Proyecto integrador en Bioinformática

Entrega 2

Laura del Sol González
Sara Flórez Hernández
Maria Camila Gómez Villegas
Sergio Andrés Morales Toro

Universidad EIA
Envigado – Antioquia
Octubre 2025

Punto 1: Descripción de clasificación taxonómica

Descripción del flujo de trabajo:

Para este punto se hizo la clasificación taxonómica de las lecturas que no se mapearon al genoma de referencia. El análisis se realizó con Kraken2, Bracken y Krona, siguiendo el procedimiento descrito en el archivo Scripts.pdf del punto 1. Primero se convirtieron los archivos .sam a .bam y se extrajeron las lecturas no mapeadas usando Samtools. Después, con Kraken2 y la base de datos minikraken2_v2_8GB_201904_UPDATE, se asignaron las lecturas a sus posibles grupos taxonómicos. Luego, Bracken permitió refinar los conteos para obtener resultados más precisos a nivel de especie. Por último se generaron las visualizaciones con Krona, obteniendo los archivos evol1_krona.html y evol2_krona.html, donde se puede observar la composición taxonómica de las muestras.

Resultados obtenidos:

Evol1:

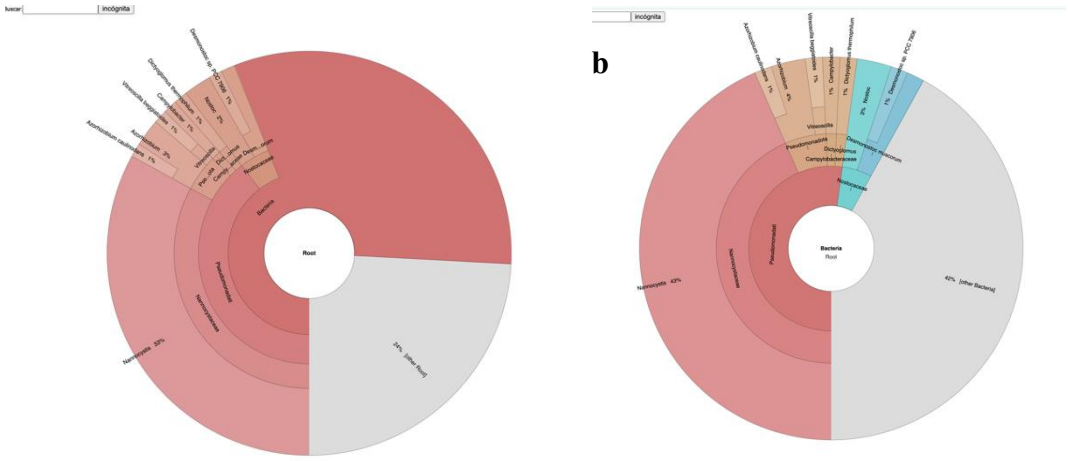

Figura1. Clasificación taxonómica de las lecturas no mapeadas de la muestra evol1 visualizada con Krona. (a) Vista general de la composición bacteriana. (b) Detalle ampliado de los géneros predominantes.

En la muestra evol1 se observó que la mayoría de las lecturas no mapeadas pertenecen al dominio Bacteria, lo que muestra que gran parte de las secuencias provienen de microorganismos bacterianos. El filo más abundante fue Pseudomonadota, con una alta presencia del género Nannocystis, que representó aproximadamente entre el 40 y el 45 % del total. Este resultado sugiere que las lecturas corresponden a bacterias típicas de ambientes naturales y suelos húmedos, ya que Nannocystis pertenece al grupo Myxococcota, conocido por formar estructuras cooperativas y participar en la degradación de materia orgánica. Estas bacterias suelen ser importantes en ecosistemas terrestres porque ayudan a reciclar nutrientes y descomponer compuestos complejos, lo que podría explicar su alta abundancia en la muestra. Además de Nannocystis, se encontraron otros géneros en menor proporción como Azorhizobium, Vitreoscilla, Dictyoglomus thermophilum y Campylobacter. Nos parecio interesante que aparezcan bacterias de ambientes tan diferentes, ya que eso podría indicar que la muestra contiene ADN ambiental mezclado o trazas de microorganismos de distintos entornos.

Tabla 1. Principales géneros bacterianos encontrados en la muestra evol1 y su posible función ecológica.

