciar que el ensamblaje realizado con Flye logro métricas superiores como un 91.4% de Complete BUSCOs (C) a comparación de 81.5% obtenido por Canu, así mismo Flye presentó un 5.2% de secuencias fragmentadas (F) y un 3.4% de secuencias no encontradas (M), a su vez que Canu presentó 10.7% (F) y 7.8% (M). Esto indica que a pesar de que los dos ensamblajes evaluados por QUAST presentaron métricas relativamente similares, usando BUSCO se evidencia una notable diferencia a la hora de identificación de genes altamente conservados en los ensamblajes para el dataset vibrionales_odb10, ubicando al ensamblaje realizado por Flye como el ensamblaje más óptimo para continuar el pipeline (Figura 1).

Para mejorar la calidad del mejor ensamblaje obtenido (Flye), se procedió a realizar un pulido (*Polishing*) a través de Pilon (Walker et al., 2014) usando las secuencias forward y reverse SRR21422416 de Illumina. Estas lecturas presentan un índice de calidad alto con un Phred Score de 30 (Figura 4). Para realizar el proceso de pulido fue necesario generar las entradas de Pilon las cuales consistían en el mapeo e indexado de lecturas cortas al genoma ensamblado (Flye) con las herramientas Bowtie2 (Langmead & Salzberg, 2012a) y Samtools (Walker et al., 2014), para una posterior sustitución basada en la identificación de errores con el software Pilon.

Como archivo de salida generado por Pilon, se obtuvo un archivo fasta el cual es el ensamblaje de Flye pulido (flye_pilon). Para determinar si el ensamblaje de Flye pulido presentaba mejores métricas de calidad que el no pulido, se realizó una evaluación de calidad usando QUAST y BUSCO. Los resultados se observan en la Tabla 6.

Tabla 6

Tabla comparativa de las estadísticas del ensamblaje sin pulir junto con el ensamblaje pulido mediante la herramienta QUAST

Referencia estadística	***flye_ensamblaje***	***flye_pilon***
# contigs	4	4
# contigs (>= 0 bp)	4	4
# contigs (>= 1000 bp)	4	4
# contigs (>= 5000 bp)	4	4

# contigs (>= 10000 bp)	4	4
# contigs (>= 25000 bp)	4	4
# contigs (>= 50000 bp)	3	3
Largest contig	3118351	3118981
Total length	5175891	5176947
Total length (>= 0 bp)	5175891	5176947
Total length (>= 1000 bp)	5175891	5176947
Total length (>= 5000 bp)	5175891	5176947
Total length (>= 10000 bp)	5175891	5176947
Total length (>= 25000 bp)	5175891	5176947
Total length (>= 50000 bp)	5127870	5128911
N50	3118351	3118981
N90	1830892	1831277
auN	2532992	2533509

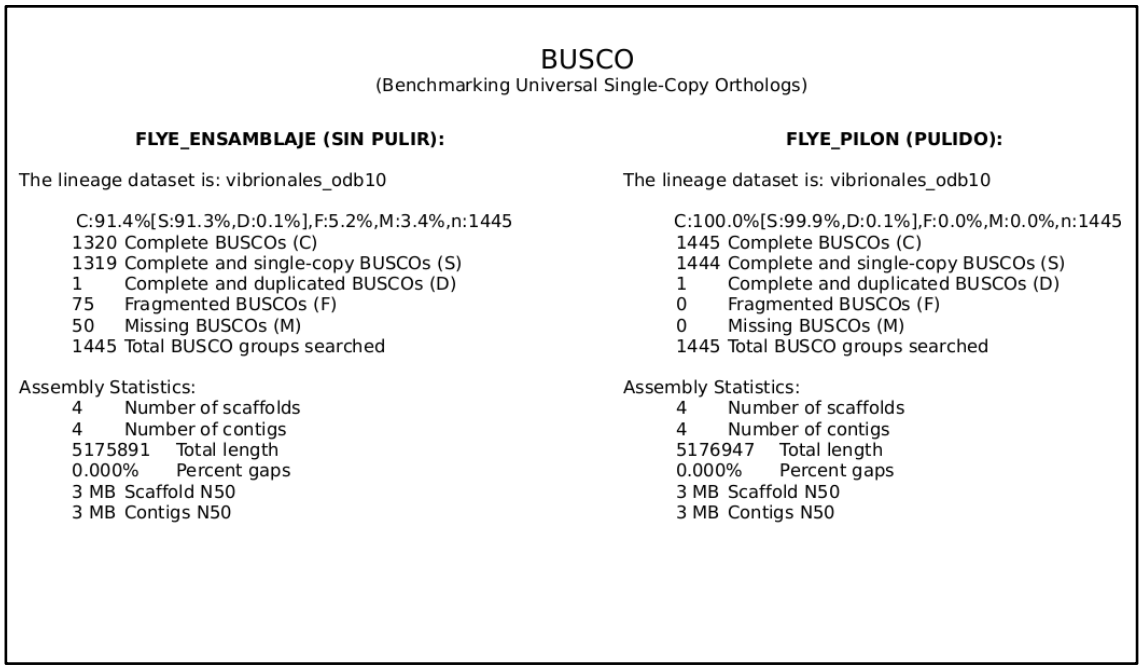
L50	1	1
L90	2	2
GC (%)	44.87	44.87
Mismatches		
# N's per 100 kbp	0	0
# N's	0	0

Nota: Se puede observar todas las métricas proporcionadas por la herramienta QUAST para evaluar la calidad del ensamblaje Flye sin pulir vs el ensamblaje Flye pulido (con Pilon) a fin de poder realizar una comparación completa de la calidad entre los mismos.

Como se puede observar el ensamblaje de Flye pulido (flye_pilon) mejoró sustancialmente en comparación con el ensamblaje de Flye sin pulir (flye_ensamblaje). Entre las métricas a destacar se encuentran, el contig más largo ensamblado con flye_pilon fue 3,118,981 bp siendo mayor al obtenido por flye_ensamblaje con 3,118,351 bp, el total de longitud ensamblado con flye_pilon fue de 5,176,947 bp a comparación del obtenido con flye_ensamblaje de 5,175,891 bp, y un N50 para flye_pilon con 3,118,981 pb y para flye_ensamblaje de 3,118,351 pb, siendo el primero mayor, lo que indica que el proceso de pulido realizado al ensamblaje de Flye agregó pares de bases de las secuencias de Illumina que mejoraron la longitud general de los contigs, y obtuvo mejores estadísticas en comparación con el ensamblaje de Flye sin pulir.

Figura 7

Comparación de las estadísticas del ensamblaje sin pulir junto con el ensamblaje pulido mediante la herramienta BUSCO



Nota: En esta figura se puede observar la comparación de las diversas métricas proporcionado por la herramienta BUSCO en el dataset vibrionales_odb10 para el ensamblaje de Flye sin pulir y el ensamblaje Flye pulido.

Mediante la evaluación realizado por BUSCO con el linaje de vibrionales_odb10, se encontró que el ensamblaje flye_pilon aumentó el número de Complete BUSCOs en 100.0% a comparación del flye_ensamblaje con 91.4%. Lo que resalta una mejoría en la identificación de genes altamente conservados mediante BUSCO del ensamblaje original por medio del proceso de pulido (*Polishing*) mediante el software de Pilon.

Una vez evaluada la calidad de nuestro ensamblaje final (pulido), se procedió a ingresar el archivo fasta en el Bacterial and Viral Bioinformatics Resource Center (BV-BRC) para su análisis mediante la

herramienta Análisis Integral del Genoma (CGA) dicha herramienta es considerada un metaservicio simplificado de análisis integral del genoma admitiendo secuencias ya ensambladas, proporcionando una descripción completa del genoma analizado. El resultado incluye una evaluación de la calidad del genoma, genes relacionados con resistencia antimicrobiana (AMR), predicciones de fenotipos, genes especializados, descripción general del subsistema, identificación de las secuencias genómicas más cercanas representado por un árbol filogenético (Olson et al., 2022); a continuación, se describen los resultados emitidos por el Análisis Integral del Genoma (CGA) para nuestro genoma ensamblado.

8.1 ANÁLISIS INTEGRAL DEL GENOMA (CGA)

Se cargo el genoma ensamblado de Flye pulido al servicio de Análisis Integral del Genoma en PATRIC (Wattam et al., 2017). Según las estadísticas de anotación y una comparación con otros genomas en PATRIC dentro de esta misma especie, este genoma es de buena calidad. Los detalles del análisis, incluidos los genes de interés (genes especializados), una categorización funcional (subsistemas) y un árbol filogenético (análisis filogenético) se proporcionan a continuación.

