BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS QUÍMICAS

DEPARTAMENTO DE MICROBIOLOGÍA

# Ensamblaje del genoma y caracterización de *Bacillus paralicheniformis* proveniente de muestras ambientales de la Ciudad de Puebla

**TESIS**

PARA OBTENER EL GRADO DE LICENCIADO EN QUÍMICO FARMACOBIÓLOGO

PRESENTA

**JESÚS EMMANUEL MOYOTL GÓMEZ**

DIRECTOR DE TESIS

**M.C. ERICK ACOCAL JUÁREZ**

CODIRECTORA

**D. C. ANA MARTA DE LOS ÁNGELES LOBO SÁNCHEZ**

ASESOR EXTERNO

**D. C. MAURA CÁRDENAS GARCÍA**

Puebla, Pue. Enero 2026

# INDICE GENERAL

# INDICE DE TABLAS

# INDICE DE TABLAS

1. RESUMEN

Los bioaerosoles representan un componente relevante del ambiente urbano debido a su influencia en la salud pública y en los procesos ecológicos. Entre ellos, las bacterias del género *Bacillus* se caracterizan por su resistencia y capacidad de adaptación a condiciones atmosféricas variables. En este estudio se realizó el ensamblaje genómico y la caracterización comparativa de cuatro aislados ambientales de *Bacillus paralicheniformis* obtenidos del aire de la Ciudad de Puebla. Mediante un análisis bioinformático integral, que incluyó la evaluación de calidad ("FastQC"), filtrado ("Trimmomatic"), ensamblaje ("SPAdes"), anotación ("Prokka") y comparación genómica, se identificaron patrones de estabilidad y variabilidad en los genomas analizados. Los aislados presentaron un número de Secuencias Codificantes (CDS) estable (4305–4323), pero variaciones notables en genes de ARN, particularmente tRNA, lo que sugiere posibles adaptaciones a microambientes urbanos. Aproximadamente un tercio de los genes anotados correspondieron a proteínas hipotéticas, indicando la presencia de regiones funcionales aún no caracterizadas. El análisis comparativo reveló 32 CDS específicos, así como genes compartidos con longitudes variables, consistentes con un pangenoma abierto y con procesos de microevolución intraespecífica. Genes relacionados con la respiración celular, la partición cromosómica y el transporte de nutrientes destacaron como posibles mecanismos adaptativos. El análisis filogenético confirmó diversidad intraespecífica y relaciones evolutivas diferenciadas entre los aislados. En conjunto, los resultados evidencian que *B. paralicheniformis* presenta un genoma central conservado acompañado de regiones accesorias variables que favorecen su adaptación a las condiciones atmosféricas de la Ciudad de Puebla. Este estudio contribuye al entendimiento de la diversidad microbiana presente en bioaerosoles urbanos y demuestra la utilidad del análisis genómico para explorar su potencial ecológico y funcional.

## 2. INTRODUCCIÓN

El estudio de los bioaerosoles en entornos urbanos es fundamental para comprender su impacto tanto en la salud humana como en los procesos ecológicos que ocurren dentro de las ciudades. Diversas investigaciones, como la realizada por Zhao et al. (2022), han demostrado que estos microorganismos transportados por el aire pueden actuar como agentes infecciosos, alérgenos o desencadenantes de enfermedades respiratorias; sin embargo, también desempeñan funciones clave en los ciclos biogeoquímicos y en la dinámica ambiental. Su abundancia y distribución responden a factores como la variabilidad climática, la calidad del aire, la presencia de partículas suspendidas y la existencia de fuentes naturales y antropogénicas, lo que los convierte en indicadores relevantes del estado ecológico urbano. El análisis de microorganismos presentes en bioaerosoles permite identificar potenciales riesgos para la salud, así como comprender procesos de adaptación microbiana a condiciones ambientales fluctuantes. El estudio genómico, en particular, ofrece una herramienta poderosa para explorar la diversidad microbiana, sus capacidades funcionales y las estrategias evolutivas que favorecen su persistencia en la atmósfera urbana. El presente trabajo tuvo como objetivo realizar el ensamblaje genómico y determinar los genes diferenciales entre cuatro aislados bacterianos obtenidos del aire de la Ciudad de Puebla, identificados fenotípica y genotípicamente como *Bacillus paralicheniformis*. Para ello, se emplearon herramientas bioinformáticas especializadas como “FastQC”, “Trimmomatic”, “SPAdes” y “Prokka”, que permitieron evaluar la calidad de las secuencias, ensamblar los genomas, realizar la anotación funcional de los genes y comparar de manera precisa las características genómicas entre los aislados. El análisis realizado permitió obtener una caracterización genómica robusta que reveló patrones clave sobre la diversidad intraespecífica y las posibles adaptaciones funcionales de los aislados estudiados. La anotación realizada con “Prokka” mostró que el número de Secuencias Codificantes (CDS) se mantuvo relativamente estable entre las cuatro cepas, lo que indica la presencia de un genoma central altamente conservado en *B. paralicheniformis*. No obstante, se observaron variaciones notables en el número de genes de ARN, particularmente en genes de tRNA, lo que sugiere adaptaciones a microambientes específicos del aire urbano, posiblemente relacionadas con la eficiencia metabólica o la regulación de la traducción. Aproximadamente un tercio de los CDS en cada aislado correspondieron a proteínas hipotéticas, reflejando la presencia de regiones genómicas aún

no caracterizadas y evidenciando un potencial metabólico y funcional que no ha sido explorado previamente. El análisis comparativo reveló 32 CDS específicos distribuidos entre los aislados y un conjunto adicional de genes compartidos de manera parcial o con longitudes variables. Estos patrones son consistentes con la existencia de un pangenoma abierto en el género *Bacillus*, donde un núcleo genómico estable coexiste con regiones accesorias que otorgan plasticidad funcional y capacidad adaptativa. Entre los genes compartidos con tamaños variables destacaron aquellos relacionados con la respiración celular (subunidades del citocromo C oxidasa y la deshidrogenasa de NADH) y con la partición cromosómica (ParA y Smc); hallazgos que sugieren procesos de microevolución intraespecífica que permiten ajustar mecanismos de producción de energía y estabilidad genómica frente a condiciones cambiantes del ambiente atmosférico. Asimismo, se identificaron genes asociados al transporte de aminoácidos y metales, como permeasas GntP, que indican la capacidad de las cepas para responder a la disponibilidad irregular de nutrientes en el aire urbano. El análisis filogenético realizado con "IQ-TREE 2" mostró una clara separación entre los cuatro aislados, confirmando la presencia de diversidad intraespecífica. Las relaciones evolutivas observadas (la cercanía entre los aislados 3 y 4 y la divergencia del aislado 1) sugieren que estos microorganismos han experimentado procesos de diferenciación influenciados por mutaciones, recombinación o adaptación a microambientes atmosféricos específicos. Los resultados obtenidos respaldan la hipótesis de que los aislados ambientales de *B. paralicheniformis* presentan variaciones principalmente en genes accesorios, lo cual refleja su plasticidad genómica y su capacidad de adaptación a las condiciones atmosféricas de la Ciudad de Puebla. Esta investigación aporta evidencia sobre la diversidad y el potencial funcional de bacterias presentes en bioaerosoles urbanos y subraya la importancia de integrar enfoques microbiológicos, bioinformáticos y ecológicos para comprender la dinámica microbiana en ambientes urbanos.

## 3. MARCO TEÓRICO

Las interacciones de los microorganismos en la atmósfera representan un área de investigación cada vez más relevante en la microbiología. Estos microorganismos, que incluyen bacterias, hongos, virus y esporas, juegan un papel fundamental en diversos procesos biogeoquímicos, ya que, a través del aire pueden dispersarse por grandes distancias, lo que les permite interactuar con otros ecosistemas y organismos. Los genomas de estos microorganismos son muy diversos, con una amplia gama de adaptaciones que les permiten sobrevivir en un entorno tan variable y dinámico como la atmósfera. Algunas especies poseen características genéticas que les otorgan resistencia a condiciones extremas como la radiación ultravioleta, la desecación o las variaciones de temperatura, lo que les permite persistir y propagarse en la atmósfera. Las consecuencias ecológicas de estas interacciones permiten que los microorganismos aerotransportados pueden influir en procesos como la formación de nubes, la distribución de nutrientes y la fertilización de suelos, afectando así la biodiversidad y los ciclos biogeoquímicos. También pueden tener implicaciones para la salud humana, contribuyendo al desarrollo de enfermedades respiratorias, alergias y otras afecciones, especialmente cuando están presentes en concentraciones elevadas o en condiciones propicias para su propagación.

### 3.1. Bacterias aerotransportadas

Los microorganismos aerotransportados están ampliamente presentes y metabólicamente activos en la atmósfera y representan una parte importante de las partículas transportadas por el aire (*Jaenicke, 2005*). El aire generalmente se ha considerado solo un conducto para la vida microbiana terrestre y acuática; sin embargo, también es un hábitat para microorganismos como las bacterias con más de $1x10^4$ células bacterianas/$m^3$ y cientos de taxones únicos (*Burrows et al., 2009*). La estructura de las comunidades bacterianas parece perturbarse más fácilmente, debido a los procesos estocásticos, los factores meteorológicos y la calidad del aire; en las zonas urbanas, los impactos humanos debilitan la importancia relativa de las fuentes vegetales de bacterias transportadas por el aire y elevan la aparición de patógenos potenciales de origen antropogénico. Las bacterias aerotransportadas son los principales componentes de los bioaerosoles, por lo que juegan un papel importante en la salud de los humanos, animales y plantas. La presencia de las bacterias en los diferentes ambientes varía

en función de las características que presenta su hábitat, por ejemplo, en el suelo la diversidad microbiana está influenciada por los cambios de pH o la temperatura; en los ambientes marinos, el factor determinante es la salinidad. Sin embargo, los mecanismos que impulsan la dinámica de las comunidades bacterianas aerotransportadas aún no se han caracterizado a nivel mundial. Por lo tanto, es necesario realizar investigaciones sobre las estructuras de las comunidades microbianas, los patrones biogeográficos y los mecanismos impulsores a escala global para comprender los microbiomas atmosféricos (*Zhao et al., 2022*). Entre las bacterias aerotransportadas se encuentran *Mycobacterium tuberculosis, Streptococcus pneumoniae y Bacillus anthracis,* todos ellos agentes patógenos. Sin embargo, también es posible hallar microorganismos no patógenos u oportunistas suspendidos en el aire, cuya identificación y análisis genómico puede contribuir a comprender mejor la estructura de la comunidad bacteriana (*González, 2015*).