Género / Especie	Porcentaje estimado	Descripción biológica y relevancia
Nannocystis	40–45 %	Bacteria del suelo perteneciente al filo *Myxococcota*. Participa en la degradación de materia orgánica y procesos cooperativos microbianos.
Azorhizobium caulinodans	3–5 %	Fijadora de nitrógeno, asociada a raíces de plantas leguminosas; su presencia puede reflejar restos de ADN ambiental.
Vitreoscilla beggiatoides	2–3 %	Bacteria aerobia productora de hemoglobina bacteriana (*Vitreoscilla hemoglobin*), importante para la respiración en ambientes pobres en oxígeno.
Dictyoglomus thermophilum	2–3 %	Bacteria termófila, habitante de ambientes calientes, posiblemente presente por contaminación cruzada o secuencias ambientales.
Campylobacter	1–2 %	Bacteria microaerófila presente en ambientes intestinales o acuáticos; su baja abundancia indica posible contaminación ambiental.
No asignados	20–25 %	Lecturas sin clasificación en la base de datos, posiblemente por falta de genomas de referencia o secuencias incompletas.

Con estos resultados se puede ver que la muestra evol1 contiene principalmente bacterias ambientales y saprófitas que son microorganismos que viven en el suelo o en medios naturales y participan en la descomposición de materia orgánica. El hecho de que una parte de las secuencias (20–25 %) no haya podido clasificarse es algo normal, ya que la base de datos MiniKraken2 no tiene todos los genomas disponibles.

Evol 2:

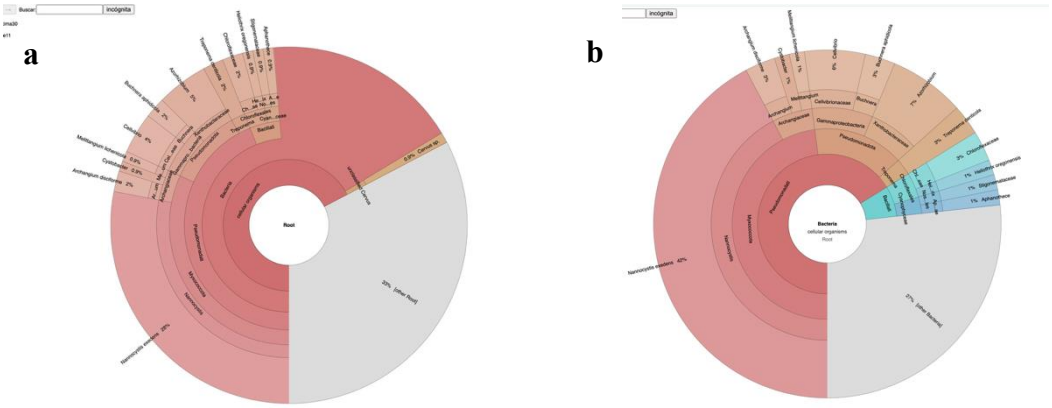

Figura 2. Clasificación taxonómica de las lecturas no mapeadas de la muestra evol2 visualizada con Krona.(a) Vista general de la composición bacteriana. (b) Detalle ampliado de los géneros predominantes.

En la muestra evol2 se observó que, al igual que en evol1, la mayoría de las lecturas no mapeadas pertenecen al dominio Bacteria. Sin embargo, esta muestra presentó una mayor diversidad microbiana, ya que se identificaron más géneros bacterianos y varios filos distintos. El filo predominante fue Pseudomonadota, con una fuerte presencia de la clase Myxococcia y del género Nannocystis exedens, que representó aproximadamente entre un 40 % y 45 % del total de lecturas clasificadas. La alta abundancia de Nannocystis exedens muestra que probablemente se trata de una comunidad bacteriana típica de suelos y ambientes naturales, donde esta especie participa en la

degradación de materia orgánica y en procesos cooperativos entre microorganismos. Además, se detectaron géneros como Archangium disciforme, Melittangium lichenicola, Cystobacter, Cellvibrio y Buchnera aphidicola, entre otros. Estos resultados reflejan que en evol2 hay una mayor variedad de bacterias con funciones ecológicas diferentes, como la descomposición de compuestos vegetales, la fijación de nitrógeno o relaciones simbióticas con otros organismos. Nos pareció curioso que también se detectaran secuencias clasificadas como Cervus sp. ya que probablemente corresponden a contaminación cruzada o a errores de asignación por similitud entre fragmentos de ADN.