8.1.1 Anotación del genoma

El genoma de Vibrionales vibrionales_pilon se anotó utilizando el kit de herramientas RAST (RASTtk) (Brettin et al., 2015) y asignó un identificador de genoma único de 135623.175. La taxonomía de este genoma es:

organismos celulares > Bacterias > Proteobacterias > Gammaproteobacterias > Vibrionales

El genoma anotado se constituye de 4 Contigs (Tabla 7), una longitud del genoma de 5,176,947 bp, contenido de GC 44.86% y un N50 de 3118981.

Tabla 7

Estadísticas de los contigs ensamblados

Descripción	***Contenido %GC***	***Longitud (bp)***
contig_2_pilon	44.72	3118981
contig_3_pilon	44.95	1831277
contig_5_pilon	46.45	48036
contig_6_pilon	46.21	178653

Nota: en la tabla se puede observar las características estadísticas por cada contig ensamblado, los cuales servirán para ensamblar todo el genoma posteriormente.

Este genoma tiene 4,791 secuencias de codificación de proteínas (CDS), 125 genes de ARN de transferencia (ARNt) y 33 genes de ARN ribosómico (ARNr). Las características anotadas se resumen en la Tabla 8.

Tabla 8

Características generales del genoma anotado

CDS	4,791
tARN	125
Regiones repetitivas	100
rARN	33

Nota: La tabla muestra las diversas características generales del genoma anotado a partir del ensamblaje de Flye pulido mediante el Análisis integral del Genoma (CGA)

La anotación incluía 1112 proteínas hipotéticas y 3679 proteínas con asignaciones funcionales (Tabla 9). Se encontraron 1086 proteínas con números de la Comisión de Enzimas (EC). Las enzimas se consideran el grupo más grande y diverso de todas las proteínas e intervienen en todas las reacciones químicas del metabolismo de todos los organismos, jugando un papel crucial en la regulación metabólica dentro de la célula. Con el desarrollo y progreso de proyectos de genómica y metabolómica estructural y funcional, la recopilación, accesibilidad y procesamiento sistemáticos de datos de enzimas se vuelven aún más importantes para analizar y comprender los procesos biológicos. Por medio de la base de datos BRaunschweig ENzyme DAtabase (BRENDA) se pueden asignar funciones a las proteínas encontradas (Schomburg et al., 2004). Se encontraron 905 con asignaciones de Gene Ontology (GO) que difieren de las anteriores en que el sistema EC no aborda la clasificación de proteínas no enzimáticas (Ashburner et al., 2000), y 794 proteínas que se asignaron a las vías KEGG, es decir, las funciones de las proteínas encontrados se asocian con grupos ortólogos y se almacenan en la base de datos KEGG Orthology (KO) para su uso ((Kanehisa et al., 2016).

La anotación PATRIC incluye dos tipos de familias de proteínas (Davis et al., 2016). Este método utiliza las asignaciones de funciones basadas en k-mer disponibles a través de RAST (Overbeek et al., 2014) para guiar rápidamente la formación de familias, y luego diferencia los grupos basados en funciones en familias utilizando un algoritmo de clúster de Markov (MCL). Se encontró que este genoma tiene 0 proteínas que pertenecen a las familias de proteínas específicas de género (PLFams) esto se debe a que no fue especificado la especie al momento de hacer la entrada de información para el análisis, y 4648 proteínas que pertenecen a las familias de proteínas de género cruzado (PGFams): las familias de proteínas de géneros cruzados se calculan agrupando proteínas representativas de las familias específicas

de género con criterios ligeramente relajados (MCL = 1,1). Esto permite que los homólogos distantes o de géneros cruzados se agrupen, lo cual es necesario para respaldar el análisis comparativo de géneros cruzados en todos los genomas microbianos, permitiendo agrupar nuestro genoma problema en un grupo de género e incluso especies cercanas.

Tabla 9

Características a nivel proteico del genoma anotado

Proteínas hipotéticas	1,112
Proteínas con asignaciones funcionales	3,679
Proteínas con asignaciones de números EC	1,086
Proteínas con asignaciones GO	905
Proteínas con asignaciones de rutas	794
Proteínas con asignaciones de familias específicas de género (PLfam) PATRIC	0
Proteínas con asignaciones de familia de género cruzado PATRIC (PGfam)	4,648

Nota: La tabla resumen la cantidad de proteínas encontradas en base a diversas clasificaciones al realizar la anotación del genoma a partir del ensamblaje de Flye pulido mediante el Análisis integral del Genoma (CGA)

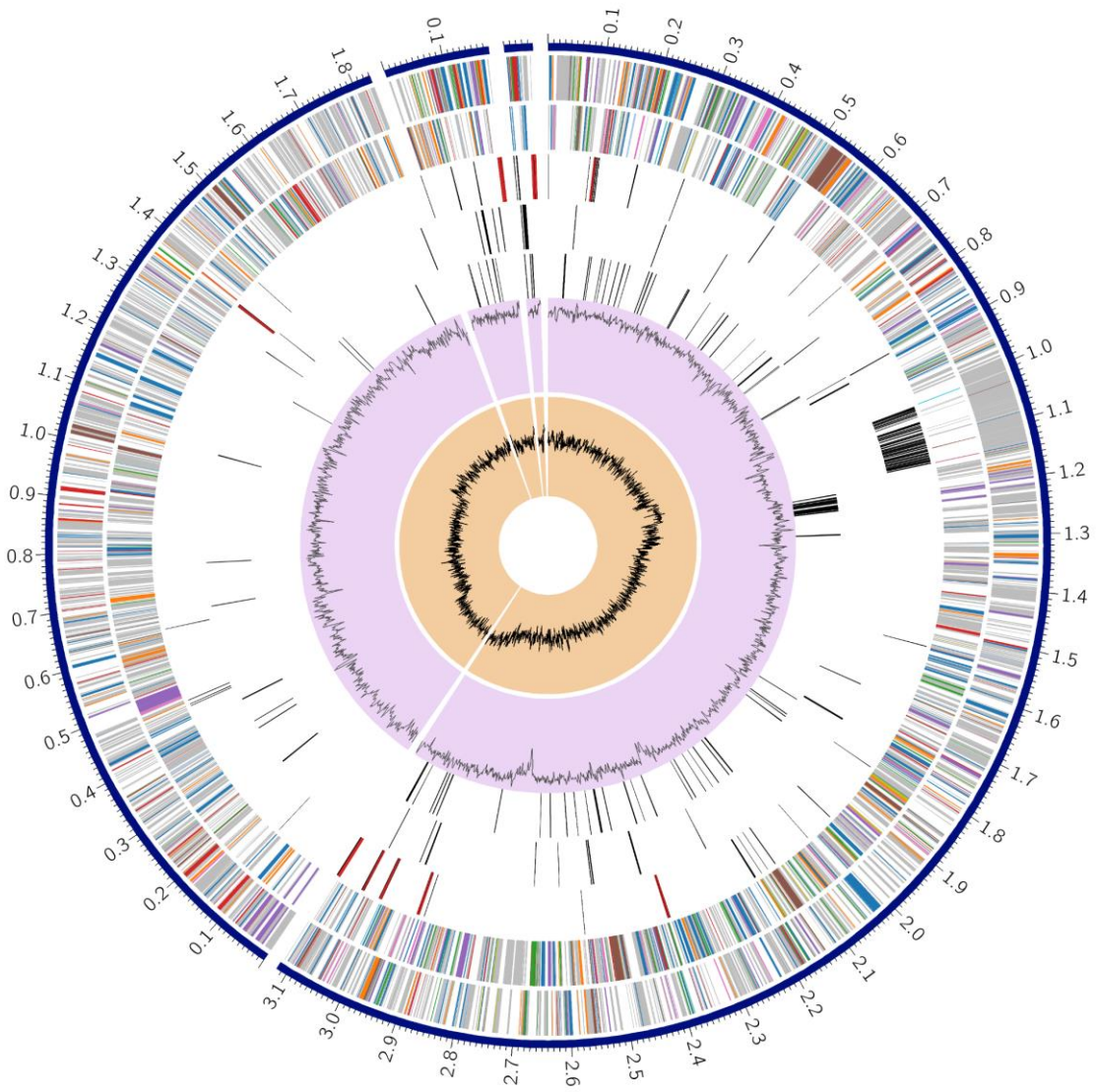
El CGA proporciona una visualización gráfica circular de la distribución de las anotaciones del genoma (Figura 8). Esto incluye, de los anillos externos a los internos, los contigs que representan los cromosomas, región codificante (CDS) en la cadena directa (forward), CDS en la cadena inversa (reverse), genes de ARN, CDS con genes de resistencia antimicrobianos conocidos, CDS con homología para factores de virulencia, contenido de GC y sesgo de GC. Los colores del CDS en la cadena directa e inversa indican el subsistema al que pertenecen estos genes (Figura 9).