### 3.2. *Bacillus paralicheniformis*

*B. paralicheniformis* es una bacteria grampositiva, esporulante y aeróbica perteneciente a la familia *Bacillaceae* y al filo *Firmicutes*. Este microorganismo se encuentra ampliamente distribuido en diversos ambientes como suelos, agua dulce, sedimentos y la microbiota de ciertos animales y humanos. Su adaptabilidad y capacidad de supervivencia en condiciones extremas han captado el interés de la comunidad científica, especialmente por su potencial en aplicaciones biotecnológicas, industriales y médicas. Estas aplicaciones incluyen la producción de enzimas industriales, biopolímeros y compuestos antimicrobianos. A nivel genómico, la cepa *B. paralicheniformis* MDJK30 presenta un genoma de aproximadamente 4.2 millones de pares de bases (bp), un tamaño comparable al de otras especies cercanas del género *Bacillus*. Su contenido de guanina-citosina (GC) se sitúa en torno al 46–46.7 %, lo que evidencia una organización genética que favorece su diversidad funcional y adaptación a nichos ecológicos variados. Estudios recientes han identificado entre 4,200 y 4,300 genes en promedio en el genoma de esta especie, distribuidos entre genes codificantes y ARN estructurales. Además, la plasticidad genómica de *Bacillus paralicheniformis* está impulsada por elementos genéticos móviles, como plásmidos, transposones y secuencias de inserción, que facilitan la transferencia horizontal de genes y le otorgan ventajas adaptativas significativas (*Du et al.*, 2019).

### 3.2.1. Relación genómica de *Bacillus paralicheniformis* con otras especies

En el contexto de la taxonomía molecular, *B. paralicheniformis* guarda una estrecha relación con *B. licheniformis*, una especie ampliamente estudiada. Sin embargo, se pueden distinguir mediante análisis genómicos detallados. Por ejemplo, se han identificado diferencias en regiones específicas del genoma, incluyendo variaciones en secuencias repetitivas y elementos estructurales, así como la presencia de clústeres de genes secundarios. Entre estos, el clúster para la biosíntesis de fengicina, que involucra genes específicos como fenC y fenD, permite diferenciar entre ambas especies (*Du et al*., 2019).

### 3.2.2. Perspectiva evolutiva de *Bacillus paralicheniformis*

La diversidad genómica dentro de *B. paralicheniformis* refleja adaptaciones evolutivas que han favorecido su persistencia en distintos nichos ecológicos. Estas adaptaciones incluyen genes asociados a la esporulación, resistencia a condiciones ambientales hostiles y la producción de metabolitos secundarios. El análisis comparativo entre genomas de diferentes aislados ha revelado variaciones importantes en genes relacionados con la resistencia antimicrobiana y la biosíntesis de compuestos especializados, lo que resalta su potencial como modelo para estudios biotecnológicos y ecológicos (*Du et al*., 2019).

## 3.3. Análisis bioinformático

El análisis bioinformático constituye un pilar central en la caracterización genómica de bacterias como *B. paralicheniformis*. A través de este enfoque es posible evaluar la calidad de las lecturas, ensamblar genomas, identificar genes y comparar diferentes aislados para comprender su variabilidad genética y capacidades funcionales. Herramientas ampliamente utilizadas, como “FastQC” y “Trimmomatic”, permiten asegurar la calidad inicial de las secuencias, mientras que ensambladores como “SPAdes” facilitan la reconstrucción del genoma a partir de datos de secuenciación. Posteriormente, plataformas como “Prokka” y “RAST” permiten la anotación funcional, proporcionando una visión detallada del repertorio génico presente en cada cepa.

Un ejemplo reciente del uso de estas metodologías es el estudio de *B. paralicheniformis* AA1, aislada en la región sur de Sonora, México. En dicho trabajo, Chávez-Almanza *et al*.

aplicaron herramientas de ensamblaje y anotación comparables, logrando obtener un genoma superior a 4.3 Mb y caracterizar un conjunto diverso de genes asociados con metabolismo, estrés y producción de compuestos bioactivos. Además, el análisis comparativo con otras cepas del género evidenció la presencia de un número considerable de genes compartidos y, simultáneamente, variabilidad en el contenido génico entre aislados, lo que refuerza la utilidad del análisis bioinformático para comprender la plasticidad genómica dentro de esta especie (Chávez-Almanza et al., 2025).

Este tipo de investigaciones demuestra cómo la combinación de distintas herramientas bioinformáticas permite no solo caracterizar de manera integral a *B. paralicheniformis*, sino también situar nuevos aislados dentro de la diversidad genética de la especie. En conjunto, estas metodologías proporcionan una base sólida para explorar su potencial ecológico, evolutivo y biotecnológico.

Es importante señalar que todas las etapas del procesamiento bioinformático, incluyendo la evaluación de calidad, el recorte de secuencias, el ensamblaje y la anotación, se realizaron de manera independiente para cada una de las cuatro cepas analizadas. En ningún momento se mezclaron lecturas, *contigs* ni archivos de anotación en ninguna fase previa al análisis comparativo.

La integración de datos entre los aislados ocurrió únicamente durante la etapa de comparación genómica, mediante el análisis de presencia/ausencia de genes y las herramientas para pangenomas. De esta manera se garantiza la trazabilidad y reproducibilidad del análisis para cada genoma individual.

### 3.3.1 Herramientas para el análisis bioinformático

La secuenciación de ADN es una técnica que ha revolucionado la biología molecular y la genómica al permitir la lectura precisa del orden de los nucleótidos en una cadena de ADN. Existen diversas tecnologías de secuenciación. Las plataformas de Next-Generation Sequencing (NGS) son las más utilizadas hoy en día debido a su capacidad para generar

grandes cantidades de datos en un tiempo relativamente corto y con un costo más accesible en comparación con métodos convencionales (Goodwin *et al.,* 2016). Sin embargo, la obtención de secuencias de ADN no es un fin en sí mismo, sino un paso inicial que requiere un análisis bioinformático profundo para obtener información biológica relevante. Uno de los aspectos fundamentales de este proceso es la anotación. La anotación de secuencias genómicas se refiere a la identificación de regiones funcionales dentro del ADN, tales como genes, promotores, ARN de transferencia y otras estructuras (*Koonin et al., 2003*). En organismos bacterianos, por ejemplo, la anotación de genes permite predecir su capacidad metabólica, sus mecanismos de resistencia a antibióticos o su potencial biotecnológico (*Aziz et al., 2008*). Este paso es particularmente relevante en estudios comparativos entre diferentes aislados, ya que posibilita identificar genes compartidos o exclusivos que pueden ser responsables de características fenotípicas distintivas. Un paso crucial previo a la anotación es la evaluación de la calidad de las secuencias. Las secuencias genómicas generadas mediante NGS suelen contener fragmentos no deseados, como adaptadores, que son secuencias artificiales añadidas durante la preparación de las bibliotecas de ADN (*Andrews, 2010*). Los adaptadores facilitan el proceso de secuenciación, pero deben eliminarse antes del análisis final, ya que su presencia podría interferir con la correcta interpretación de los datos. Asimismo, es importante realizar un control de calidad detallado para identificar y corregir errores como bases de baja calidad, sesgos en el contenido de bases nitrogenadas o secuencias contaminantes que podrían haber sido introducidas en las muestras. Herramientas como "FastQC" proporcionan un análisis detallado de estos parámetros, generando gráficos y métricas que permiten tomar decisiones informadas sobre el tratamiento posterior de los datos (*Bolger et al., 2014*). Los datos obtenidos por las tecnologías de nueva generación (NGS) están fragmentados y desordenados por lo que requiere un proceso de ensamblaje posterior, donde se reconstruyen las secuencias originales. Este ensamblaje puede realizarse *de novo*, es decir, sin un genoma de referencia, o utilizando secuencias previamente conocidas como guía para ordenar y unir los fragmentos (*Bankevich et al., 2012*). El ensamblaje de secuencias es un proceso algorítmico que busca alinear y unir las lecturas de ADN para formar secuencias contiguas (o *contigs*), que representen porciones más largas y coherentes del genoma (*Li et al., 2010*). El uso de ensambladores como "SPAdes" permite optimizar este proceso, generando un mayor número de *contigs* de alta calidad que sirven

como base para la anotación y el análisis funcional subsecuente (*Nurk et al., 2013*). Este paso es esencial para estudios en organismos con genomas complejos o poco caracterizados, donde el ensamblaje adecuado puede marcar la diferencia entre obtener información genómica útil o confusa. Con esta introducción, queda claro que el análisis bioinformático de secuencias no solo implica el uso de herramientas automatizadas, sino una comprensión profunda de cada etapa del proceso, desde la obtención de datos crudos hasta su interpretación biológica.

### 3.4 Contigs

En el contexto del ensamblaje genómico, un *contig* (del inglés *contiguous sequence*) es una secuencia continua de ADN generada al alinear y unir múltiples lecturas (*reads*) de secuenciación que se solapan entre sí. Estas estructuras representan segmentos del genoma objetivo que han sido ensamblados sin interrupciones aparentes y con un alto grado de confianza (*Miller et al., 2010).*

Los *contigs* aparecen como resultado del proceso de ensamblaje, en el cual los algoritmos computacionales reconstruyen la secuencia genómica a partir de fragmentos cortos de ADN obtenidos mediante tecnologías de secuenciación masiva. Debido a que estas lecturas son más cortas que la longitud de los genomas que se desea ensamblar es necesario generar una gran cantidad de lecturas que cubran múltiples veces el genoma para que sea posible detectar regiones solapadas entre ellas. Estas regiones de solapamiento permiten la construcción de *contigs*. La calidad de un ensamblaje suele evaluarse por el número y tamaño de los *contigs* generados, siendo deseable un número menor de *contigs* y con mayor longitud. Los *contigs* forman la base para estructuras más complejas llamadas *scaffolds* o andamiajes, que agrupan y ordenan *contigs* utilizando información adicional, como lecturas de extremo pareado (*paired-end reads*), para inferir su posición relativa y la distancia entre ellos (*Miller et al., 2010).*

## 4. MARCO DE REFERENCIA

El análisis comparativo de genomas es una metodología esencial para identificar y comprender las variaciones genéticas entre diferentes aislados bacterianos. En el caso de *B. paralicheniformis*, dichas variaciones incluyen la presencia de genes únicos, la ausencia de rutas metabólicas específicas y la adquisición de elementos genéticos móviles que pueden

influir en la expresión de fenotipos diferenciados, tales como la producción de metabolitos antimicrobianos, enzimas extracelulares o la resistencia a condiciones ambientales adversas. Este tipo de divergencias está directamente relacionado con los procesos evolutivos propios de cada cepa y con la adaptación a nichos ecológicos particulares.

También se ha demostrado que ciertos aislados de *B. paralicheniformis* poseen genes asociados con la síntesis de compuestos bioactivos de interés médico e industrial. Songnaka *et al*. (2024) identificaron un péptido antimicrobiano similar a la bacitracina con actividad contra *Staphylococcus aureus* resistente a meticilina (MRSA), mientras que Rao, M. P. N., *et al.* (2024) reportaron exopolisacáridos con propiedades antioxidantes y quelantes de hierro. Estos compuestos refuerzan la relevancia biotecnológica de esta especie y justifican el interés en estudiar su diversidad genómica.

A nivel internacional, Du *et al*. (2019) realizaron un análisis comparativo del genoma de *B. paralicheniformis* MDJK30 frente a 55 cepas relacionadas. Sus resultados revelaron que esta especie posee un pangenoma abierto, con variaciones importantes en secuencias de inserción, profagos, islas genómicas y operones de metabolitos secundarios. Encontraron además que los grupos génicos de fengicina, bacitracina y ciertos lantipéptidos están presentes exclusivamente en *B. paralicheniformis*, adquiridos por transferencia horizontal y utilizables como marcadores genéticos para diferenciarla de *B. licheniformis*. De manera complementaria, Kushmitha *et al*. (2025) reportaron once clústeres biosintéticos de metabolitos secundarios en la cepa NB stem 4, incluyendo fengicina y liquenisina, los cuales explican su relevante actividad antifúngica, lo que posiciona a la especie como un agente biocontrolador prometedor en agricultura.