Tabla 2. Principales géneros bacterianos encontrados en la muestra evol2 y su posible función ecológica

Género / Especie	Porcentaje estimado	Descripción biológica y relevancia
Nannocystis exedens	40–45 %	Bacteria del suelo del filo *Myxococcota*, asociada a la degradación de materia orgánica.
Archangium disciforme	2–3 %	Bacteria depredadora del suelo que participa en la descomposición de materia orgánica.
Melittangium lichenicola	1–2 %	Bacteria con capacidad de formar esporas, común en ambientes naturales y húmedos.
Cystobacter	1–2 %	Bacteria del suelo con comportamiento cooperativo; forma cuerpos fructíferos.
Cellvibrio	6 %	Degrada polisacáridos vegetales, típica de suelos con abundante materia orgánica.
Buchnera aphidicola	3–4 %	Endosimbionte de insectos; su detección puede deberse a contaminación ambiental o ADN residual.
Treponema denticola	3 %	Bacteria anaerobia, presente en microbiotas animales; posible contaminación ambiental.
Heliothrix oregonensis / Aphanothece	1 %	Bacterias fotosintéticas, probablemente detectadas por ADN ambiental.
No asignados	27–33 %	Lecturas sin clasificación específica en la base de datos, posiblemente por falta de genomas de referencia.

Nos pareció interesante ver que, aunque evol1 y evol2 compartieron géneros similares, evol2 presentó más diversidad bacteriana y funciones ecológicas diferentes, lo que refleja que cada muestra tiene su propio perfil microbiano.

Punto 2: Tabla y discusión de variantes

En los archivos evol1_variantes.vcf, evol1_variantes_filtradas.vcf, evol2_variantes.vcf y evol2_variantes_filtradas.vcf se encuentran los resultados del análisis de variantes detectadas en cada muestra. Los archivos sin filtrar (evol1_variantes.vcf y evol2_variantes.vcf) contienen todas las variantes identificadas por bcftools call, mientras que los archivos filtrados (_variantes_filtradas.vcf) solo incluyen aquellas que cumplieron con los criterios de calidad establecidos (QUAL > 30 y DP > 10). Estas son las variantes consideradas más confiables y precisas.

Posteriormente, se utilizó Prokka, una herramienta ampliamente empleada para la anotación de genomas bacterianos, permitiendo identificar genes codificantes y otros elementos genómicos en el ensamblaje ancestral de Escherichia coli, generando archivos de salida en formato .gff y .faa con la información estructurada necesaria para la base de datos de SnpEff, herramienta utilizada para la predicción funcional de variantes genéticas. Aquí se integraron los archivos .vcf provenientes del análisis de variantes de las cepas evolucionadas (evol_1 y evol_2). Y, se construyó una base de datos a partir de la anotación generada por Prokka y las secuencias del genoma ancestral, todo este proceso se encuentra en el desarrollo de los scripts_punto_3_y_4.pdf. Donde las variantes detectadas fueron clasificadas según su impacto funcional (HIGH, MODERATE, LOW), su calidad, ect. y se

relacionaron con genes específicos. Con bcftools se extrajo la información más relévante de estos archivos (posición, referencia, alternativa, calidad y profundidad) y se convirtió en .csv para poder hacer una visualización en Jupyter Notebook y para hacer un filtrado manual. En este filtrado primero se extrajeron todas las variantes con una calidad mayor a 30 para evol_1 y 37 para evol_2, donde un poco menos de la mitad de las variantes fueron depuradas

A partir de los datos obtenidos del filtrado manual, se eligieron las variantes con los valores de calidad más altos para construir las tablas de resultados que se muestran a continuación. En ellas se resumen las variantes más representativas de cada muestra y su posible interpretación biológica.

En las tablas se pueden ver los campos principales obtenidos del archivo VCF, como la secuencia (contig), la posición genómica, el cambio de bases, el nombre del gen afectado y el tipo de variante detectada, junto con una breve descripción del posible efecto biológico asociado.

Tabla 3. Variantes seleccionadas en la muestra evol1 y su posible interpretación biológica

Secuencia	Posición (bp)	Cambio de bases	Calidad	Tipo de Variante	Gen Afectado	Posible efecto biológico
NODE_12_length_31822_cov_4.522649	3530	T → TG	38.0	frameshift_variant	malT	Regulador del uso de maltosa.
NODE_16_length_28105_cov_4.645312	23540	T→ A	48.0	stop_gained	panF	Transporta pantotenato (vitamina B5).
NODE_27_length_24632_cov_4.384058	21716	TG → TGG	36.0	frameshift_variant	cadA	Descarboxila lisina (tolerancia ácida).
NODE_47_length_19935_cov_4.144920	15935	GAAA → GAAAAA	42.0	frameshift_variant	argC	Síntesis de arginina.
NODE_106_length_12735_cov_4.212618	5233	T → A	30.0	Stop lost and splice region variant	mdoC	Forma glucanos periplásmicos.
NODE_108_length_12718_cov_4.092711	12416	A → C	31.0	Stop lost and splice region variant	yceD	Estabiliza ARN ribosomal.
NODE_124_length_11842_cov_4.158395	9973	T CGGCGCGGCGCGGCGCGG → TCGGCGCGGCGCGG	53.0	frameshift_variant	dcuS	Sensor de fumarato/succinato.