8.1.2 Anotación de subsistemas

Un subsistema es un conjunto de proteínas que juntas implementan un proceso biológico específico o un complejo estructural (Overbeek et al., 2005) y la anotación incluye un análisis de los subsistemas del genoma y además se proporciona una descripción general de los subsistemas de este genoma que están activos, en la Figura 9 se puede apreciar las superclases de los subsistemas y una indicación del número de subsistemas dentro de esa superclase (primer número) y el número de genes anotados que forman parte de la superclase (segundo número).

Figura 8

Visualización gráfica circular de la distribución de las anotaciones del genoma

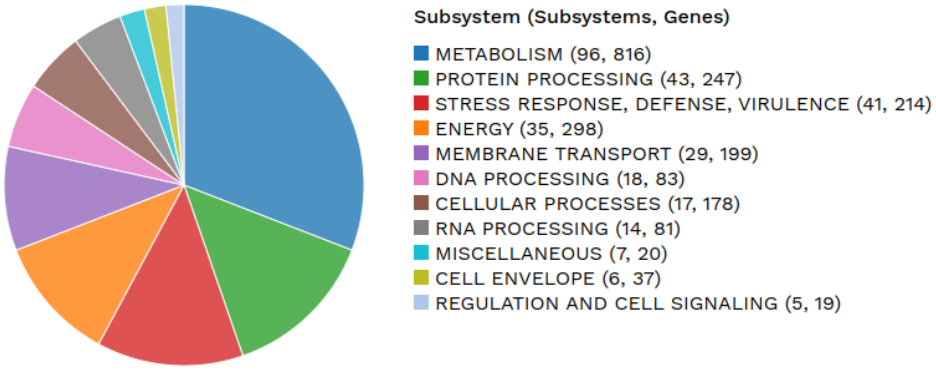


Nota: El gráfico ilustra una representación gráfica circular del genoma anotado a partir del ensamblaje de Flye pulido mediante el Análisis integral del Genoma (CGA), donde se puede apreciar en la parte más exterior (línea color azul oscuro) dos secciones de gran tamaño que corresponde a los dos cromosomas principales de *Vibrio sp*. y dos posibles plásmidos, así mismo los diversos niveles de anotación realizados.

Nota: En la siguiente figura se puede correlacionar los colores de los CDS en la cadena directa e inversa que indica el subsistema al que pertenecen estos genes.

Figura 9

Descripción general de los subsistemas del genoma por colores



Nota: Se presenta una representación gráfica mediante colores de los subsistemas encontrados en el genoma anotado a partir del ensamblaje de Flye pulido mediante el Análisis integral del Genoma (CGA)

8.1.3 Genes especiales

Muchos de los genes anotados tienen homología con diversos genes que el BV-BRC considera especiales (Tabla 10) en este caso los grupos de genes especiales son los siguientes:

Genes que conceden capacidades de resistencia a antibióticos (AMR): donde el BV-BRC separa según la base de datos donde son hallados estos genes, en el primer caso hace referencia a la base de datos integral de investigación de antibióticos (CARD) detectando 8 genes, la cual integra datos moleculares y de secuencia dispares, proporcionando un principio de organización único en forma de ontología de resistencia a antibióticos (ARO) (McArthur et al., 2013). Posteriormente, la Base de datos nacional de organismos resistentes a los antibióticos (NDARO) del NCBI detectando 3 genes y por último

la base de datos del Centro de Integración de Recursos de PathoSystems (PATRIC) (Davis et al., 2019) la cual ha sido integrada al BV-BRC, encontrando 33 genes.

Genes que son objetivos farmacológicos (Drug Target): la base de datos de objetivos terapéuticos (TTD) brinda información sobre los principales objetivos de los medicamentos aprobados y experimentales, y a través de la búsqueda de similitud de objetivos mediante el algoritmo BLAST, determina el nivel de similitud entre la secuencia de una proteína de entrada y la secuencia de cada una de las entradas diana TTD (Zhu et al., 2010). Con esta base de datos encontramos 5 genes a diferencia del DrugBank que detectó 35 genes lo que denota la mayor magnitud de datos alojada en esta base de datos la cual desde 2014, DrugBank 4.0 mejoró sustancialmente su tecnología bioinformática para la detección de drogas, sus metabolitos y sus efectos posteriores. En particular, se han realizado mejoras significativas y adiciones a gran escala en las áreas de QSAR (relaciones cuantitativas entre estructura y actividad), ADMET (absorción, distribución, metabolismo, excreción y toxicidad), farmacometabolómica y farmacogenómica (Law et al., 2014), siendo así la base de datos de mayor relevancia en el ámbito farmacológico.

Genes que conceden capacidades de sintetizar proteínas transportadoras (Transportador): La base de datos de clasificación de transportadores (TCDB) es una base de datos de referencia de libre acceso para la investigación de proteínas de transporte, que proporciona información estructural, funcional, mecánica, evolutiva y de enfermedades sobre transportadores de organismos de todo tipo, en nuestro caso encontrando 64 genes directamente relacionados con la síntesis de distintas proteínas transportadoras (Saier et al., 2016).

Genes que conceden factores de virulencia (Virulence Factor): Los factores de virulencia (VF) son una clase importante de productos génicos que ayudan a los patógenos a evadir los mecanismos defensivos de un huésped específico para establecer la infección en una condición ambiental específica,

lo que resulta en un estado de enfermedad para ese huésped. El estudio de las FV facilita nuestra comprensión de la patogenicidad y los mecanismos de las enfermedades infecciosas, en el 2015 se desarrolló una biblioteca de factores de virulencia bacteriana (VF) altamente seleccionada en PATRIC para apoyar la investigación de enfermedades infecciosas. Esta base de datos encontró para nuestro genoma 15 genes directamente relacionados con factores de virulencia (C. Mao et al., 2015). Así mismo, la base de datos de factores de virulencia (VFDB) se dedica a proporcionar conocimientos actualizados sobre los factores de virulencia (FV) de varios patógenos bacterianos humanos. Desde sus inicios, la VFDB ha servido como un depósito completo de FV bacterianas durante más de una década (L. Chen et al., 2016), la misma encontró 33 FV. Por último, se encuentra Victors la cual es una novedosa base de datos integral, curada manualmente, siendo un gran recurso de análisis para VF de patógenos que causan enfermedades infecciosas en humanos y animales. Además, la base de datos Victors VF tiene un mayor énfasis en la anotación manual y el análisis de VF. Cada uno de los VF en Victors está asociado con evidencia anotada manualmente de al menos una cita revisada por pares (Sayers et al., 2019), la misma determinó 66 genes relacionados con FV para nuestro genoma. Los factores de virulencia encontrados por la base de datos de PATRIC_VF pueden ser visualizados en la Tabla 11

Tabla 10

Genes especiales encontrados en el genoma anotado

Categoría	***Fuente***	***Genes***
Resistencia antibiótica	CARD	8
Resistencia antibiótica	NDARO	3
Resistencia antibiótica	PATRIC	33

Objetivo farmacológico	DrugBank	35
Objetivo farmacológico	TTD	5
Transportador	TCDB	64
Factor de virulencia	PATRIC_VF	15
Factor de virulencia	VFDB	33
Factor de virulencia	Victors	66

Nota: La tabla detalla la categoría de genes especiales identificados en el Análisis Integral del Genoma (CGA), seguido de la fuente o base de datos donde se encontraron dichos genes y por último la cantidad de genes encontrados por cada categoría y fuente.

Tabla 11

Genes relacionados con factores de virulencia encontrados en la base de datos de PATRIC_VF

Gen	***Producto***	***Clasificación***
dksA	RNA polymerase-binding transcription factor DksA	Propagación de célula a célula, Estrés, Regulación de la expresión génica
guaB	Inosine-5'-monophosphate dehydrogenase (EC 1.1.1.205) / CBS domain	Virulencia e Invasión
carB	Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)	Virulencia
ruvB	Holliday junction ATP-dependent DNA helicase RuvB (EC 3.6.4.12)	Supervivencia intracelular y replicación

grxD	Monothiol glutaredoxin GrxD	Regulación de la expresión génica
fur	Ferric uptake regulation protein FUR	Virulencia
carA	Carbamoyl-phosphate synthase small chain (EC 6.3.5.5)	Virulencia
ompR	Two-component system response regulator OmpR	Virulencia
clpB	Chaperone protein ClpB (ATP-dependent unfoldase)	Virulencia
trxA	Thioredoxin	Modula la respuesta inmune del huésped
hfq	RNA-binding protein Hfq	Regulación de la expresión génica
cap	Cyclic AMP receptor protein	Supervivencia intracelular y replicación
rpoS	RNA polymerase sigma factor RpoS	Virulencia
gmhA	D-sedoheptulose 7-phosphate isomerase (EC 5.3.1.28)	Virulencia
trpB	Tryptophan synthase beta chain (EC 4.2.1.20)	Virulencia

Nota: Se identifican los genes a través del nombre, producto o función y clasificación de los mismos asociados a factores de virulencia encontrados por la base de datos PATRIC_VF descrita anteriormente.