En México, el Centro de Investigación en Alimentación y Desarrollo (CIAD) ha desarrollado estudios sobre ensamblaje genómico empleando estrategias híbridas basadas en lecturas cortas y largas. En su informe de Bioinformática CIAD (2024), demostraron que la combinación de plataformas Illumina y Nanopore permite mejorar la resolución de regiones repetitivas y estructuras complejas del genoma, incrementando la precisión en la identificación de genes clínicamente relevantes. De forma similar, Chávez-Almanza *et al*. (2025) analizaron el genoma de *B. paralicheniformis* AA1, aislada de un sistema de milpa en Sonora, revelando un repertorio funcional diverso y confirmando su clasificación mediante taxonomía basada en genoma completo. Otros trabajos del CIAD han demostrado que el uso

de herramientas como “SPAdes” acoplado a corrección de errores con “Pilon” permite obtener ensamblajes más íntegros y precisos, lo que es fundamental para análisis comparativos posteriores.

En el contexto regional, existen investigaciones realizadas en el estado de Puebla que contribuyen a comprender la dinámica microbiana en ambientes urbanos y atmosféricos. Un estudio conducido por la Benemérita Universidad Autónoma de Puebla (BUAP) analizó la formación y composición de aerosoles atmosféricos en zonas urbano-forestales, destacando la influencia de factores ambientales locales en la presencia y dispersión de material biológico en el aire (López-Mendoza & Pérez-Guerrero, 2023). En un trabajo previo, investigadores de la BUAP aislaron microorganismos cultivables del aire en diversos puntos de la ciudad de Puebla, identificando numerosos géneros bacterianos presentes en bioaerosoles urbanos y modelando su transporte aerobiológico (Soto-Ramírez, 2017). A nivel nacional, una revisión reciente sobre microbiota aerotransportada en México mostró que la diversidad bacteriana presente en bioaerosoles varía significativamente entre regiones y depende de factores ambientales como urbanización, polvo en suspensión y emisiones antropogénicas (Martínez-Sánchez *et al*., 2024).

Estas investigaciones son particularmente relevantes para el presente estudio, ya que los aislados de *B. paralicheniformis* analizados provienen de bioaerosoles en la Ciudad de Puebla. El ensamblaje genómico y la comparación entre las cuatro cepas permiten identificar genes únicos y compartidos, así como variaciones estructurales derivadas de condiciones ambientales locales. Considerando que *B. paralicheniformis* posee un pangenoma abierto, la caracterización de aislados provenientes de ambientes urbanos constituye una oportunidad para aportar nueva información sobre la diversidad genética de esta especie y su adaptación al entorno atmosférico de Puebla.

En conjunto, los antecedentes internacionales, nacionales y regionales coinciden en la importancia de emplear estrategias de ensamblaje de *novo* y análisis comparativo para caracterizar la variabilidad genética entre cepas bacterianas. La integración de estudios provenientes de la literatura y del contexto local refuerza la pertinencia del enfoque utilizado, proporcionando una base conceptual y metodológica sólida para la interpretación de los resultados genómicos obtenidos en este trabajo.

## 5. PLANTEAMIENTO DEL PROBLEMA

*Bacillus paralicheniformis* destaca por su capacidad para producir enzimas y compuestos bioactivos. Sin embargo, a pesar de su relevancia potencial, aún existe un conocimiento limitado sobre su genoma y sobre las características de aislados provenientes de ambientes específicos, como el aire. Las condiciones atmosféricas como la variabilidad de temperatura, la humedad, la escasez de nutrientes y la radiación pueden influir significativamente en la diversidad genética y en las propiedades funcionales de las bacterias presentes en este medio. No obstante, los mecanismos que sustentan la plasticidad genómica de *B. paralicheniformis* en ambientes aerotransportados siguen siendo poco estudiados, lo que limita el aprovechamiento informado de su potencial biotecnológico. Si bien las herramientas de secuenciación y el análisis bioinformático han permitido explorar genes asociados con funciones clave, las diferencias genéticas entre aislados de *B. paralicheniformis* provenientes del aire continúan siendo poco conocidas. Esta falta de información dificulta comprender sus procesos de adaptación ambiental, así como identificar genes responsables de propiedades funcionales únicas que podrían ser esenciales para su optimización en aplicaciones biotecnológicas.

## 6. JUSTIFICACIÓN

El estudio de microorganismos presentes en el aire urbano es fundamental para comprender los procesos ecológicos, adaptativos y funcionales que ocurren en este tipo de ambientes. *Bacillus paralicheniformis* se ha reportado como una especie bacteriana de interés biotecnológico, ha sido escasamente caracterizada cuando se encuentra en condiciones atmosféricas, particularmente en ciudades con dinámicas ambientales complejas como la Ciudad de Puebla. La atmósfera urbana representa un entorno altamente variable, donde factores como las fluctuaciones de temperatura, humedad, radiación y disponibilidad limitada de nutrientes pueden ejercer presiones selectivas que favorecen la diversificación genética. Estas condiciones pueden generar variantes locales con atributos funcionales distintos a los descritos en cepas provenientes de otros ecosistemas. El ensamblaje del genoma completo y la caracterización comparativa de aislados ambientales provenientes de la Ciudad de Puebla permitirá identificar genes asociados con la adaptación a condiciones aerotransportadas, así como detectar rasgos genéticos diferenciales que puedan aportar ventajas funcionales. Esta

información es esencial no solo para ampliar el conocimiento sobre la diversidad microbiana del aire urbano, sino también para revelar el potencial biotecnológico de estas cepas locales, contribuyendo al desarrollo de aplicaciones en áreas como la biorremediación, la producción de compuestos bioactivos y la industria enzimática. La realización de este estudio proporciona una base científica sólida para la comprensión de la variabilidad genética y funcional de *B. paralicheniformis* en contextos ambientales urbanos y su aprovechamiento en distintos sectores tecnológicos.

## 7. OBJETIVOS

### 7.1 Objetivo General

Ensamblar y caracterizar el genoma de *Bacillus paralicheniformis* aislado de muestras ambientales de la Ciudad de Puebla, mediante herramientas de secuenciación y análisis bioinformático

### 7.2 Objetivos Específicos

1. Evaluar la calidad de las secuencias genómicas de los aislados utilizando “FastQC”.
2. Realizar el filtrado, recorte y limpieza de las lecturas utilizando “Trimmomatic”.
3. Ensamblar los genomas de los aislados utilizando “SPAdes”.
4. Anotar los genes de los aislados utilizando “Prokka”.
5. Comparar los perfiles genómicos de los aislados mediante análisis comparativos para identificar genes compartidos, genes exclusivos y posibles elementos asociados con la adaptación a condiciones atmosféricas de la Ciudad de Puebla.

## 8. HIPÓTESIS

8.1. H1 (alternativa): Los aislados de *Bacillus paralicheniformis* obtenidos de muestras ambientales de la Ciudad de Puebla presentan variación genómica significativa entre sí y contienen genes asociados con la adaptación a condiciones atmosféricas específicas.

8.2. H0 (nula): Los aislados de *B. paralicheniformis* provenientes de la Ciudad de Puebla no presentan diferencias genómicas significativas entre sí ni poseen genes diferenciadores asociados a la adaptación a condiciones atmosféricas.

## 9. DISEÑO DE LA INVESTIGACIÓN

### 9.1 Tipo de Estudio

Observacional, prospectivo, transversal y descriptivo.

### 9.2. Universo del estudio

- Cuatro cepas aisladas del aire en la Ciudad de Puebla, Puebla, identificadas fenotípica y genotípicamente como *Bacillus paralicheniformis*, obtenidas previamente en el Departamento de Microbiología de la Facultad de Ciencias Químicas de la BUAP.
- Herramientas de bioinformática

### 9.3. Tamaño de Muestra

Cuatro aislados de *B. paralicheniformis*.

### 9.4. Sede y Lugar de Estudio

Laboratorio de Microbiología, Departamento de Microbiología, Facultad de Ciencias Químicas, Benemérita Universidad Autónoma de Puebla (BUAP).

### 9.5. Criterios de selección

- Criterios de inclusión: cepas bacterianas que sean *B. paralicheniformis*
- Criterios de exclusión: cepas que sean *B. paralicheniformis* cuya cantidad o calidad de lecturas obtenidas no sea suficiente para realizar un ensamblaje genómico confiable.

### 9.6. Diseño estadístico

Estadística descriptiva

## 10. MATERIALES Y METODOLOGÍAS

### 10.1 Equipo de trabajo

Sistema operativo y herramientas bioinformáticas

El análisis bioinformático se realizó utilizando el sistema operativo Ubuntu debido a su eficiencia y compatibilidad con herramientas bioinformáticas. Las que se utilizaron fueron:

- Ubuntu: sistema operativo utilizado para la instalación y ejecución de herramientas bioinformáticas.
- "FastQC": evaluación de la calidad de las secuencias genómicas.
- "Trimmomatic": recorte y limpieza de secuencias de ADN.
- "SPAdes": ensamblaje de secuencias genómicas.
- "Prokka": anotación de *contigs* ensamblados.

### 10.2 *Hardware y software*

- Computadora con sistema operativo Ubuntu
- CPU de alto rendimiento
- Memoria RAM de 16 GB
- Disco duro de 1 TB

### 10.3 Análisis bioinformático

#### 10.3.1 Secuenciación genómica:

Los datos de secuenciación utilizados en este estudio forman parte del trabajo de aislamiento realizado previamente en el Departamento de Microbiología de la Facultad de Ciencias Químicas de la BUAP. Se emplearon cuatro conjuntos de datos de secuenciación obtenidos a partir de aislados del aire recolectados en la Ciudad de Puebla, identificados fenotípica y genotípicamente como *Bacillus paralicheniformis*. Las secuencias se encontraban en formato FASTQ.

#### 10.3.2 Evaluación de calidad (FastQC):

La calidad de las secuencias genómicas es un aspecto crítico en el análisis bioinformático, "FastQC" es una herramienta que se utiliza para evaluar la calidad de las secuencias de ADN obtenidas mediante secuenciación de alto rendimiento. "FastQC" genera informes detallados

sobre diversos parámetros de calidad, como la distribución de la calidad de las bases, la presencia de adaptadores y el contenido de GC, entre otros (Andrews, 2010).

Se utilizó "FastQC v0.11.9" para evaluar la calidad de las bases, la presencia de adaptadores y el contenido de GC. Los informes generados permitieron identificar secuencias que requirieron limpieza.

10.3.3 Recorte y limpieza (Trimmomatic):

Se empleó "Trimmomatic v0.39" para el recorte y limpieza de secuencias, eliminando adaptadores y bases de baja calidad. Este paso fue esencial para asegurar secuencias de alta calidad previo al ensamblaje y anotación (Bolger et al., 2014).

Se emplearon los siguientes parámetros:

- ILLUMINACLIP: identifica y elimina las secuencias de los adaptadores específicos del sistema comercial TruSeq de Illumina en las lecturas.
- SLIDINGWINDOW: aplica un recorte dinámico de la lectura cuando el promedio de la calidad PHRED en una ventana de 4 bases consecutivas cae por debajo de 15.
- LEADING y TRAILING: eliminan las bases de baja calidad en los extremos 5' y 3' de cada lectura.
- MINLEN: descarta las lecturas resultantes cuyo tamaño final sea menor de 36 bases.