NODE_142_length_10787_cov_4.290719	4697	TGGGG → TGGG	70.0	frameshift_variant	nadR	Regula síntesis de NAD.
NODE_345_length_3705_cov_4.079178	749	T → A	30.0	stop_lost&splice_region_variant	pcnB	Agrega colas poliA al ARN.
NODE_422_length_2271_cov_4.346119	1828	T → A	30.0	stop_lost&splice_region_variant	ogt	Repara ADN metilado.

Tabla 4. Variantes seleccionadas en la muestra evol2 y su posible interpretación biológica

Secuencia	Posición (bp)	Cambio de bases	Calidad	Tipo de variante	Gen afectado	Posible efecto biológico
NODE_1_length_58499_cov_4.208781	7996	C → T	44	Missense variant	ccmC	Proteína C exportadora de hemo
NODE_12_length_31822_cov_4.522649	3530	T → TG	46	Frameshift variant	malT_2	Regulador del uso de maltosa
NODE_16_length_28105_cov_4.645312	23540	T → A	53	Stop gained	panF	Simportador de sodio/pantotenato
NODE_47_length_19935_cov_4.144920	15935	GAAAA → GAAAAA	39	Frameshift variant	argC	N-acetil-gamma-glutamil-fosfato reductasa
NODE_66_length_16713_cov_4.246668	14839	CA → CAA	67	Frameshift variant	fhuA	Transportador de membrana externa/receptor de fagos ferricromo
NODE_66_length_16713_cov_4.246668	14890	TAA → TA	37	Frameshift variant	fhuA	Transportador de membrana externa/receptor de fagos ferricromo
NODE_108_length_12718_cov_4.092711	12416	A → C	47	Stop lost and splice region variant	yceD_1	Proteína de acumulación de subunidades de ARN ribosómico grande YceD

En el Anexo 1 se presenta un archivo en formato Excel con información complementaria sobre los genes mencionados, incluyendo detalles de sus funciones y las variantes encontradas en cada muestra.

En este análisis se encontraron múltiples variantes en las muestras evol1 y evol2, lo que evidencia que las lecturas mapeadas presentan diferencias notorias frente al genoma de referencia. Esto puede

deberse a la diversidad natural de las bacterias presentes o a que algunas especies no están completamente representadas en la base de datos usada. Después del filtrado, se conservaron únicamente las variantes más confiables. Todas estas variantes retenidas y seleccionadas (en los archivos data_evol1.csv y data_evol2.csv) se ubican en regiones codificantes de proteínas y presentan un nivel de impacto alto según SnpEff, lo que significa que probablemente tengan consecuencias severas sobre la función del gen. Entre ellas predominan las mutaciones frameshift, stop_gained y stop_lost, además de algunas en regiones de splicing, y todas conservan codón de inicio, lo que indica una lectura de alta calidad. Estas categorías son importantes porque implican una alta probabilidad de alterar la proteína: los frameshift cambian el marco de lectura y suelen generar proteínas truncadas o no funcionales; los stop_gained introducen terminadores prematuros; y los stop_lost prolongan la traducción más allá del punto final esperado. La identificación de estas mutaciones es clave para entender la variabilidad genética entre las muestras y confirmar que el proceso de filtrado fue efectivo.

En la muestra evol1, las variantes afectan genes con funciones metabólicas y regulatorias importantes. Por ejemplo, malT actúa como regulador del uso de maltosa, y su mutación podría modificar la capacidad de la bacteria para utilizar ciertos azúcares. El gen panE, involucrado en la síntesis de vitamina B5 (pantotenato), presenta una mutación stop_gained que podría reducir la producción de esta coenzima esencial. cadA, que participa en la tolerancia a ambientes ácidos, muestra una alteración que podría ser una respuesta adaptativa al estrés por pH. argC, implicado en la síntesis de arginina, presenta un frameshift que podría afectar la regulación del metabolismo del nitrógeno, mientras que dcuS, un sensor de fumarato/succinato, podría estar modificando la detección de señales metabólicas. Además, yceD y nudC están relacionados con la estabilidad del ARN ribosomal y la síntesis de NAD, respectivamente, lo que sugiere ajustes en la eficiencia de traducción y en el metabolismo energético.