8.1.4 Genes de resistencia a los antimicrobianos

El Servicio de anotación del genoma en PATRIC utiliza el método de detección de genes AMR basado en k-mer, que utiliza la colección curada de PATRIC de variantes representativas de secuencias de genes AMR (Wattam et al., 2017) asignando a cada gen AMR una anotación funcional, un amplio mecanismo de resistencia a los antibióticos, una clase de fármaco y, en algunos casos, un antibiótico específico al que confiere resistencia. Se debe tener en cuenta que la presencia de genes relacionados con

AMR (incluso de longitud completa) en un genoma determinado no implica directamente un fenotipo resistente a los antibióticos. Es importante considerar los mecanismos AMR específicos y especialmente la ausencia/presencia de mutaciones SNP que transmitan resistencia. A continuación, se proporciona un resumen de los genes AMR anotados en este genoma correspondiente a 33 genes y el mecanismo AMR correspondiente en l**a** Tabla 12.

Tabla 12

Genes relacionados con mecanismos de resistencia antimicrobiana

Mecanismo AMR	***Genes***
Enzima de activación de antibióticos	KatG
Enzima de inactivación de antibióticos	CARB family
Blanco de antibióticos en especies susceptibles	Alr, Ddl, dxr, EF-G, EF-Tu, folA, Dfr, folP, gyrA, gyrB, Iso-tRNA, kasA, MurA, rho, rpoB, rpoC, S10p, S12p
Proteína de protección de la diana antibiótica	QnrB family
Proteína de reemplazo diana de antibióticos	fabV
Bomba de eflujo que confiere resistencia a los antibióticos	MdtL, Tet(35), TolC/OpmH

Gen que confiere resistencia por ausencia	gidB
Proteína que altera la carga de la pared celular que confiere resistencia a los antibióticos	GdpD, PgsA
Regulador que modula la expresión de genes de resistencia a antibióticos	H-NS, OxyR

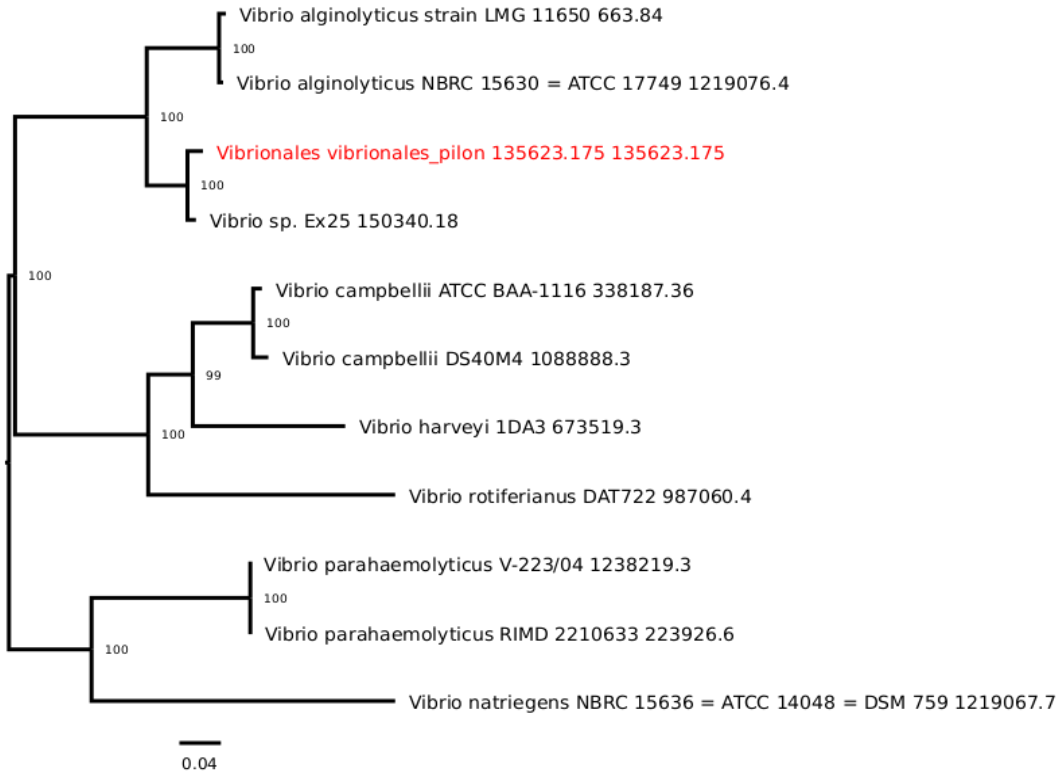
Nota: La tabla clasifica los genes según su mecanismo de resistencia antimicrobiana encontrados en la base de datos PATRIC utilizando el método de detección de genes AMR basado en k-mer (Wattam et al., 2017) del genoma anotado a partir del ensamblaje de Flye pulido mediante el Análisis integral del Genoma (CGA).

8.1.5 Análisis filogenético

Se realiza mediante una colaboración entre el Centro Nacional de Información Biotecnológica (NCBI) y PATRIC que seleccionan y categorizan manualmente los genomas de referencia, que se consideran de alta calidad e importancia para la comunidad de investigación incluyéndose en el análisis filogenético que forma parte del informe. La referencia más cercana y los genomas representativos fueron identificados por Mash/MinHash (Ondov et al., 2016). En base a las Familias de proteínas globales PATRIC (PGFams) (Davis et al., 2016) fueron seleccionados de estos genomas representativos para determinar la ubicación filogenética de nuestro genoma. Las secuencias de proteínas de estas familias se alinearon con MUSCLE (Edgar, 2004) y los nucleótidos para cada una de esas secuencias se asignaron a la alineación de proteínas. El conjunto de alineaciones de aminoácidos y nucleótidos se concatenaron usando una matriz de datos mediante RaxML (Stamatakis, 2014) con fast-bootstrapping (Stamatakis et al., 2008), para generar los valores de soporte del árbol filogenético. En la Figura 10 se observa el árbol filogenético obtenido, marcando en rojo nuestro genoma. Con este resultado, vemos que nuestra muestra es más cercana a *Vibrio alginolyticus.*

Figura 10

Árbol filogenético para el genoma ensamblado con Flye y pulido con Pilon



Nota: El CGA ubica la muestra proveniente del SRA de *Vibrio sp.* ensamblada con Flye y pulido con Pilon (Vibrionales vibrionales_pilon 135623.175 135623.175) al genoma más cercano encontrado mediante el NCBI y PATRIC (Wattam et al., 2017) que corresponde a *Vibrio alginolyticus*.

9 DISCUSIÓN

Los ensamblajes de genomas se pueden realizar mediante lecturas largas de secuenciación de Oxford Nanopore Technologies (ONT). Esta tecnología se basa en el principio de secuenciación de hebras en nanoporos en tiempo real, que aumenta la longitud de lectura entre 100 y 1000 veces. En nuestro caso, utilizando este tipo de datos, obtuvimos contigs ensamblados por Flye del tamaño completo de los cromosomas, permitiendo abarcar regiones repetitivas mucho más largas. Para los genomas bacterianos, las lecturas de Nanopore suelen ser más largas que la repetición más grande del genoma (Lu et al., 2016), lo que hace posibles ensamblajes completos (un contig por replicón) a diferencia de las lecturas cortas de Illumina las cuales no pueden resolver secuencias repetidas más largas que las lecturas de secuenciación. Al usar solo lecturas cortas, pueden fallar al ensamblar elementos genéticos móviles, duplicaciones genómicas y sobre todo, puede ser imposible saber a partir de un ensamblaje de lectura corta de Illumina incompleto si los genes de interés, como los genes de resistencia a los antimicrobianos (AMR) residen en el cromosoma o en un plásmido (Z. Chen et al., 2021).