Las líneas de código usadas en "Trimmomatic" fueron las siguientes:

```
1. java -jar trimmomatic-0.39.jar PE \
   secuenciaR1.fastq.gz secuenciaR2.fastq.gz \
   secuenciaR1_paired.fastq.gz secuencia_R1_unpaired.fastq.gz \
   secuencia_R2_paired.fastq.gz secuencia_R2_unpaired.fastq.gz \
   ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 \
   LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

10.3.4 Ensamblaje genómico (SPAdes):

"SPAdes (v3.13.1)", es una herramienta de ensamblaje de genomas que genera *contigs* a partir de secuencias de ADN. "SPAdes" es conocido por su capacidad para ensamblar genomas bacterianos de manera eficiente y precisa, proporcionando una base sólida para el análisis genómico posterior (Bankevich et al., 2012).

Las secuencias limpiadas se ensamblaron utilizando "SPAdes" para generar *contigs*. Para cada aislado, se utilizó la siguiente línea de comando, especificando los archivos pareados obtenidos de "Trimmomatic" y definiendo una carpeta de salida específica:

Línea de código para el ensamblaje para los aislados

spades.py -1 secuenciaR1_paired.fastq.gz -2 secuenciaR2_paired.fastq.gz -o nombredesecuencia_output --careful -t 4 -m 16

Este comando se replicó para cada muestra (desde el aislado 1 al 4), con sus respectivos nombres. Los parámetros -1 y -2 se utilizaron para indicar los archivos pareados (R1 Y R2), mientras que -t y -m definieron el número de núcleos de procesamiento y la memoria RAM asignada, respectivamente.

El archivo de salida más relevante fue *contigs.fasta*, el cual contiene las secuencias ensambladas. Las principales métricas analizadas para evaluar la calidad del ensamblaje fueron:

- Número total de *contigs*.
- Tamaño total del ensamblaje.
- N50: longitud mínima tal que el 50% del genoma ensamblado está contenido en *contigs* de esa longitud o mayores (indicador de ensamblaje robusto).

10.3.5 Anotación de genes (Prokka):

"Prokka (v1.14.6)", es una herramienta utilizada para la anotación rápida de *contigs* ensamblados. Proporciona una anotación integral de genes, ARN y otras características genómicas, facilitando la interpretación funcional de los datos genómicos (Seemann, 2014).

Los *contigs* ensamblados se anotaron utilizando "Prokka" para identificar genes, ARN y otras características genómicas.

El archivo *contigs.fasta* generado por "SPAdes" para cada muestra fue utilizado como entrada. Se emplearon los siguientes comandos para anotar cada uno de los genomas ensamblados:

Ejemplo de comandos para la anotación para los archivos

prokka nombredesecuenciaoutput/contigs.fasta --outdir archivodesalida_Prokka --prefix nombreinicialdearchivo --cpus 4

Este comando (usado para cada archivo) produjo un conjunto de archivos de salida, entre los que destacan:

- gff: archivo principal con toda la anotación estructurada.
- faa: secuencias de proteínas predichas.

- ffn: secuencias de nucleótidos de los genes.
- tsv: tabla con los detalles de la anotación funcional.
- txt: resumen general del análisis.

### 10.3.6 Creación de archivos mediante "Roary"

"Roary" es una herramienta diseñada para la construcción del pangenoma bacteriano a partir de múltiples genomas anotados en formato ".gff" (Page et al., 2015). Este software permite identificar los genes core, accesorios y únicos, generando archivos de salida que facilitan el análisis comparativo entre cepas.

Para este estudio, se utilizaron los archivos ".gff" generados por "Prokka" para cada uno de los aislados de *B. paralicheniformis*. Los análisis se realizaron empleando "Roary (v3.13.0)", ejecutado en entorno Linux.

La línea de comando utilizada fue la siguiente:

```
roary -e -n -v -p 4 *.gff
```

Donde el parámetro -e activa el uso de alineación rápida basada en *MAFFT*, -n genera una matriz de presencia/ausencia de genes, y -p define el número de núcleos de procesamiento. El resultado principal, *gene_presence_absence.csv*, se empleó posteriormente para analizar los genes compartidos y diferenciales entre las cepas. Este análisis es fundamental para entender la plasticidad genómica y las adaptaciones ambientales presentes en los aislados.

### 10.3.7 Uso de IQ-TREE2 para el árbol filogenético

El análisis filogenético se llevó a cabo utilizando "IQ-TREE 2 (v2.4.0)", una herramienta moderna para la inferencia de árboles filogenéticos mediante el método de máxima verosimilitud (*Maximum Likelihood*). "IQ-TREE 2" incorpora modelos evolutivos avanzados y estimaciones ultrarrápidas de soporte de rama (*bootstrap*), lo que permite generar árboles robustos y reproducibles (Minh et al., 2020).

Para construir el árbol filogenético, se utilizó como entrada el archivo de alineamiento concatenado generado por "Roary" (*core_gene_alignment.aln*). La línea de comando empleada fue la siguiente:

```
iqtree2 -s core_gene_alignment.aln -m MFP -bb 1000 -nt AUTO
```

El parámetro -m MFP permite que "IQ-TREE 2" seleccione automáticamente el mejor modelo evolutivo, mientras que -bb 1000 aplica un *bootstrap* ultrarrápido con 1000 réplicas para evaluar el soporte de las ramas.
El árbol resultante (*.treefile*) fue visualizado y editado con el *software* "FigTree" y "iTOL" (Interactive Tree of Life), lo que permitió una representación clara de las relaciones evolutivas entre los aislados de *B. paralicheniformis* analizadas.

### 10.3.8 Comparación genómica:

Los perfiles genómicos de los diferentes aislado se compararon para identificar genes específicos y compartidos. Esta comparación permitió comprender mejor la variabilidad genética y su impacto en la producción de enzimas y metabolitos secundarios.

## 11. DIAGRAMA GENERAL DEL TRABAJO

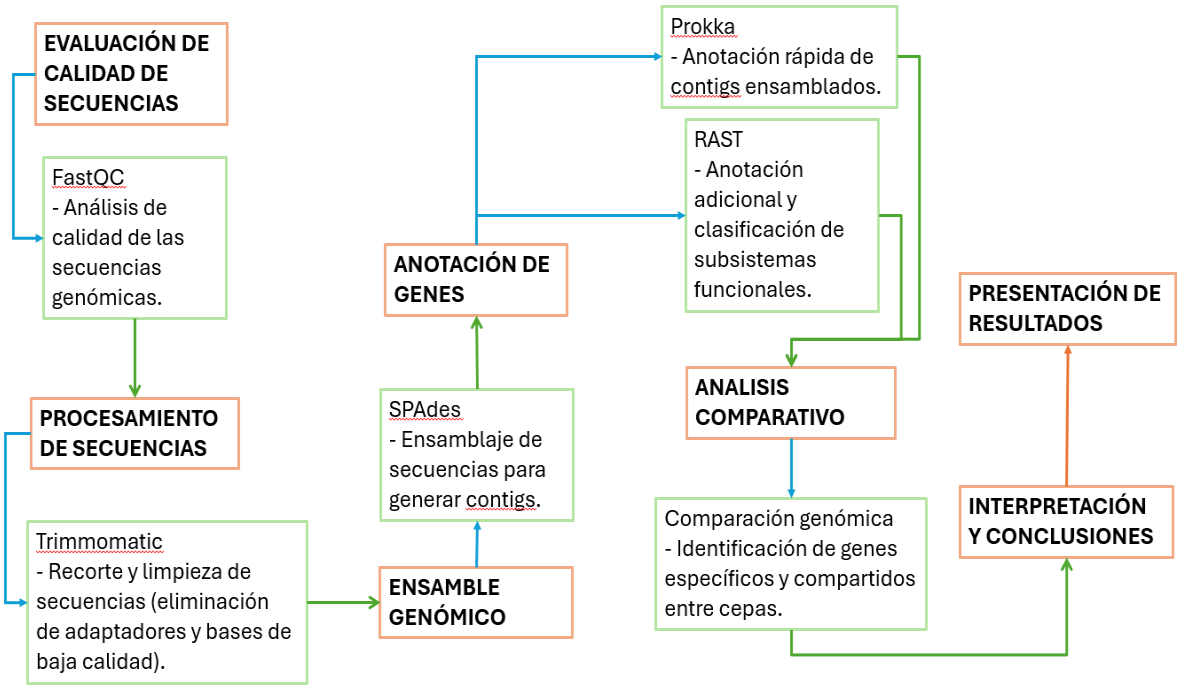


**Fig 1**. Diagrama de la metodología.

12. RESULTADOS

## 12.1 Evaluación de la calidad de las secuencias (FastQC)

Para garantizar que los análisis genómicos se realizaron con datos confiables, se efectuó primero la evaluación de calidad de las secuencias obtenidas mediante secuenciación. Esta revisión se realizó con el programa "FastQC (v0.11.9)", herramienta que permite analizar parámetros fundamentales como el valor de calidad de cada base, el contenido de GC, la presencia de adaptadores y la longitud de las lecturas. Como se observa en la Figura 2, se utilizó la terminal de Ubuntu para acceder a los archivos "fastq.gz" sobre los cuales se ejecutó "FastQC" y los análisis posteriores.

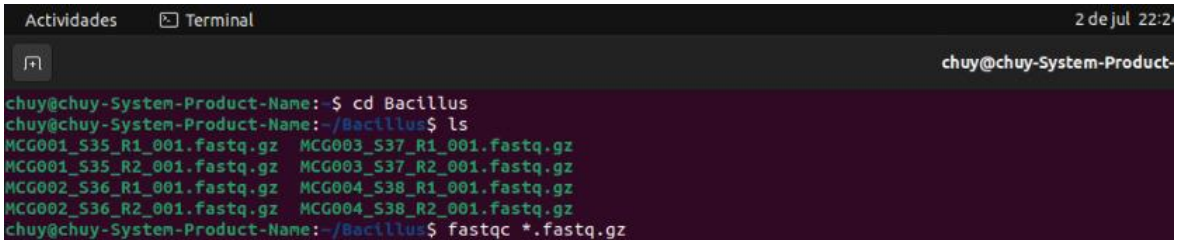


**Fig 2.** Interfaz de la terminal de Ubuntu. Ejemplificación para ingresar a la carpeta con los archivos "fastq.gz" en los cuales se correrá el análisis de calidad y todos los análisis posteriores.

Los resultados mostraron valores PHRED superiores a 30 en la mayoría de las posiciones de lectura, lo que indica una calidad alta y una baja probabilidad de error en la llamada de bases (precisión aproximada del 99.9 %; Illumina, 2023). Este tipo de puntaje, ampliamente utilizado en las plataformas de secuenciación actuales, refleja la probabilidad de error mediante una escala logarítmica, donde valores mayores representan una mayor confianza en la base identificada. Sin embargo, se observó una ligera disminución de la calidad hacia el extremo 3' de las lecturas, acompañada de señales residuales de adaptadores (Figura 3). Estos hallazgos justificaron la necesidad de un proceso de limpieza previo al ensamblaje.

En general, los archivos crudos presentaron un contenido de GC de aproximadamente 45 % y un tamaño total cercano a 1.3 Gbp, con más de 17 millones de secuencias por muestra. Estos parámetros confirman que las lecturas eran adecuadas para continuar con los análisis bioinformáticos.

| *Parametro* | Valor |
|---|---|
| *Archivo* | MCG001_S35_**R1**_001.fastq.gz |
| *Tipo de archivo* | Bases convencionales |
| *Codificador* | Sanger/Illumina 1.9 |
| *Total de secuencias* | 17675907 |
| *Total de bases* | 1.3 Gbp |
| *Secuencias de baja calidad* | 0 |
| *Longitud de secuencia* | 76 |
| *%GC* | 45 |

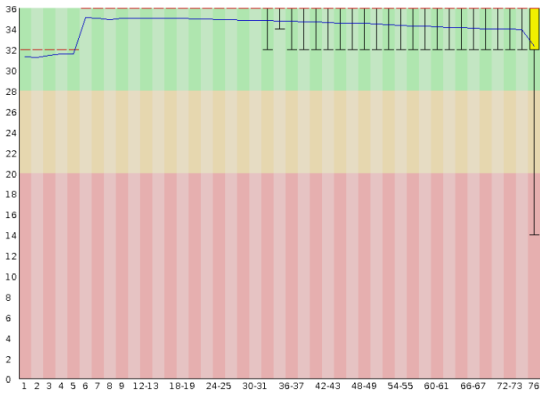


**Fig 3.** Análisis de calidad realizado con "FastQC" para las lecturas crudas del aislamiento de *B. paralicheniformis*. Gráfico de calidad del archivo R1 leído en sentido 3'-5'. En los gráficos, el eje X indica la posición de las bases en las lecturas y el eje Y la puntuación de calidad (Q-score). La escala cromática representa la calidad: verde para alta, amarillo para intermedia y rojo para baja. En general, las lecturas presentan buena calidad, aunque en las lecturas reversas se aprecia una caída más evidente hacia el extremo final de las secuencias.

Este tipo de información fue fundamental para decidir el siguiente paso: el recorte de secuencias de baja calidad y eliminación de contaminantes.

### 12.2 Recorte y limpieza de las secuencias (Trimmomatic)

Con base en los resultados del control de calidad inicial, se aplicó un proceso de recorte utilizando el programa "Trimmomatic (v0.39)". Este paso permitió eliminar adaptadores, bases con baja calidad y lecturas demasiado cortas que pudieran interferir con el ensamblaje.

Después del procesamiento, los reportes de calidad mostraron mejoría significativa: eliminación completa de adaptadores y caída de calidad en los extremos 3'. Las lecturas mantuvieron longitud promedio de 76 pb y contenido GC del 45%., sin secuencias descartadas por baja calidad (Figura 4).

| *Parametro* | Valor |
|---|---|
| *Archivo* | MCG001_S35_**R1**_paired.fastq.gz |
| *Tipo de archivo* | Bases convencionales |
| *Codificador* | Sanger/Illumina 1.9 |
| *Total de secuencias* | 17209085 |
| *Total de bases* | 1.3 Gbp |
| *Secuencias de baja calidad* | 0 |
| *Longitud de secuencia* | 76 |
| *%GC* | 45 |

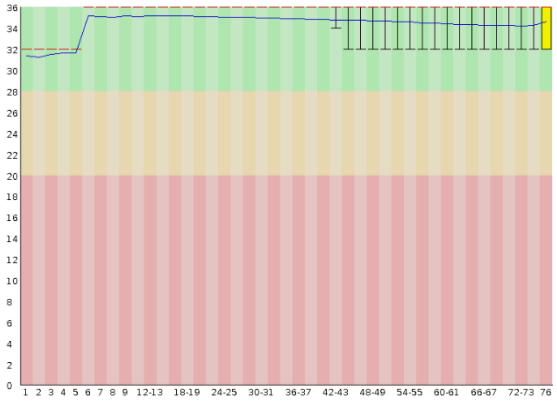


**Fig 4.** Reporte de control de calidad de lecturas crudas de *B. paralicheniformis*. Lecturas 3'– 5' (R1) con un total de 17,209,085 secuencias y 1.3 Gbp, con un contenido GC de 45 %. La gráfica de calidad muestra valores elevados (Q > 30) a lo largo de la mayoría de las posiciones, con mejoría del extremo 3' respecto a los análisis previos de las lecturas.

La mejora obtenida tras el recorte aseguró que las secuencias utilizadas en el ensamblaje fueran confiables y libres de errores comunes, garantizando la precisión de las etapas posteriores.

### 12.3 Ensamblaje genómico (SPAdes)

Las secuencias limpias se ensamblaron con "SPAdes (v3.13.1)", un programa especializado en genomas bacterianos. Este proceso reconstruyó los fragmentos de ADN en secuencias más largas llamadas *contigs*, las cuales representan porciones continuas del genoma.

| Comando o parámetro | Explicación |
|---|---|
| spades.py | Llama al programa SPAdes. No se necesita escribir python, simplemente se ejecuta desde la terminal. |
| -1 archivo_R1 | Indica el archivo de lecturas pareadas "forward" (adelante). Debe ser el archivo limpio que resultó de Trimmomatic. |
| -2 archivo_R2 | Indica el archivo de lecturas pareadas "reverse" (atrás). Igual, debe ser el recortado con Trimmomatic. |
| -o nombre_output | Carpeta de salida. SPAdes creará aquí archivos como contigs.fasta, scaffolds.fasta, etc. |
| --careful | Modo de ensamblaje más preciso: reduce errores como inserciones o eliminaciones, especialmente útil en genomas pequeños (como bacterianos). |
| -t 4 | Número de hilos o núcleos del procesador a usar. Se puede ajustar este valor según lu computadora (por ejemplo, -t 8 si se tienen 8 núcleos). |
| -m 16 | Memoria RAM en GB que SPAdes puede usar. Ajusta este valor si se tiene menos RAM disponible (por ejemplo -m 8 si tienen 8 GB libres). |

**Tabla 1.** Parámetros de "Spades". Significado de los códigos del programa "SPAdes", así como la función de cada uno.

Cada aislado de *Bacillus paralicheniformis* produjo un número distinto de *contigs* y un tamaño total de ensamblaje dentro del rango esperado para la especie (alrededor de 4.2 Mbp). La métrica N50, utilizada para evaluar la robustez del ensamblaje, mostró valores adecuados, lo que indica una buena continuidad y cobertura genómica.

Los resultados obtenidos confirman que la calidad de los datos y la configuración de "SPAdes" fueron suficientes para generar ensamblajes confiables, útiles para la posterior anotación de genes y análisis comparativo.

### 11.4 Anotación genómica (Prokka)

Una vez obtenido el ensamblaje genómico de las cepas de *B. paralicheniformis*, se procedió a realizar la anotación funcional mediante el software "Prokka (v1.14.6)", una herramienta ampliamente utilizada para la predicción, identificación y etiquetado de elementos funcionales en genomas procariotas.

La anotación genómica tiene como objetivo reconocer y clasificar distintas características dentro del genoma ensamblado, tales como:

- Genes codificantes de proteínas (CDS, *Coding DNA Sequences*),
- Genes de ARN ribosomal (rRNA),
- Genes de ARN de transferencia (tRNA),
- Regiones hipotéticas (sin anotación conocida) y otras secuencias funcionales.

"Prokka" automatiza este proceso al integrar varias herramientas bioinformáticas, entre las que destacan:

- "Prodigal", para la predicción de CDS.
- "Barrnap", para la identificación de genes rRNA.
- "Aragorn", para la anotación de tRNA.
- Búsqueda contra bases de datos como "UniProt", "RefSeq", y "HAMAP", para asignación funcional y nomenclatura estandarizada.

### 12.4 Anotación genómica (Prokka)

El ensamblaje resultante fue analizado con "Prokka (v1.14.6)", programa que predice y clasifica genes en categorías funcionales. Esta herramienta permite identificar genes (CDS), (rRNA), (tRNA) y regiones hipotéticas sin función conocida.

| Parámetro | Descripción |
|---|---|
| prokka | Llama al programa de anotación genómica Prokka. |
| contigs.fasta | Archivo de ensamblaje generado por SPAdes. Contiene las secuencias que se van a anotar. |
| --outdir MCG001_Prokka | Carpeta donde se guardarán los resultados de la anotación. |
| --prefix MCG001 | Prefijo con el que comenzarán todos los archivos de salida (por ejemplo, MCG001.gff, etc). |
| --cpus 4 | Número de núcleos del procesador a usar (puedes ajustar según tu máquina). |

**Tabla 2.** Parámetros de "Prokka". Significado de los códigos del programa "Prokka", así como la función de cada uno.

"Prokka" detectó un número total de CDS relativamente constante entre los cuatro aislados, en el aislado uno encontró 4323, el dos 4305, el tres 4314 y el cuatro 4307, mientras que las proteínas hipotéticas representaron alrededor de un tercio del total (entre 1 437 y 1 442, tabla 3).

| Característica | Aislado 1 | Aislado 2 | Aislado 3 | Aislado 4 |
|---|---|---|---|---|
| Total de CDS | 4323 | 4305 | 4314 | 4307 |
| Total de rRNA | 10 | 7 | 8 | 8 |
| Proteínas hipotéticas | 1442 | 1442 | 1437 | 1437 |
| Genes anotados (con nombre) | 2682 | 2682 | 2877 | 2870 |

**Tabla 3.** CDS's resultantes. Descripción de los datos resultantes mediante la anotación genómica con "Prokka".

Como todas los aislados corresponde a la bacteria *B. paralicheniformis*, nos preguntamos cuáles son los genes que son diferenciales, por lo que hicimos una comparación de los genes que son específicos de cada aislado.

### 12.5 Comparación de los genes específicos de cada aislado

El aislado 2 presentó 11 CDS específicos, mientras que los demás aislados presentan 7 CDS únicos cada uno.