En la muestra evol2, también se encontraron variantes de alto impacto. ccmC, asociado con la exportación de hemo, presentó una mutación missense que podría afectar la respiración celular. Nuevamente, los genes malT_2, panE y argC aparecen mutados, indicando que existen rutas metabólicas comunes bajo presión selectiva en ambas poblaciones. Por otro lado, fluA, que codifica un receptor de fago y transportador de membrana, mostró varias mutaciones frameshift, lo que podría estar relacionado con mecanismos de defensa frente a fagos o con cambios en la permeabilidad celular. Finalmente, yceD aparece otra vez, relacionado con la acumulación de subunidades de ARN ribosomal, lo que refuerza su posible papel en la estabilidad del ribosoma.

En conjunto las variantes detectadas no son aleatorias sino que se concentran en genes esenciales para la supervivencia, la regulación metabólica y la respuesta al entorno. Estos resultados sugieren que las bacterias podrían estar adaptándose a condiciones específicas de cultivo o estrés, modificando vías metabólicas y reguladoras críticas.

Bibliografía:

- Center for Computational Biology, Johns Hopkins University. (2019). *Kraken2: Metagenomic sequence classification system*. Recuperado de https://ccb.jhu.edu/software/kraken2/
- Center for Computational Biology, Johns Hopkins University. (2019). *Bracken: Bayesian reestimation of abundance with Kraken*. Recuperado de https://ccb.jhu.edu/software/bracken/
- National Center for Biotechnology Information (NCBI). (2018). *Krona: Interactive metagenomic visualization*. Recuperado de https://github.com/marbl/Krona/wiki
- Wellcome Sanger Institute. (2021). *SAMtools: Sequence Alignment/Map tools*. Recuperado de http://www.htslib.org/

- Victorian Bioinformatics Consortium. (2023). *Prokka: Rapid prokaryotic genome annotation*. Recuperado de https://github.com/tseemann/prokka
- Universidad Nacional de Córdoba. (2023). *SnpEff: Genetic variant annotation and functional effect prediction toolbox*. Recuperado de https://pcingola.github.io/SnpEff/
- MicrobeWiki. (2023). *Nannocystis exedens*. Recuperado de https://microbewiki.kenyon.edu/index.php/Nannocystis_exedens
- MicrobeWiki. (2023). *Azorhizobium caulinodans*. Recuperado de https://microbewiki.kenyon.edu/index.php/Azorhizobium_caulinodans
- National Center for Biotechnology Information (NCBI). (2023). *Vitreoscilla beggiatoides (Taxonomy ID: 41428)*. Recuperado de https://www.ncbi.nlm.nih.gov/
- Madigan, M. T., Bender, K. S., Buckley, D. H., Sattley, W. M., & Stahl, D. A. (2019). *Brock: Biología de los microorganismos* (15.a ed.). Pearson Educación.
- MicrobeWiki. (2023). *Archangium disciforme*. Recuperado de https://microbewiki.kenyon.edu/index.php/Archangium_disciforme
- MicrobeWiki. (2023). *Melittangium lichenicola*. Recuperado de https://microbewiki.kenyon.edu/index.php/Melittangium_lichenicola
- MicrobeWiki. (2023). *Cystobacter fuscus*. Recuperado de https://microbewiki.kenyon.edu/index.php/Cystobacter_fuscus
- MicrobeWiki. (2023). *Cellvibrio mixtus*. Recuperado de https://microbewiki.kenyon.edu/index.php/Cellvibrio_mixtus
- National Center for Biotechnology Information (NCBI). (2023). *Buchnera aphidicola (Taxonomy ID: 9)*. Recuperado de https://www.ncbi.nlm.nih.gov/
- National Center for Biotechnology Information (NCBI). (2023). *Treponema denticola (Taxonomy ID: 158)*. Recuperado de https://www.ncbi.nlm.nih.gov/
- National Center for Biotechnology Information (NCBI). (2023). *Heliothrix oregonensis (Taxonomy ID: 1097)*. Recuperado de https://www.ncbi.nlm.nih.gov/
- National Center for Biotechnology Information (NCBI). (2023). *Aphanothece sp. (Taxonomy ID: 1194)*. Recuperado de https://www.ncbi.nlm.nih.gov/