Por otro lado, los errores sistemáticos en lecturas largas pueden conducir a cientos de errores residuales, la mayoría de los cuales son indels en secuencias de homopolímeros (Rubio et al., 2020), debido a que las lecturas largas de Oxford Nanopore tienen una alta tasa de error del 10 % al 15 % (12.2% en nuestro caso) a diferencia de la las lecturas de Illumina que tienen menos de 1% de error. Ahora bien, cuando se ensambla con lecturas de Nanopore pueden ocurrir errores en las secuencias de codificación de proteínas, que provocan cambios de marco en el marco de lectura abierto, lo que genera problemas con la anotación del genoma y limita la utilidad de los ensamblajes de lectura largos, mientras que las lecturas cortas de Illumina no sufren los mismos errores en las secuencias de homopolímeros (Wick & Holt, 2022). Por tal motivo los enfoques integrados que utilizan de manera eficiente las lecturas largas y cortas pueden superar este problema (Walker et al., 2014). Una solución es pulir el ensamblaje de lectura larga Oxford Nanopore con lecturas cortas Illumina de alta precisión para mejorar la precisión como se

realizó en este caso. Este enfoque no se basa en ensambladores independientes específicos, pero se puede lograr usando ensambladores de lectura larga para crear un borrador de ensamblaje como es el caso de Canu y Flye, seguido de un paso de pulido con lecturas cortas de Illumina mediante Pilon. En esta aproximación, comenzamos con un genoma de entrada y usamos evidencia de alineamientos de lecturas cortas para identificar diferencias específicas del genoma de entrada respaldado por los datos de secuenciación y aplicando cambios al ensamblaje de entrada dando como resultados un ensamblaje mejorado. Dado lo anterior, el pipeline desarrollado cumple con los objetivos de crear un pipeline eficiente que permita obtener un ensamblaje de alta calidad y además destaca la importancia de las lecturas cortas de Illumina para mejorar la integridad y precisión del genoma.

Entre las ventajas del pipeline está el hecho de contar con múltiples pasos que permiten mejorar la calidad de lecturas largas de Nanopore con la herramienta Porechop y posteriormente ir evaluando la calidad post ensamblaje hasta determinar cuál es el mejor de los ensambladores de *novo* usados (Canu y Flye) y proceder con un pulido (Pilon) que de igual manera es evaluado con dos herramientas (QUAST y BUSCO). EL pipeline puede ser ejecutado sin grandes requerimientos computacionales, especialmente para el caso de bacterias, lo que se traduce en un pipeline completo, actualizado y sobre todo práctico que permitirá al investigador realizar una anotación funcional completa. Adicionalmente, cabe resaltar que el pipeline no fue diseñado solo para bacterias del género *Vibrio sp.*, y podría ser replicado con otros patógenos bacterianos como fue demostrado en estudios similares, donde realizaron ensamblajes pulidos para distintos géneros bacterianos teniendo un resultado similar al encontrado en este estudio, donde el pulido con Pilon mejoró la integridad del genoma y la precisión de los ensamblajes de lectura larga de Nanopore, debido a las correcciones de errores de secuencia (Z. Chen et al., 2021; Taylor et al., 2019; Walker et al., 2014). Lo anterior sugiere que el pulido es una estrategia eficaz para generar ensamblajes bacterianos precisos.

Nuestro genoma contó con 4 contigs, donde el contig más grande ha sido mencionado en otras investigaciones como el cromosoma I grande y el segundo contig más largo como cromosoma II pequeño. En esta muestra dicho cromosoma I grande contó con 3,118,981 bp (3,12 Mb) y 44.72% de contenido de GC y el cromosoma II pequeño contó con 1,831,277 bp (1,83 Mb) y 44.95% de contenido de GC, al compararlo con otros estudios de pangenomas de *Vibrio alginolyticus*, el primero reporta que los genomas contenía un cromosoma I grande de 3,27 a 3,40 Mb y un cromosoma II pequeño de 1,81 a 1,89 Mb y el contenido de GC osciló entre 44,5 y 44,8 % (Xue et al., 2022), así mismo el siguiente estudio denota que los genomas analizados contenían un cromosoma I de 3,47 Mb y un cromosoma 2 de 1,88 Mb y 44% de contenido de GC en promedio (Chibani et al., 2020); lo cual refleja que nuestro genoma tiene un cromosoma I grande de menor tamaño al reportado en dichos estudios, mientras que el cromosoma II pequeño y el porcentaje de contenido de GC son consistentes con los mismos.

Adicionalmente nuestro genoma contiene 2 contig pequeños, uno con 178,653 pb (178 kb) y 46.21% de contenido de GC y otro con 48,036 pb (48 kb) y 46.45% de contenido de GC; esto puede ser explicado por hallazgos encontrados en los estudios antes mencionados donde se reportó que los genomas analizados tenían uno o dos plásmidos extracromosómicos (34,0–93,4 kb), con contenidos de GC que oscilaban entre 39,7 y 46,2 % (Xue et al., 2022); mientras que la otra investigación logró determinar que los genomas poseen plásmidos con un tamaño de 0,9 a 290 kb (Chibani et al., 2020). Lo cual podría explicar la presencia de dichos contigs en nuestro genoma y catalogarlos como plásmidos extracromosómicos.

Adicionalmente el Análisis Integral del Genoma anotó para nuestro genoma 4791 secuencias de codificación de proteínas (CDS), 125 genes de ARN de transferencia (ARNt) y 33 genes de ARN ribosómico (ARNr) lo cual al compararlo con el estudio realizado por Xue y col., en el 2022 donde el promedio para 7

especies de *Vibrio alginolyticus* obtuvieron 4649 CDS, ARNt 129 ARNt y 37 ARNr se puede observar características de anotación bastante semejantes entre nuestra muestra y el pangenoma realizado.

Al realizar el Análisis Integral del Genoma (CGA) mediante el BV-BRC, se determinó que la especie más cercana a nivel de genómico de nuestra muestra es *Vibrio alginolyticus* la cual es una bacteria gramnegativa que se encuentra de forma ubicua en los hábitats acuáticos y marinos y plantea riesgos considerables para la salud de los animales marinos y los seres humanos a través de las cadenas alimentarias locales (F. Mao et al., 2021). Según un número creciente de estudios, *V. alginolyticus* no solo se limita a infectar especies marinas como ostras, meros y camarones (Oberbeckmann et al., 2011; Wang et al., 2016) pero también está emergiendo como un patógeno oportunista que infecta a los humanos. La infección por *V. alginolyticus* se asocia principalmente con trastornos inflamatorios en humanos, por ejemplo, gastroenteritis, otitis media, otitis externa y septicemia (Feingold & Kumar, 2004; Gaüzère et al., 2016; Issack et al., 2008; Uh et al., 2001).

Por tal motivo concuerda que la muestra analizada, perteneciente a la especie de *V. alginolyticus,* haya sido anotado con factores de virulencia para 3 bases de datos distintas lo cual refleja la capacidad patógena de la misma, por ejemplo el gen dksA o Factor de transcripción de unión a ARN polimerasa DksA, ha sido reportado para *Vibrio cholerae* y *Shigella flexneri*, el cual en un entorno carente de nutrientes experimentan un cambio metabólico llamado respuesta estricta donde se produce un aumento en la expresión de dksA, favoreciendo la propagación de célula a célula, motilidad, formación de biocapa, resistencia al estrés oxidativo y regulación de la expresión génica según las condiciones ambientales (Mogull et al., 2001; Sofia & Dziejman, 2021).

Igualmente se identificaron genes asociados a resistencia antimicrobiana en 3 bases de datos distintas (CARD, NDARO y PATRIC) y se clasificaron en base al mecanismo de acción aquellos genes encontrados por PATRIC encontrando 33 genes. Esto es de gran relevancia ya que los antibióticos son compuestos químicos que inhiben la multiplicación bacteriana (bacteriostáticos) o matan las especies bacterianas (bactericidas), son en su mayoría productos naturales, sintetizados por las especies bacterianas o fúngicas como armas químicas para matar otros microbios en el microambiente cercano y mantener el equilibrio en las comunidades microbianas en los ecosistemas naturales (Skliros et al., 2021). Siendo los objetivos de los antibióticos generalmente exclusivos de las bacterias o significativamente diferentes de sus contrapartes eucariotas y son esenciales para la inhibición del crecimiento y la supervivencia de las bacterias. La mayoría de los antibióticos funcionan ampliamente inhibiendo la síntesis o el ensamblaje de la pared celular, interrumpiendo la integridad de la membrana celular, evitando la síntesis de ADN, ARN y proteínas, interrumpiendo rutas metabólicas celulares esenciales. La especificidad de los antibióticos hacia las maquinarias microbianas es la clave para su uso en la práctica clínica para prevenir y curar infecciones microbianas y su implementación terapéutica revolucionó la historia de la medicina. Sin embargo, desde su introducción, se reconoció que los antibióticos tienen dos propiedades distintas: la primera, inhibir el crecimiento de microbios al interferir con las funciones esenciales o procesos celulares y la segunda, e indeseable, que promueve la aparición de patógenos AMR al proporcionar un entorno adecuado para su crecimiento mediante la eliminación de variantes sensibles (Das et al., 2020); entre algunos de los genes asociados a resistencia antimicrobiana (AMR) como el gen KatG Catalase-peroxidase (EC 1.11.1.21) identificado en *Mycobacterium tuberculosis*, el cual codifica el enzima catalasa-peroxidasa e impiden la transformación del fármaco isoniacida en su principio activo, inhibiendo su acción antimicrobiana (Morlock et al., 2003) lo que podría indicar una transferencia de genes horizontal entre *Mycobacterium* y *Vibrio*, otro gen encontrado es el CARB family que es un beta-lactamase carbenicillin hydrolyzing (EC 3.5.2.6) que ha sido descrito anteriormente en *Vibrio cholerae*