En total el programa logró identificar 32 CDS específicos de cada aislado, de los que en el aislado dos se encuentran 11, mientras que, en el uno, tres y cuatro se encuentran 7, la mayoría de ellos fueron anotados como proteínas hipotéticas. Además, el programa logró identificar otros genes que se pueden encontrar en algunos aislados, pero en otros no.

| **CDS** | **tamaño** | **aislado** | | | |
|---|---|---|---|---|---|
| | **(Nt)** | 1 | 2 | 3 | 4 |
| ARN ribosómico 23S (parcial) | 1980 | | | | ✓ |
| ARN ribosómico 23S (parcial) | 1644 | | | ✓ | |
| ARN ribosómico 23S (parcial) | 1465 | | | ✓ | |
| Proteína hipotética | 1392 | ✓ | | | |
| Aspartato fosfatasa reguladora de respuesta I | 1104 | | ✓ | | |
| Proteína hipotética | 1020 | | | ✓ | |
| ARN ribosómico 23S (parcial) | 1001 | | | | ✓ |
| ARN ribosómico 23S (parcial) | 990 | | | ✓ | |
| ARN ribosómico 23S (parcial) | 983 | ✓ | | | |
| ARN ribosómico 23S (parcial) | 893 | | | | ✓ |
| ARN ribosómico 23S (parcial) | 867 | ✓ | | | |
| ARN ribosómico 16S (parcial) | 738 | | | | ✓ |
| ARN ribosómico 16S (parcial) | 722 | | ✓ | | |
| ARN ribosómico 16S (parcial) | 702 | | | | ✓ |
| ARN ribosómico 16S (parcial) | 606 | ✓ | | | |
| Proteína hipotética | 582 | | ✓ | | |
| Proteína hipotética | 453 | | ✓ | | |
| Proteína hipotética | 423 | ✓ | | | |
| Proteína hipotética | 372 | | | | ✓ |
| Subunidad 3 de la citocromo c oxidasa | 306 | | | ✓ | |
| Proteína hipotética | 288 | | | | ✓ |
| Proteína hipotética | 285 | | | ✓ | |
| Proteína hipotética | 270 | | ✓ | | |

| | | | | | |
|---|---|---|---|---|---|
| Proteína hipotética | 264 | | ✓ | | |
| Deshidrogenasa de NADH | 249 | | ✓ | | |
| Probable L,D-transpeptidasa YcfS | 207 | | | ✓ | |
| Proteína hipotética | 207 | ✓ | | | |
| Proteína hipotética | 165 | ✓ | | | |
| Proteína hipotética | 117 | | ✓ | | |
| Proteína hipotética | 114 | | ✓ | | |
| ARN ribosómico 5S | 101 | | ✓ | | |
| ARNt-Ser(gct) | 91 | | ✓ | | |

**Tabla 4.** CDS específicos de cada aislado. Genes específicos que nos dieron como resultado de la anotación con "Prokka". Los símbolos ✓ indican la presencia del gen en cada aislado.

Entre los genes específicos identificados en los cuatro aislados de *B. paralicheniformis*, se observaron CDS parciales correspondientes a componentes ribosomales (ocho 23S, cuatro 16S y uno 5S), 14 CDS hipotéticos, una secuencia correspondiente al ARNt-Ser(gct) y genes asociados con procesos metabólicos y de regulación, la aspartato fosfatasa reguladora de respuesta I, la probable L,D-transpeptidasa YcfS, la deshidrogenasa de NADH y la subunidad 3 de la citocromo c oxidasa. Estas últimas participan en mecanismos de señalización celular, respiración aeróbica y mantenimiento de la integridad de la pared celular.

Por otra parte, con los resultados obtenidos después de la anotación se lograron identificar CDS compartidos entre algunos aislados de *B. paralicheniformis*. En total se detectaron 14 CDS, de los cuales 10 fueron anotados como proteínas hipotéticas. El aislado 1 presentó coincidencias con los demás aislados, incluyendo dos ARN ribosómicos 23S parciales, varias proteínas hipotéticas de diferentes tamaños (1701, 930, 417, 315, 312, 252 y 156 nt), y el gen aspartato fosfatasa reguladora de respuesta I (1068 nt). El aislado 2 compartió CDS con los aislados 1 y 3, principalmente proteínas hipotéticas y un ARN ribosómico 23S parcial. El aislado 3 presentó coincidencias con los aislados 1, 2 y 4, destacando la aspartato fosfatasa reguladora de respuesta I, la subunidad 2 de la citocromo c oxidasa (264 nt) y varias proteínas hipotéticas (930, 735, 408 y 366 nt). Finalmente, el aislado 4 compartió CDS con los aislados 1 y 3, incluyendo una aspartato fosfatasa reguladora de respuesta I, una subunidad 2 de la

citocromo c oxidasa, y varias proteínas hipotéticas (1701, 735, 408, 366 y 156 nt). En cuanto a la función de los CDS compartidos, se identificaron dos secuencias correspondientes al ARN ribosómico 23S (parcial), junto con genes implicados en señalización celular y respiración aeróbica, la aspartato fosfatasa reguladora de respuesta I y la subunidad 2 de la citocromo c oxidasa.

| **CDS** | **Tamaño (Nt)** | **Aislado** | | | |
|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** |
| ARN ribosómico 23S (parcial) | 2579 | ✓ | ✓ | | |
| Proteína hipotética | 1701 | ✓ | | | ✓ |
| Aspartato fosfatasa reguladora de respuesta I | 1068 | ✓ | | ✓ | ✓ |
| ARN ribosómico 23S (parcial) | 996 | ✓ | | ✓ | |
| Proteína hipotética | 930 | ✓ | ✓ | ✓ | |
| Proteína hipotética | 735 | | | ✓ | ✓ |
| Proteína hipotética | 417 | ✓ | | ✓ | |
| Proteína hipotética | 408 | | | ✓ | ✓ |
| Proteína hipotética | 366 | | | ✓ | ✓ |
| Proteína hipotética | 315 | ✓ | ✓ | | |
| Proteína hipotética | 312 | ✓ | ✓ | | |
| Subunidad 2 de la citocromo c oxidasa | 264 | | | ✓ | ✓ |
| Proteína hipotética | 252 | ✓ | ✓ | | |
| Proteína hipotética | 156 | ✓ | | | ✓ |

**Tabla 5.** CDS identificados como genes compartidos de *B. paralicheniformis*. Los símbolos ✓ muestran la presencia del gen en cada cepa. Este grupo de CDS representa porciones del genoma accesorio compartidas entre aislados, posiblemente relacionadas con funciones metabólicas o regulatorias comunes.

Por otro lado, al analizar los CDS compartidos entre los aislados, se observó que algunos genes mantienen su presencia en más de una cepa, aunque presentan variaciones en su

longitud (Tabla 6). Estas diferencias podrían atribuirse a mutaciones puntuales, inserciones o eliminaciones, las cuales reflejan procesos de microevolución dentro de la especie.

Estos genes son los siguientes:

- **Ligasa de D-alanina–proteína transportadora de D-alanil,** con un tamaño de 16,461 nt en el aislado 2, mientras que en los aislados 1, 3 y 4 presenta 16,224 nt.
- **Quinasa de histidina sensora ComP**, con una longitud de 2,313 nt en los aislados 1 y 2, y de 1,932 nt en los aislados 3 y 4.
- **Metiltransferasa G de la subunidad pequeña del ARN ribosómico**, con 753 nt en los aislados 1 y 2, y 720 nt en los aislados 3 y 4.
- **Pectato liasa C**, que mostró 669 nt en los aislados 1 y 4, mientras que en el 2 y 3 fueron de 666 nt.
- Finalmente, la **proteína de partición cromosómica Smc** presentó 618 nt en los aislados 1 y 2, y 576 nt en los aislados 3 y 4.

| **CDS** | **Aislado (tamaño nt)** | | | |
|---|---|---|---|---|
| | **1** | **2** | **3** | **4** |
| Ligasa de D-alanina--proteína transportadora de D-alanil | 16224 | 16461 | 16224 | 16224 |
| Quinasa de histidina sensora ComP | 2313 | 2313 | 1932 | 1932 |
| Metiltransferasa G de la subunidad pequeña del ARN ribosómico | 753 | 753 | 720 | 720 |
| Pectato liasa C | 669 | 666 | 666 | 669 |
| Proteína de partición cromosómica Smc | 618 | 618 | 576 | 576 |

**Tabla 6.** CDS compartidos entre los diferentes aislados de *B. paralicheniformis*. Los símbolos ✓ indican la presencia del gen en cada aislado. Las variaciones en el tamaño de los CDS sugieren posibles diferencias estructurales o ajustes adaptativos entre cepas.

En resumen, el análisis comparativo de los cuatro aislados de *Bacillus paralicheniformis* permitió identificar diferencias notables en el número y tipo de CDS presentes. El aislado 1 presentó 7 CDS específicos y 15 compartidos, con un total de 22 CDS. El aislado 2 mostró 11 CDS específicos y 10 compartidos, sumando 21 CDS. El aislado 3 registró 7 CDS

específicos y 13 compartidos, con 20 CDS en total, mientras que el aislado 4 presentó 7 CDS específicos y 12 compartidos, para un total de 19 CDS. Estos patrones fueron visualizados mediante diagramas de Venn generados con Venny 2.1 (Oliveros, 2007-2015), lo que permitió identificar de manera más clara las regiones exclusivas y compartidas entre los aislados.

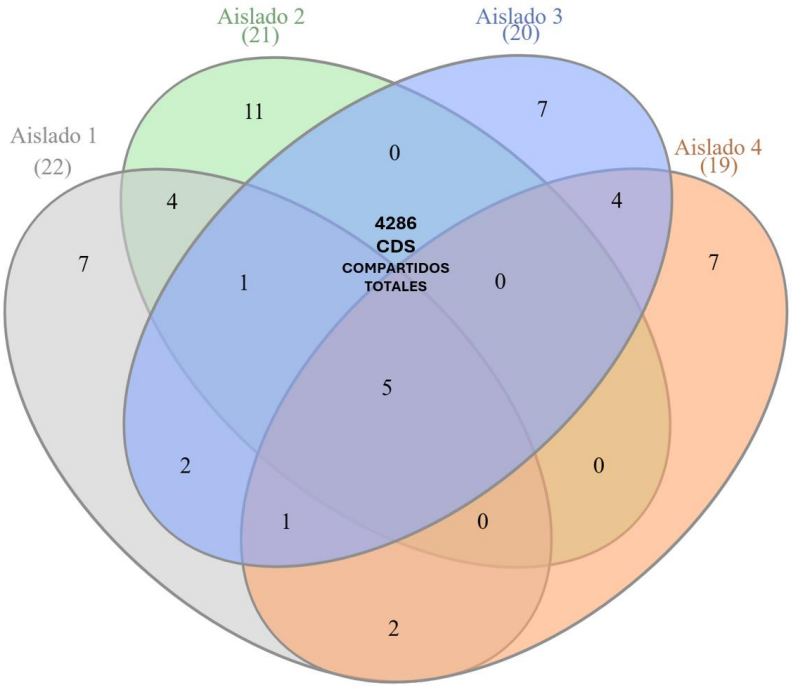


**Fig 5.** Comparación de los conteos de genes (CDS). En el diagrama se puede observar tanto los genes específicos de cada aislado como los genes compartidos entre ellos. El solapamiento de las regiones representa la cantidad de CDS comunes, mientras que las áreas no superpuestas muestran los CDS únicos de cada cepa. Este análisis permite identificar patrones de conservación y variación genómica entre los cuatro aislados de *Bacillus paralicheniformis*, destacando la presencia de un conjunto central de genes compartidos y un número variable de genes diferenciales que podrían estar relacionados con adaptaciones específicas.

### 12.6 Construcción del árbol filogenético con IQ-TREE 2

Para representar las relaciones evolutivas entre las cepas analizadas de *B. paralicheniformis*, se empleó el software "IQ-TREE 2" (Minh et al., 2020), un programa de inferencia filogenética de máxima verosimilitud que ofrece alta precisión y rapidez computacional. Este análisis se basó en el archivo *core_gene_alignment.aln* generado por "Roary", que contiene un alineamiento múltiple de los genes del genoma central compartido por todas los aislados.

La ejecución se realizó con el siguiente comando:

“iqtree2 -s core_gene_alignment.aln -nt AUTO -m GTR+G -bb 1000 -alrt 1000”

Este comando tiene ciertas características especiales para iqtree2 como se describen en la siguiente tabla:

| Parámetro | Significado / Función |
|---|---|
| -s core_gene_alignment.aln | Indica el archivo de alineamiento múltiple de los genes centrales generado por Roary. |
| -nt AUTO | Detecta y utiliza automáticamente todos los núcleos del procesador disponibles. |
| -m GTR+G | Especifica el modelo de sustitución **GTR (General Time Reversible)** con **distribución gamma** para heterogeneidad entre sitios. |
| -bb 1000 | Realiza **1000 réplicas de bootstrap ultrarrápido** para evaluar el soporte estadístico de las ramas. |
| -alrt 1000 | Ejecuta **1000 pruebas SH-aLRT** (test de razón de verosimilitud aproximada) para evaluar la confiabilidad de las ramas. |

**Tabla 7.** Parámetros de IQ-TREE2. Significado de los códigos del programa IQ-TREE2, así como la función de cada uno.

El resultado principal fue un archivo en formato "Newick” (*core_gene_alignment.aln.treefile*), el cual representa el árbol filogenético estimado. Este archivo puede visualizarse posteriormente con herramientas como “iTOL” o “FigTree”. El uso de “IQ-TREE 2” permitió obtener un árbol altamente resolutivo que facilita la interpretación de la cercanía evolutiva entre cepas ambientales, apoyando así la caracterización genética del aislado estudiado.