muestran que es una enzima hidrolizante de carbenicilina que induce resistencia mediada por plásmidos a los β-lactámicos que son susceptibles a los efectos de los inhibidores de las β-lactamasas (Choury et al., 1999), otro gen encontrado de gran relevancia es el QnrB family que induce resistencia a las quinolonas (ciprofloxacina, gatifloxacina, levofloxacina, moxifloxacina, ácido nalidíxico, norfloxacina, esparfloxacina) las cuales son potentes agentes antibacterianos que tienen como objetivo las siguientes topoisomerasas: ADN girasa y la topoisomerasa IV en bacterias. Estudios previos han demostrado que la resistencia a las quinolonas se originan por mutaciones en los genes cromosómicos de *Escherichia coli* (Tran & Jacoby, 2002) y el gen gidB (EC 2.1.1.170) que codifica una 7-metilguanosina (m(7)G) metiltransferasa conservada específica para el ARNr 16S, confiriendo resistencia a la estreptomicina (Okamoto et al., 2007).

Por tal motivo la presencia de genes AMR en nuestra muestra sumado a los genes que promueven factores de virulencia da un claro panorama de la capacidad patogénica y de supervivencia de nuestra muestra perteneciente a la especie *V. alginolyticus*.

10 CONCLUSIÓN

El advenimiento de las tecnologías de secuenciación de alto rendimiento ha revolucionado el campo de la genómica, permitiendo la generación rápida y rentable de datos de secuencias a escala del genoma con una resolución y precisión muy alta, sin embargo, uno de los mayores retos consiste en aprovechar correctamente los datos secuenciados. Es claro que no existe una sola herramienta bioinformática que permita realizar todo el procesamiento de la muestra desde su análisis de calidad hasta la anotación funcional por medio de la consola GNU/Linux. Por tal motivo el diseño de un pipeline como el estructurado en esta investigación permite servir de guía para futuras investigaciones en el campo de la ciencia aplicada a la genómica, donde sea necesario extraer la mayor cantidad de información de los datos genómicos. Concluimos que este pipeline cumplió con los objetivos planteados y se logró realizar un ensamblaje de *novo* con lecturas largas de Nanopore. Según las herramientas de calidad de ensamblaje QUAST y BUSCO se determinó que el mejor ensamblaje fue realizado por Flye y además se procedió al pulido del mismo con lecturas cortas de Illumina usando Pilon. Obtuvimos un ensamblaje de excelente calidad con 4 contigs que revelan la presencia de dos cromosomas, un cromosoma I grande con 44.72% de contenido GC y una longitud de 3,118,981 pb junto con un cromosoma II pequeño con 44.95% de contenido de GC y una longitud de 1,831,277 pb; además se encontraron dos regiones de más pequeñas con 46.45% y 46.21% de contenido de GC y una longitud de 48,036 pb y 178,653 pb respectivamente las cuales son asociadas a plásmidos, teniendo el genoma una longitud total de 5,176,947 pb, un N50 de 3,118,981 y con un 100.0% de complementariedad en los conjuntos de ortólogos universales de copia única de evaluación comparativa de OrthoDB para el dataset vibrionales_odb10 de BUSCO. La anotación funcional realizada mediante el servicio de Análisis Integral de Genoma (CGA) del servidor BV-BRC presenta un amplio panorama de información genómica que va desde la estadística estructural del genoma, representaciones gráficas del genoma anotado, genes especiales incluidos factores de virulencia, caracterización de genes asociados a mecanismos de resistencia antimicrobiana (AMR) y por último un

análisis filogenético que permitió correlacionar la muestra con *Vibrio alginolyticus*, especie oportunista conocida por su alto grado de patogenicidad en humanos.

11 MATERIAL SUPLEMENTARIO

11.1 Códigos de programación utilizados para cada herramienta bioinformática

Creación de ambiente conda

```
conda config --add channels conda-forge
conda config --add channels bioconda
conda config --show
conda create -n <nombre_env>
```

Instalación de paquetes dentro del env

```
conda install -c bioconda <nombre_del_paquete>
```

Porechop
Se instala mediante Conda
Su input son las lecturas crudas fastq

```
porechop -i <archivo entrada> -o <archivo salida> --format fasta -t 10
```

Canu
Se instala mediante Conda

```
canu [-haplotype|-correct|-trim] [-s <assembly-specifications-file>] -p <assembly-
prefix> -d <assembly-directory> genomeSize=<number>[g|m|k] [-trimmed|-untrimmed|-
raw|-corrected] [-pacbio|-nanopore|-pacbio-hifi] *fastq
```

Flye
Se instala mediante Conda

```
flye --nano-raw <archivo entrada fastq> --genome-size <tamaño genoma> --out <dir/
archivo salida>
```

El archivo de salida de los ensambladores son archivos fasta *(PRIMERA ENTRADA PARA PILON)*

QUAST
Se descarga el quast-5.2.0.tar.gz de la página
https://sourceforge.net/projects/quast/files/
una vez descargado se ubica la ruta y se ejecuta el .py seguido del ensamblaje a evaluar y -o el nombre de la carpeta de salida

```
<ruta completa>/quast-5.2.0/./quast.py <archivo entrada> -o <archivo salida>
```

BUSCO
se instala mediante Conda

```
busco -i <archivo entrada> --auto-lineage-prok -o <archivo salida> -m genome
```

Bowtie2 y Samtools
Se instalan mediante Conda

```
bowtie2-build -f <archivo entrada-ensamblaje.fasta> <archivo salida-index.fa>
```

```
bowtie2 -x <archivo entrada index.fa> -1 <archivo entrada-lecturas illumina forward>
-2 <archivo entrada-lecturas illumina reverse> -S <archivo salida indexado.sam>
```

Transformar el archivo sam a archivo bam

```
samtools view -S -b indexado.sam > indexado.bam
```

Ordenar el bam a sorted.bam *(SEGUNDA ENTRADA PARA PILON)*

```
samtools sort indexado.bam -o indexado.sorted.bam
```

Indexar el sorted.bam

```
samtools index indexado.sorted.bam
```

Crear archivo fasta.fai

```
samtools faidx <archivo entrada-ensamblaje.fasta>
```

Pilon

Descargar el jar

https://github.com/broadinstitute/pilon/releases/tag/v1.2

Es necesario estar ubicado en la carpeta donde se crearon los 2 INPUTS

```
java -Xmx16G -jar pilon-1.24.jar --genome <archivo entrada ensamblaje.fasta> --frags
indexado.sorted.bam --output <nombre ensamblaje pulido> --fix all --mindepth 0.5 --
changes --verbose --threads 4
```

El output de Pilon es un archivo fasta

12 REFERENCIAS BIBLIOGRAFICAS

Andrews, S. (2010). *A Quality Control Tool for High Throughput Sequence Data*.

Arias, F. (2006). *El proyecto de investigación: Introducción a la investigación científica* (5.ª ed.). Editorial Episteme.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25-29. https://doi.org/10.1038/75556

Baker-Austin, C., Oliver, J. D., Alam, M., Ali, A., Waldor, M. K., Qadri, F., & Martinez-Urtaza, J. (2018). Vibrio spp. infections. *Nature Reviews Disease Primers*, *4*(1), 1-19. https://doi.org/10.1038/s41572-018-0005-8

Baumann, P., Baumann, L., Woolkalis, M. J., & Bang, S. S. (1983). EVOLUTIONARY RELATIONSHIPS IN VIBRIO AND PHOTOBACTERIUM: A BASIS FOR A NATURAL CLASSIFICATION. *Annual Review of Microbiology*, *37*(1), 369-398. https://doi.org/10.1146/annurev.mi.37.100183.002101

Bonenfant, Q., Noé, L., & Touzet, H. (2022). Porechop_ABI: discovering unknown adapters in ONT sequencing reads for downstream trimming. *BioRxiv*. https://doi.org/10.1101/2022.07.07.499093

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., Stevens, R., Vonstein, V., Wattam, A. R., & Xia, F. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, *5*(1), 8365. https://doi.org/10.1038/srep08365

Caro, M., Rios, A., & Calero, C. (2005). Análisis y revisión de la literatura en el contexto de proyectos de fin de carrera: Una propuesta. *Revista Electrónica de La Sociedad Chilena de Ciencia de La Computación*, *6*(1). https://www.researchgate.net/publication/251671565

Castillo, D., Vandieken, V., Engelen, B., Engelhardt, T., & Middelboe, M. (2018). Draft Genome Sequences of Six Vibrio diazotrophicus Strains Isolated from Deep Subsurface Sediments of the Baltic Sea. *Genome Announcements*, *6*(10), e00081-18. https://doi.org/10.1128/genomeA.00081-18

Chen, L., Zheng, D., Liu, B., Yang, J., & Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Research*, *44*(D1), D694-D697. https://doi.org/10.1093/nar/gkv1239

Chen, Z., Erickson, D. L., & Meng, J. (2021). Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses. *Genomics*, *113*(3), 1366-1377. https://doi.org/https://doi.org/10.1016/j.ygeno.2021.03.018

Chibani, C. M., Roth, O., Liesegang, H., & Wendling, C. C. (2020). Genomic variation among closely related Vibrio alginolyticus strains is located on mobile genetic elements. *BMC Genomics*, *21*(1), 354. https://doi.org/10.1186/s12864-020-6735-5

Choury, L., Rald Aubert, G., Szajnert, M., Azibi, K., Delpech, M., Rard Paul, G., & Alger-Ouest, C. (1999). *Characterization and Nucleotide Sequence of CARB-6, a New Carbenicillin-Hydrolyzing-Lactamase from Vibrio cholerae* (Vol. 43, Issue 2).

Cumbicos, D., & Ruiz, J. (2018). Proliferation cycle of bacterial strains Vibrio ssp and Pseudomonas ssp of young black seashell (Anadara tuberculosa). *ESPACIOS*, *39*(13), 14.

D'Ancona, M., & Angeles, M. (2012). *Fundamentos y aplicaciones en metodología cuantitativa* (Sintesis, Ed.).

Das, B., Verma, J., Kumar, P., Ghosh, A., & Ramamurthy, T. (2020). Antibiotic resistance in Vibrio cholerae: Understanding the ecology of resistance genes and mechanisms. *Vaccine*, *38*, A83-A92. https://doi.org/https://doi.org/10.1016/j.vaccine.2019.06.031

Davis, J. J., Gerdes, S., Olsen, G. J., Olson, R., Pusch, G. D., Shukla, M., Vonstein, V., Wattam, A. R., & Yoo, H. (2016). PATtyFams: Protein Families for the Microbial Genomes in the PATRIC Database. *Frontiers in Microbiology*, *7*. https://doi.org/10.3389/fmicb.2016.00118

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E. M., Gabbard, J. L., Gerdes, S., Guard, A., Kenyon, R. W., Machi, D., Mao, C., Murphy-Olson, D., Nguyen, M., Nordberg, E. K., … Stevens, R. (2019). The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz943

de Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, *34*(15), 2666-2669. https://doi.org/10.1093/bioinformatics/bty149

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792-1797. https://doi.org/10.1093/nar/gkh340

Faruque, S. M., Biswas, K., Nashir Udden, S. M., Shafi Ahmad, Q., Sack, D. A., Balakrish Nair, G., & Mekalanos, J. J. (2006). *Transmissibility of cholera: In vivo-formed biofilms and their relationship to infectivity and persistence in the environment*. www.pnas.orgcgidoi10.1073pnas.0601277103

Feingold, M. H., & Kumar, M. L. (2004). Otitis Media Associated With Vibrio alginolyticus in a Child With Pressure-Equalizing Tubes. *The Pediatric Infectious Disease Journal*, *23*(5).

https://journals.lww.com/pidj/Fulltext/2004/05000/Otitis_Media_Associated_With_Vibrio_alginolyticus.23.aspx

Gaüzère, B.-A., Chanareille, P., & Vandroux, D. (2016). Septicémie à Vibrio alginolyticus au décours d'une presque noyade à La Réunion (océan Indien). *Bulletin de La Société de Pathologie Exotique*, *109*(3), 151-154. https://doi.org/10.1007/s13149-016-0505-2

Greene, J. M., Collins, F., Lefkowitz, E. J., Roos, D., Scheuermann, R. H., Sobral, B., Stevens, R., White, O., & di Francesco, V. (2007). National Institute of Allergy and Infectious Diseases Bioinformatics Resource Centers: New Assets for Pathogen Informatics. *Infection and Immunity*, *75*(7), 3212-3219. https://doi.org/10.1128/IAI.00105-07

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075. https://doi.org/10.1093/bioinformatics/btt086

Hernández-Sampieri, R., Fernández-Collado, C., & Baptista-Lucio, M. (2014). *Metodologia de la Investigacion* (6ta ed.). McGRAW-HILL.

Issack, M. I., Appiah, D., Rassoul, A., Unuth, M. N., & Unuth-Lutchun, N. (2008). Extraintestinal Vibrio infections in Mauritius. *The Journal of Infection in Developing Countries*, *2*(05). https://doi.org/10.3855/jidc.205

Janecko, N., Bloomfield, S. J., Palau, R., & Mather, A. E. (2021). Whole genome sequencing reveals great diversity of vibrio spp in prawns at retail. *Microbial Genomics*, *7*(9). https://doi.org/10.1099/mgen.0.000647

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457-D462. https://doi.org/10.1093/nar/gkv1070

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540-546. https://doi.org/10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research*, *27*(5), 722-736. https://doi.org/10.1101/gr.215087.116

Langmead, B., & Salzberg, S. L. (2012a). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923

Langmead, B., & Salzberg, S. L. (2012b). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z. T., Han, B., Zhou, Y., & Wishart, D. S. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, *42*(D1), D1091-D1097. https://doi.org/10.1093/nar/gkt1068

Lekshmi, N., Joseph, I., Ramamurthy, T., & Thomas, S. (2018). Changing facades of vibrio cholerae: An enigma in the epidemiology of cholera. En *Indian Journal of Medical Research* (Vol. 147, Issue February, pp. 133-141). Indian Council of Medical Research. https://doi.org/10.4103/ijmr.IJMR_280_17

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265-279. https://doi.org/https://doi.org/10.1016/j.gpb.2016.05.004

Lukjancenko, O., & Ussery, D. W. (2014). Vibrio chromosome-specific families. *Frontiers in Microbiology*, *5*(MAR). https://doi.org/10.3389/fmicb.2014.00073

Mao, C., Abraham, D., Wattam, A. R., Wilson, M. J. C., Shukla, M., Yoo, H. S., & Sobral, B. W. (2015). Curation, integration and visualization of bacterial virulence factors in PATRIC. *Bioinformatics*, *31*(2), 252-258. https://doi.org/10.1093/bioinformatics/btu631

Mao, F., Liu, K., Wong, N. K., Zhang, X., Yi, W., Xiang, Z., Xiao, S., Yu, Z., & Zhang, Y. (2021). Virulence of Vibrio alginolyticus Accentuates Apoptosis and Immune Rigor in the Oyster Crassostrea hongkongensis. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.746017

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., de Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J. v., Spanogiannopoulos, P., … Wright, G. D. (2013). The Comprehensive Antibiotic Resistance Database. *Antimicrobial Agents and Chemotherapy*, *57*(7), 3348-3357. https://doi.org/10.1128/AAC.00419-13

Mogull, S. A., Runyen-Janecky, L. J., Hong, M., & Payne, S. M. (2001). dksA is required for intercellular spread of Shigella flexneri via an RpoS-independent mechanism. *Infection and Immunity*, *69*(9), 5742-5751. https://doi.org/10.1128/IAI.69.9.5742-5751.2001

Morlock, G. P., Metchock, B., Sikes, D., Crawford, J. T., & Cooksey, R. C. (2003). ethA, inhA, and katG Loci of Ethionamide-Resistant Clinical Mycobacterium tuberculosis Isolates. *Antimicrobial Agents and Chemotherapy*, *47*(12), 3799-3805. https://doi.org/10.1128/AAC.47.12.3799-3805.2003

Morris, J. G. (2003). Cholera and Other Types of Vibriosis: A Story of Human Pandemics and Oysters on the Half Shell. En *July) • FOOD SAFETY*. https://academic.oup.com/cid/article/37/2/272/302936

Noreña, L., Alcaraz-Moreno, N., Rojas, J. G., & Rebolledo-Malpica, D. (2012). Applicability of the Criteria of Rigor and Ethics in Qualitative Research. *AÑO*, *12*, 263-274.