### 12.7 Visualización del árbol filogenético

El archivo “.treefile” generado por “IQ-TREE 2” fue visualizado utilizando la plataforma “iTOL” (*Interactive Tree Of Life*), una herramienta que permite la representación gráfica interactiva de árboles filogenéticos. Esta plataforma facilita la interpretación de relaciones evolutivas, ya que permite resaltar ramas, modificar estilos visuales, incorporar metadatos

(como el origen de las cepas) y exportar figuras en alta resolución, adecuadas para su inclusión en trabajos científicos.

Se compararon los cuatro aislados de *B. paralicheniformis*. Se personalizaron las ramas mediante el archivo *itol_tree_colors.txt*, asignando a cada cepa un color único y un grosor cinco veces superior al predeterminado, con el fin de mejorar su visibilidad y diferenciación.

Existen genomas completos de *Bacillus paralicheniformis* disponibles en bases de datos públicas como NCBI. Sin embargo, no fueron incluidos en el análisis comparativo por dos razones principales. Primero, el objetivo del estudio se centró en comparar exclusivamente los aislados ambientales obtenidos en este trabajo, para evaluar su variación genómica local y su relación evolutiva entre sí. Segundo, la inclusión de genomas completos de referencia (particularmente aquellos con mayor calidad o distintos métodos de secuenciación) puede introducir sesgos en la longitud de ramas y en el cálculo de distancias evolutivas, lo que dificultaría interpretar la microevolución únicamente entre los aislados ambientales. No obstante, se reconoce que incorporar genomas de referencia podría ser útil en futuros análisis para contextualizar filogenéticamente a los aislados dentro de la diversidad global de *B. paralicheniformis*. La relación filogenética entre las cuatro cepas analizadas se resume en la Figura 5, construida a partir del árbol generado con "IQ-TREE 2" y visualizado en "iTOL" (Letunic & Bork, 2021).

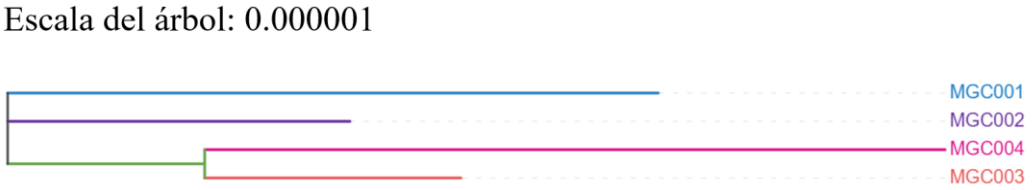


**Fig 6.** Árbol filogenético de los cuatro aislados de *Bacillus paralicheniformis*, generado con "IQ-TREE 2" y visualizado en "iTOL".

El árbol se construyó a partir del alineamiento del genoma central. Para mejorar la visualización, las ramas fueron coloreadas mediante el archivo *itol_tree_colors.txt*, asignando un color y grosor específico a cada aislado. Esta representación permitió observar la diversidad genómica y la separación entre las cepas analizadas.

La estructura del árbol filogenético evidenció una clara separación entre las cepas ambientales analizadas, lo que sugiere una diversidad genómica dentro del grupo *B.*

*paralicheniformis*. Las longitudes de las ramas reflejan las distancias evolutivas: ramas más largas indican un mayor número de cambios o mutaciones acumuladas entre cepas.

Específicamente, los aislados identificadas como aislado 3 y aislado 4 compartieron un nodo interno, lo que indica una relación evolutiva más estrecha entre ellas. El aislado 2 mostró una posición intermedia, mientras que el aislado 1 se ubicó más alejado del resto, con una rama notoriamente más larga, lo que sugiere un linaje más divergente o la posible existencia de una variación genómica más significativa.

Las diferencias genómicas observadas entre los cuatro aislados pueden cuantificarse utilizando los datos derivados del ensamblaje y anotación. De acuerdo con la Tabla 4, el número total de proteínas anotadas varió ligeramente entre los genomas, con valores que oscilaron entre 3385 y 3388 genes codificantes. De manera similar, el número de proteínas hipotéticas mostró pequeñas diferencias, de 1437 a 1442 por aislado.

Además, el análisis comparativo de CDS (Tablas 5–7) reveló variaciones tanto en presencia/ausencia como en el tamaño en nucleótidos de ciertos genes. Por ejemplo, la subunidad pequeña del ARN ribosomal presentó longitudes de 753 nt en los aislados 1 y 2, pero 720 nt en los aislados 3 y 4. Otros genes, como la pectato liasa C, mostraron diferencias menores en longitud (666–669 nt), mientras que genes como la proteína Smc presentaron variaciones más marcadas (618 nt vs. 576 nt).

Estas diferencias, aunque sutiles, son suficientes para generar las variaciones observadas en las longitudes de rama del árbol filogenético, ya que reflejan mutaciones, inserciones o deleciones acumuladas entre los aislados. En conjunto, los cambios en el número de genes anotados y en las longitudes específicas de diversos CDS explican cuantitativamente la diversidad genómica identificada entre los cuatro genomas estudiados.

Este enfoque basado en el genoma central ha sido ampliamente validado en estudios recientes sobre *B. paralicheniformis* (Liu *et al*., 2019), y proporciona un mayor poder de resolución evolutiva en comparación con análisis basados en secuencias individuales. En los ensamblajes obtenidos no se identificaron secuencias de plásmidos. Esto puede deberse a que los aislados analizados podrían no portar plásmidos naturales o a limitaciones metodológicas asociadas al uso de lecturas cortas, que tienden a fragmentar o incluso perder plásmidos

pequeños durante el proceso de ensamblaje. Además, en este estudio no se emplearon herramientas especializadas para la identificación y reconstrucción de plásmidos, como PlasmidSPAdes, Recycler, mlplasmids o plasFlow, las cuales están diseñadas para recuperar secuencias plasmídicas con mayor precisión. Al igual que sería importante utilizar tecnologías de secuenciación de lectura larga que permiten recuperar de manera más fiel la estructura completa de plásmidos y confirmar su presencia o ausencia.

## 13. DISCUSIÓN DE RESULTADOS

El análisis bioinformático realizado en este estudio permitió obtener una caracterización genómica robusta de los cuatro aislados ambientales de *Bacillus paralicheniformis* provenientes de la Ciudad de Puebla. Los resultados obtenidos en cada etapa (desde la evaluación de calidad hasta la comparación genómica) aportan información relevante sobre la estabilidad, variabilidad y posibles adaptaciones funcionales de esta especie, así como sobre su potencial ecológico y biotecnológico.

La anotación genómica mediante "Prokka (v1.14.6)" mostró un número de CDS relativamente estable entre los cuatro aislados, con valores que oscilaron entre 4305 y 4323 (Tabla 1), cifras consistentes con lo reportado en estudios previos (Othman et al., 2024). Esta estabilidad refleja un genoma central altamente conservado, característico de *B. paralicheniformis*. No obstante, se observaron diferencias en el número de genes de ARN, especialmente los tRNA, que variaron de 55 a 121 entre los aislados. Estas diferencias podrían reflejar variaciones en la capacidad metabólica o en la eficiencia traduccional de cada cepa, sugiriendo adaptaciones específicas a los microambientes del aire urbano.

Las proteínas hipotéticas representaron aproximadamente un tercio del total (alrededor de 1,440 por aislado), lo que evidencia la existencia de regiones genómicas aún no caracterizadas funcionalmente en esta especie. Este patrón ha sido descrito también en otros genomas de *Bacillus* (Kanehisa et al., 2016).

El análisis comparativo de genes (Tablas 4–6) reveló la existencia de 32 CDS específicos distribuidos entre los cuatro aislados y un conjunto adicional de genes compartidos parcialmente o con longitudes variables, lo que refuerza la noción de un pangenoma abierto descrito para el género *Bacillus*, en el cual las cepas comparten un núcleo genómico conservado, pero poseen regiones accesorias variables que determinan su plasticidad funcional (Page et al., 2015). Dichas regiones accesorias probablemente influyen en la capacidad de los aislados para adaptarse a condiciones ambientales particulares.

El análisis de genes compartidos con longitudes distintas (Tabla 6) sugiere la presencia de variantes alélicas o ajustes estructurales dentro de genes conservados, un patrón característico de procesos de microevolución intraespecífica. Entre los genes identificados destacan aquellos relacionados con el metabolismo energético y la respiración celular, como las

subunidades 2 y 3 de la citocromo c oxidasa y la deshidrogenasa de NADH. Estas enzimas participan en la transferencia de electrones y en el mantenimiento del gradiente protónico, elementos clave para la respiración aeróbica y la generación de energía (He et al., 2022; Zhou et al., 2022). Su presencia refuerza la idea de que los aislados conservan rutas metabólicas esenciales que facilitan su supervivencia en entornos con variaciones en la concentración de oxígeno. Asimismo, la identificación de proteínas de la familia ParA y Smc sugiere mecanismos de partición cromosómica conservados, que contribuyen a la estabilidad del genoma bacteriano en condiciones ambientales adversas (Gao et al., 2023; Soberón et al., 2024).

Los genes relacionados con el transporte de aminoácidos y metales, como las permeasas GntP, reflejan la capacidad de las cepas para mantener un equilibrio nutricional frente a fluctuaciones en la disponibilidad de nutrientes (Sharma et al., 2022; Wang et al., 2024). Estos mecanismos adaptativos son consistentes con la presencia estable de *B. paralicheniformis* en bioaerosoles urbanos, donde las condiciones pueden ser limitantes y variables. En conjunto, los resultados evidencian que los aislados comparten un genoma central altamente conservado, pero presentan diferencias específicas en el contenido de genes accesorios, lo que respalda la hipótesis de una plasticidad genómica que favorece su adaptación a los microambientes atmosféricos.

El árbol filogenético obtenido con “IQ-TREE 2 (v2.4.0)” y visualizado en “iTOL” mostró una clara separación entre las cuatro cepas analizadas, confirmando la existencia de diversidad intraespecífica. Las cepas identificadas como aislado 3 y aislado 4 mostraron una relación evolutiva más estrecha, mientras que el aislado 1 se ubicó como el más divergente, lo que podría deberse a diferencias acumuladas por mutaciones, recombinación o adaptación local. Este patrón ha sido documentado en otros estudios de *B. paralicheniformis*, donde se reporta una estructura filogenética diversificada impulsada por transferencia horizontal de genes y selección ambiental (Du et al., 2019; Othman et al., 2024).

En síntesis, los resultados confirman la hipótesis planteada: los aislados de *B. paralicheniformis* presentan variaciones genómicas principalmente en genes accesorios, lo que refleja procesos de adaptación evolutiva a las condiciones atmosféricas de la Ciudad de Puebla. Estas diferencias podrían representar ajustes genéticos frente a factores ambientales