Oberbeckmann, S., Wichels, A., Wiltshire, K. H., & Gerdts, G. (2011). Occurrence of Vibrio parahaemolyticus and Vibrio alginolyticus in the German Bight over a seasonal cycle. *Antonie van Leeuwenhoek*, *100*(2), 291-307. https://doi.org/10.1007/s10482-011-9586-x

Okamoto, S., Tamaru, A., Nakajima, C., Nishimura, K., Tanaka, Y., Tokuyama, S., Suzuki, Y., & Ochi, K. (2007). Loss of a conserved 7-methylguanosine modification in 16S rRNA confers low-level streptomycin resistance in bacteria. *Molecular Microbiology*, *63*(4), 1096-1106. https://doi.org/10.1111/j.1365-2958.2006.05585.x

Olson, R. D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, J. J., Dempsey, D. M., Dickerman, A., Dietrich, E. M., Kenyon, R. W., Kuscuoglu, M., Lefkowitz, E. J., Lu, J., Machi, D., Macken, C., Mao, C., Niewiadomska, A., Nguyen, M., Olsen, G. J., … Stevens, R. L. (2022). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkac1003

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. v, Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., … Vonstein, V. (2005). The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Research*, *33*(17), 5691-5702. https://doi.org/10.1093/nar/gki866

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., & Stevens, R. (2014). The SEED and the Rapid

Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, *42*(D1), D206-D214. https://doi.org/10.1093/nar/gkt1226

Percival, S., & Williams, D. (2014). Chapter Twelve - Vibrio. En *Microbiology of Waterborne Diseases* (2.ª ed., Vol. 1, pp. 237-248).

Pérez-Duque, A., Gonzalez-Muñoz, A., Arboleda-Valencia, J., Vivas-Aguas, L. J., Córdoba-Meza, T., Rodriguez-Rey, G. T., Díaz-Guevara, P., Martinez-Urtaza, J., & Wiesner-Reyes, M. (2021). Comparative Genomics of Clinical and Environmental Isolates of Vibrio spp. of Colombia: Implications of Traits Associated with Virulence and Resistance. *Pathogens*, *10*(12). https://doi.org/10.3390/pathogens10121605

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C. N., Dietrich, J., Klem, E. B., & Scheuermann, R. H. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, *40*(D1), D593-D598. https://doi.org/10.1093/nar/gkr859

Rubio, S., Pacheco-Orozco, R. A., Gómez, A. M., Perdomo, S., & García-Robles, R. (2020). Secuenciación de nueva generación (NGS) de ADN: presente y futuro en la práctica clínica. *Universitas Médica*, *61*(2). https://doi.org/10.11144/Javeriana.umed61-2.sngs

Saier, M. H., Reddy, V. S., Tsu, B. v., Ahmed, M. S., Li, C., & Moreno-Hagelsieb, G. (2016). The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Research*, *44*(D1), D372-D379. https://doi.org/10.1093/nar/gkv1103

Sayers, S., Li, L., Ong, E., Deng, S., Fu, G., Lin, Y., Yang, B., Zhang, S., Fa, Z., Zhao, B., Xiang, Z., Li, Y., Zhao, X.-M., Olszewski, M. A., Chen, L., & He, Y. (2019). Victors: a web-based knowledge base of virulence

factors in human and animal pathogens. *Nucleic Acids Research*, *47*(D1), D693-D700. https://doi.org/10.1093/nar/gky999

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., & Schomburg, D. (2004). BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, *32*(suppl_1), D431-D433. https://doi.org/10.1093/nar/gkh081

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. v., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212. https://doi.org/10.1093/bioinformatics/btv351

Skliros, D., Kalatzis, P. G., Kalloniati, C., Komaitis, F., Papathanasiou, S., Kouri, E. D., Udvardi, M. K., Kokkari, C., Katharios, P., & Flemetakis, E. (2021). The Development of Bacteriophage Resistance in Vibrio alginolyticus Depends on a Complex Metabolic Adaptation Strategy. *Viruses*, *13*(4), 656. https://doi.org/10.3390/v13040656

Sofia, M. K., & Dziejman, M. (2021). Dksa coordinates bile-mediated regulation of virulence-associated phenotypes in type three secretion system-positive vibrio cholerae. *Microbiology (United Kingdom)*, *167*(2). https://doi.org/10.1099/mic.0.001006

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Systematic Biology*, *57*(5), 758-771. https://doi.org/10.1080/10635150802429642

Taylor, T. L., Volkening, J. D., DeJesus, E., Simmons, M., Dimitrov, K. M., Tillman, G. E., Suarez, D. L., & Afonso, C. L. (2019). Rapid, multiplexed, whole genome and plasmid sequencing of foodborne

pathogens using long-read nanopore technology. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-52424-x

Thompson, F., Iida, T., & Swings, J. (2004). Biodiversity of Vibrios. *Microbiology and Molecular Biology Reviews*, *68*(3), 403-431. https://doi.org/10.1128/MMBR.68.3.403-431.2004

Tran, J. H., & Jacoby, G. A. (2002). Mechanism of plasmid-mediated quinolone resistance. *Proceedings of the National Academy of Sciences*, *99*(8), 5638-5642. https://doi.org/10.1073/pnas.082092899

Uh, Y., Park, J. S., Hwang, G. Y., Jang, I. H., Yoon, K. J., Park, H. C., & Hwang, S. O. (2001). Vibrio alginolyticus acute gastroenteritis: Report of two cases [2]. En *Clinical Microbiology and Infection* (Vol. 7, Issue 2, pp. 104-106). Blackwell Publishing Ltd. https://doi.org/10.1046/j.1469-0691.2001.00207.x

Vadillo, S., Piriz, S., & Mateos, E. (2002). *Manual de microbiologia veterinaria* (1.ª ed.). McGraw-Hill Interamericana de España.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wang, Z., Wang, B., Chen, G., Jian, J., Lu, Y., Xu, Y., & Wu, Z. (2016). Transcriptome analysis of the pearl oyster (Pinctada fucata) hemocytes in response to Vibrio alginolyticus infection. *Gene*, *575*(2, Part 2), 421-428. https://doi.org/https://doi.org/10.1016/j.gene.2015.09.014

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E. M., Disz, T., Gabbard, J. L., Gerdes, S., Henry, C. S., Kenyon, R. W., Machi, D., Mao, C., Nordberg, E. K., Olsen, G. J., Murphy-Olson, D. E., Olson, R., … Stevens, R. L. (2017). Improvements to PATRIC, the all-

bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Research*, *45*(D1), D535-D542. https://doi.org/10.1093/nar/gkw1017

Watve, S. S., Chande, A. T., Rishishwar, L., Mariño-Ramírez, L., Jordan, I. K., & Hammer, B. K. (2016). Whole-genome sequences of 26 Vibrio cholerae isolates. *Genome Announcements*, *4*(6). https://doi.org/10.1128/genomeA.01396-16

WHO. (2015). Cholera. *WEEKLY EPIDEMIOLOGICAL RECORD*, 517-528. http://www.who.int/wer2015,90,517-544No.40

WHO. (2017). Cholera vaccines: WHO position paper. *Weekly Epidemiological Record*, *34*. http://www.who.

Wick, R. R., & Holt, K. E. (2022). Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLOS Computational Biology*, *18*(1), e1009802-. https://doi.org/10.1371/journal.pcbi.1009802

Xue, M., Huang, X., Xue, J., He, R., Liang, G., Liang, H., Liu, J., & Wen, C. (2022). Comparative Genomic Analysis of Seven Vibrio alginolyticus Strains Isolated From Shrimp Larviculture Water With Emphasis on Chitin Utilization. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.925747

Yang, C., Pei, X., Wu, Y., Yan, L., Yan, Y., Song, Y., Coyle, N. M., Martinez-Urtaza, J., Quince, C., Hu, Q., Jiang, M., Feil, E., Yang, D., Song, Y., Zhou, D., Yang, R., Falush, D., & Cui, Y. (2019). Recent mixing of Vibrio parahaemolyticus populations. *ISME Journal*, *13*(10), 2578-2588. https://doi.org/10.1038/s41396-019-0461-5

Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C. N., Lee, A. J., Li, X., Macken, C., Mahaffey, C., Pickett, B. E., Reardon, B., Smith, T., Stewart, L.,

Suloway, C., Sun, G., … Scheuermann, R. H. (2017). Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, *45*(D1), D466-D474. https://doi.org/10.1093/nar/gkw857

Zhu, F., Han, B., Kumar, P., Liu, X., Ma, X., Wei, X., Huang, L., Guo, Y., Han, L., Zheng, C., & Chen, Y. (2010). Update of TTD: Therapeutic Target Database. *Nucleic Acids Research*, *38*(suppl_1), D787-D791. https://doi.org/10.1093/nar/gkp1014