característicos de los entornos urbanos, como la radiación UV, la variabilidad térmica y la disponibilidad irregular de nutrientes.

## 14. CONCLUSIÓN

El presente trabajo permitió ensamblar, anotar y comparar el genoma de cuatro aislados ambientales de *Bacillus paralicheniformis* obtenidos de bioaerosoles de la Ciudad de Puebla, revelándose una estructura genómica central conservada junto con un conjunto variable de genes accesorios que reflejan la plasticidad y capacidad adaptativa de esta especie.

El empleo de herramientas bioinformáticas especializadas ("FastQC", "Trimmomatic", "SPAdes", "Prokka", "Roary" e "IQ-TREE 2") garantizó la calidad del ensamblaje y permitió realizar un análisis genómico comparativo confiable. La identificación de genes asociados con procesos respiratorios (citocromo c oxidasa, NADH deshidrogenasa), transporte de nutrientes (permeasas GntP) y estabilidad cromosómica (ParA, Smc) confirma que los aislados poseen mecanismos genéticos que favorecen su supervivencia frente a condiciones ambientales fluctuantes.

El análisis filogenético evidenció diferenciación evolutiva entre los aislados, lo que sugiere la existencia de variaciones locales dentro de un mismo entorno urbano. En conjunto, los resultados respaldan la hipótesis de que la diversidad genómica de *B. paralicheniformis* surge de la interacción entre un núcleo conservado y un genoma accesorio dinámico, responsable de su versatilidad ecológica y potencial adaptativo.

Finalmente, este estudio demuestra que las cepas ambientales de *B. paralicheniformis* constituyen un modelo útil para comprender los mecanismos de adaptación microbiana en la atmósfera, y representan un recurso biológico valioso con potencial biotecnológico y ecológico. Los genes identificados pueden servir como base para futuras investigaciones orientadas al aprovechamiento de *B. paralicheniformis* en la biotecnología ambiental e industrial, así como para profundizar en su rol dentro de la ecología microbiana urbana.

15. REFERENCIAS


**1.** Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Babraham Bioinformatics.

**2.** Aziz, R., et al. (2008). The RAST server: Rapid annotations using subsystems technology. BMC Genomics, 9, 75. https://doi.org/10.1186/1471-2164-9-75

**3.** *Bacillus paralicheniformis* GMB0681 | GMBANK. (s.f.). GMBANK.

**4.** Bankevich, A., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

**5.** Bolger, A., et al. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

**6.** Burrows, S., et al. (2009). Bacteria in the global atmosphere – Part 1: Review and synthesis of literature data for different ecosystems. Atmospheric Chemistry and Physics, 9(23), 9263–9280. https://doi.org/10.5194/acp-9-9263-2009

**7.** Chávez-Almanza, A., Villa-Rodríguez, E., Rojas-Padilla, J., Cantú-Soto, E., Díaz-Quiroz, C., & Verdugo-Fuentes, A. (2025). Genome analysis of *Bacillus paralicheniformis* AA1 isolated from a conventional milpa farming system in the southern region of Sonora, Mexico. Biotecnia, 27, e2552.

**8.** Diene, S. M., et al. (2023). Transcriptional regulation and metabolic pathway control in *Bacillus* species: Insights into antibiotic biosynthesis networks. Frontiers in Microbiology, 14, 1175–1186. https://doi.org/10.3389/fmicb.2023.1175186

**9.** Du, Y., et al. (2019). Comparative genomic analysis of *Bacillus paralicheniformis* MDJK30 with its closely related species. BMC Genomics, 20, 283. https://doi.org/10.1186/s12864-019-5646-9

**10.** Foster, P. L., et al. (2021). DNA mismatch and very-short-patch repair mechanisms in bacteria. Annual Review of Microbiology, 75, 65–85. https://doi.org/10.1146/annurev-micro-042920-112020

**11.** Fronczek, C. F., & Yoon, J.-Y. (2015). Biosensors for monitoring airborne pathogens. SLAS Technology, 20(4), 390–410. https://doi.org/10.1177/2211068215580935

**12.** Gao, T., et al. (2023). Molecular mechanism of ParA-mediated plasmid segregation in *Bacillus* species. Nucleic Acids Research, 51(2), 612–625. https://doi.org/10.1093/nar/gkac123

**13.** Goodwin, S., et al. (2016). Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6), 333–351. https://doi.org/10.1038/nrg.2016.49

**14.** He, J., et al. (2022). Functional and structural analysis of NADH dehydrogenase subunits in *Bacillus paralicheniformis*. FEBS Letters, 596(21), 2648–2661. https://doi.org/10.1002/1873-3468.14371

**15.** Huang, L., et al. (2025). Amino acid uptake systems and adaptive metabolism in environmental *Bacillus* isolates. Frontiers in Microbiology, 16, 1456783. https://doi.org/10.3389/fmicb.2025.1456783

**16.** Illumina, Inc. (2023). Quality scoring in Illumina sequencing systems. Illumina Technical Documentation.

**17.** Jaenicke, R. (2005). Abundance of cellular material and proteins in the atmosphere. Science, 308(5718), 73. https://doi.org/10.1126/science.1106335

**18.** Kushmitha, B., *et al.* (2025). Whole genome sequencing of rice endophyte *Bacillus paralicheniformis* NB stem 4: A potential biocontrol agent for the suppression of pearl millet blast disease. *Physiological and Molecular Plant Pathology, 138*, 102663.

**19.** Kanehisa, M., et al. (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

**20.** Kim, H., et al. (2024). Electron transport and proton gradient regulation by NADH dehydrogenase complex I in Gram-positive bacteria. Journal of Biological Chemistry, 299(4), 104751. https://doi.org/10.1016/j.jbc.2024.104751

**21.** Koonin, E., et al. (2003). Sequence-evolution-function: Computational approaches in comparative genomics. Kluwer Academic.

**22.** Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. Nucleic Acids Research, 49(W1), W293–W296.

**23.** Li, H., et al. (2010). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

**24.** Li, J., et al. (2024). Metabolic adaptation via organic acid transporters in *Bacillus paralicheniformis* under nutrient stress. Applied and Environmental Microbiology, 90(2), e01567–23. https://doi.org/10.1128/aem.01567-23

**25.** Lu, D., et al. (2024). Functional dynamics of cytochrome c oxidase in *Bacillus paralicheniformis* under aerobic growth. Frontiers in Cellular and Infection Microbiology, 14, 1245752. https://doi.org/10.3389/fcimb.2024.1245752

**26.** López-Mendoza, A., & Pérez-Guerrero, J. (2023). *Aerosoles atmosféricos: relevancia en el clima del planeta y su dinámica en zonas urbano-forestales de Puebla*. Revista de Divulgación Científica BUAP.

**27.** Martínez-Sánchez, P., Reynoso-Valencia, D., & Torres-García, M. (2024). Studies on airborne microbiota in Mexico: A review. *Revista Internacional de Contaminación Ambiental, 40*(1), 1–25.

**28.** Miller, J., et al. (2010). Assembly algorithms for next-generation sequencing data. Genomics, 95(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

**29.** Minh, B. Q., et al. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution, 37(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

**30.** Nurk, S., et al. (2013). Assembling genomes and mini-metagenomes from highly chimeric reads. Journal of Computational Biology, 20(10), 714–737. https://doi.org/10.1089/cmb.2013.0084

**31.** Oliveros, J. C. (2007-2015). *Venny 2.1: An interactive tool for comparing lists with Venn's diagrams*. BioinfoGP, CNB-CSIC.

**32.** Othman, M. A., et al. (2024). Comparative genomics of *Bacillus paralicheniformis* reveals regulatory networks for secondary metabolite production. BMC Genomics, 25, 246. https://doi.org/10.1186/s12864-024-10246-9

**33.** Page, A., et al. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

**34.** Palzkill, T., et al. (2023). Metallo-β-lactamases: Structure, function, and evolution. ACS Infectious Diseases, 9(5), 853–864. https://doi.org/10.1021/acsinfecdis.3c00045

**35.** Park, S., et al. (2023). The role of very-short-patch repair endonuclease (Vsr) in genome stability of *Bacillus* species. DNA Repair, 126, 103474. https://doi.org/10.1016/j.dnarep.2023.103474

**36.** Rao, M. P. N., *et al.* (2024). Genome-based approach to evaluate the metabolic potentials and exopolysaccharides production of *Bacillus paralicheniformis* CamBx3 isolated from a Chilean hot spring. *Frontiers in Microbiology, 15*, Article 1377965. https://doi.org/10.3389/fmicb.2024.1377965

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

**38.** Sharma, R., et al. (2022). Functional analysis of GntP family permeases in Gram-positive bacteria. Microbial Physiology, 32, 301–312. https://doi.org/10.1159/000526447

**39.** Soberón, C., et al. (2024). Chromosome partitioning systems in *Bacillus paralicheniformis* and their role in genetic stability. Microbial Genomics, 10(6), 001347. https://doi.org/10.1099/mgen.0.001347

**40.** Songnaka, N., et al. (2024). A novel bacitracin-like peptide from *Bacillus paralicheniformis* NNS4-3 against MRSA and its genomic insights. Antibiotics, 13(8), 716. https://doi.org/10.3390/antibiotics13080716

**41.** Soto-Ramírez, C. (2017). Microorganismos cultivables aislados del aire en la ciudad de Puebla y modelado matemático del transporte aerobiológico (Tesis de maestría). Benemérita Universidad Autónoma de Puebla.

**42.** Wang, L., et al. (2024). Genomic insights into metal-dependent hydrolases and antibiotic resistance determinants in *Bacillus* species. Journal of Bacteriology, 206(7), e00221–24. https://doi.org/10.1128/jb.00221-24

**43.** Zhang, P., et al. (2021). Comparison of de novo assembly strategies for bacterial genomes. International Journal of Molecular Sciences, 22(14), 7668. https://doi.org/10.3390/ijms22147668

**44.** Zhang, X., et al. (2023). Characterization of amino acid transporters in *Bacillus* species and their role in nitrogen metabolism. Microbiology Spectrum, 11(2), e04873-22. https://doi.org/10.1128/spectrum.04873-22

**45.** Zhao, J., et al. (2022). Global airborne bacterial community—interactions with Earth's microbiomes and anthropogenic activities. Proceedings of the National Academy of Sciences, 119(42), e2204465119. https://doi.org/10.1073/pnas.2204465119

**46.** Zhou, Y., et al. (2022). Respiratory chain organization and cytochrome oxidase diversity in *Bacillus* species. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1863(9), 148962. https://doi.org/10.1016/j.bbabio.2022.148962

**47.** Bioinformática CIAD. (2024). Ensamblaje de un genoma de Escherichia coli: métodos y optimización. Centro de Investigación en Alimentación y Desarrollo.